22-24 June 2021

Zilina - Slovakia

Proceedings of

The International Conference on

Information and Digital Technologies 2021

IEEE ESRA SRDA CeBMI educational centre

The International Conference on

Information and Digital Technologies 2021

22-24 JUNE 2021
ZILINA, SLOVAKIA

ISBN 978-1-6654-3692-2

ISSN 2575-677X

IEEE Catalog Number CFP21CDT-ART

The International Conference on Information and Digital Technologies 2021

IEEE Catalog Number CFP21CDT-ART
ISBN 978-1-6654-3692-2
ISSN 2575-677X

Conference Secretariat Address for Contacts:
IDT'2021 Organizing Committee
Department of Informatics
University of Zilina
Univerzinta 1, 01026, Zilina, Slovakia
email: idt@fri.uniza.sk

IDT 2021 Programme Committee

Chair **Soda Paolo,** Italy

Co-Chairs: **Deserno Thomas**, Germany

Pancerz Krzysztof, Poland

Zaitseva Elena, Slovakia

Ablameyko Serge, Belarus
Barach Paul, USA
Baraldi Piero, Italy
Berenguer Christophe, France
Brinzei Nicolae, France
Bris Radim, Czech Republic
Cariow Aleksander, Poland
Cepin Marko, Slovenia
Coolen Frank, United Kingdom
Czapp Stanislaw, Poland
Di Maio Francesco, Italy
El-Nouty Charles, France
Filatova Daria, Poland
Fiser Petr, Czech Republic
Janota Ales, Slovakia
Kameyama Michitaka, Japan
Kharchenko Vyacheslav, Ukraine
Kohani Michal, Slovakia
Kor Ah-Lian, United Kingdom
Kostolny Jozef, Slovakia
Kovalenko Andriy, Ukraine
Krsak Emil, Slovakia
Kvassay Miroslav, Slovakia
Kvet Marek, Slovakia
Kvet Michal, Slovakia
Levashenko Vitaly, Slovakia
Levitin Gregory, Israel
Lukac Martin, Kazakhstan
Lukyanchuk Igor, France
Luntovskyy Andriy, Germany
Majernik Jaroslav, Slovakia
Moraga Claudio, Germany
Muhamedyev Ravil, Kazakhstan
Navara Mirko, Czech Republic
Pastor Luis, Spain
Perkowski Marek, USA
Podofillini Luca, Switzerland
Rauzy Antoine, Norway
Shmerko Vlad, Canada
Simic Zdenko, Netherlands
Stankevich Sergey, Ukraine
Stankovic Radomir, Serbia
Steinbach Bernd, Germany
Subbotin Sergey, Ukraine
Sztrik Janos, Hungary
van Gulijk Coen, UK
Xie Min, Hong Kong
Yanushkevich Svetlana, Canada
Zio Enrico, France/Italy

Organizing Committee

Chair: Rabcan Jan, Slovakia
Michal Mrena, Slovakia
Rusnak Patrik, Slovakia
Sedlacek Peter, Slovakia

Reviewers

Ablameyko Sergey
Berenguer Christophe
Bohacik Jan
Brinzei Nicolae
Bris Radim
Caceres Cesar
Cepin Marko
Cimrak Ivan
Cocherová Elena
Coolen Frank
Czapp Stanislaw
Druh Alex
El Nouty Charles
Filatova Darya
Fišer Petr
Izonin Ivan
Janota Aleš
Jämsä Timo
Kharin Yuriy
Kor Ah Lian
Kostolny Jozef
Kovalenko Andriy
Kvassay Miroslav
Kvassayova Nika
Kvet Marek
Kvet Michal
Levashenko Vitaly
Luntovskyy Andriy
Majer Tomas
Majernik Jaroslav
Muhamedyev Ravil
Marton Peter
Palukha Uladzimir
Pancerz Krzysztof
Pastor Luis
Perkowski Marek
Rabcan Jan
Rusnak Patrik
Sedlacek Peter
Shmerko Vlad
Simic Zdenko
Skrinarova Jarmila
Soda Paolo
Stankevich Sergey
Stankovic Radomir
Steinbach Bernd
Subbotin Sergey
Sujar Aaron
Sztrik Janos
Varga Michal
Vaclavkova Monika
Yanushkevich Svetlana
Zaitseva Elena

CONTENTS

PREFACE

Dear participants,

We have the pleasure to present the Program of International Conference on Information and Digital Technologies (IDT 2021). IDT 2021 provides a forum for presentation and discussion of scientific contributions covering the theories and methods in the field of information and digital technologies, and their application to a wide range of industrial, civil and social sectors and problem areas. IDT 2021 is also an opportunity for researchers, practitioners, academics and engineers to meet, exchange ideas, and gain insights from each other. IDT 2021 offers a multidisciplinary platform to address the technological, societal and financial aspects of information systems.

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

- *International Workshop on Biomedical Technologies (BT),*
- *International Workshop on Reliability and Safety Technologies (RaST),*
- *International Workshop on ACeSYRI: Modern Experience for PhD students and Young Researchers (ACeSYRI),*
- *Industrial Centre (IC) - Exhibition and Special Discussion Section on Information and Digital Technologies, etc.*

These workshops thematically extend the conference subjects. We would like to thank colleagues who organized these Workshops. Special mention should be made of the projects and grants with the support of which these Workshops were prepared.

The Int. Workshop on Biomedical Technologies was organized under the project *D̈evelopment of methods of health-care system risk and reliability evaluation under coronavirus outbreak¨* reg. no. PP-COVID-20-0013 of Slovak Research and Development Agency and project KEGA reg. no. 009ŽU-4/2020 titled *C̈reation of methodological and learning materials for Biomedical Informatics – a new engineering program at the UNIZA"*. Need to say about the ERASMUS+ project CeBMI *Ün̈iversity-Industry Educational Centre in Advanced Biomedical and Medical Informatics"* (reg. no. 600973-EPP-1-2018-1-SK-EPPKA2-KA), which activities were coordinated with the organization of the Workshop BT and Industrial Centre.

The Workshop ACeSYRI was organized as the event of the ERASMUS+ project ACeSYRI: *Advanced Centre for PhD Students and Young Researchers in Informatics,* (reg.no. 610166-EPP-1-2019-1-SK-EPPKA2-CBHE-JP).

In the framework of the project *N̈ew methods development for reliability analysis of complex system"* of Slovak Research and Development Agency reg.no. APVV-18-0027 was prepared the conference on IDT 2021 and the Workshop on Reliability and Safety Technologies (RaST), in particular.

Initially, about a hundred papers were submitted for review. Approximately half of these submissions have been recommended by reviewers for the presentation at the Conference and publication in the proceedings. The review process was made by a large number of reviewers, which are gratefully acknowledged for their contributions to the improvement of the quality of the accepted papers. Each paper was reviewed by at least two anonymous reviewers to ensure fair and high-quality reviews. In addition to regular sessions, IDT 2021 offers distinguished Keynote lectures.

We thank Keynote Speakers for offering their unique perspectives on information technologies at the Conference:

- ***Analyzing physical activity data collected with accelerometers by prof. Timo Jämsä (University of Oulu, Oulu, Finland);***,
- ***Cognitive systems for smart applications by prof. Kor Ah-Lian (Leeds Beckett University, United Kingdom)***,
- ***Novel WSN Protocols for Health Care and Critical Applications by Dr. Korhan Cengiz (Trakya University, Edirne, Turkey)***,
- ***Optimization of Convolutional Neural Networks by Dr. Martin Lukac (Nazarbayev University, Kazakhstan)***,
- ***Infinite Petri Nets for Cybersecurity of Intelligent Networks, Grid, and Clouds by prof. Dmitry A. Zaitsev (Supercomputación Castilla y León, Spain)***.

We gratefully acknowledge the Faculty of Management Science and Informatics of the University of Žilina, European Reliability and Safety Association (ESRA), the Czechoslovakia section of IEEE, IEEE Czechoslovakia Section Reliability Society Chapter, and IEEE Technical Committee on Computational Life Sciences.

Organization team of IDT 2021

Secured and Blockchained IoT

Andriy Luntovskyy
BA Dresden University of Coop. Education
Saxon Academy of Studies
Dresden, Germany
Andriy.Luntovskyy@ba-dresden.de

Tim Zobjack
Integration Experts
(Hann & Kropp Consulting GmbH & Co.KG)
Dresden, Germany
Tim.Zobjack@integration-experts.de

***Abstract*—This work examines the architectures, protocols and SW platforms for IoT devices as well as the important issue for data security and safety. Secured and Blockchained IoT can be provided only under use of the combination of the existed approaches possible like crypto-protocols, FW, IDS/ IPS and CIDN. As advanced approach, the Blockchain technology can be step-by-step deployed with the aim to compulsoriness, accountability and liability for IoT solutions. As one of the known BC problems is use of existing computing and resources in a responsible manner. The trade-off between energy-efficient IoT and BC must be found as well as the BC deployment must be appropriate and acceptable in terms of energy consumption. The case studies on the Secured and Blockchained IoT devices are discussed.**

Keywords—IoT; Blockchain; Integration Platforms; CIDN

I. Internet of Things: State-of-the-Art

In general, a distributed networking PHY system with IT components can by commonly described as IoT nowadays. This includes embedded, sensors, RFID/ NFCs, robots, computers, smartphones, tablets and further "intelligent stuff" with processing algorithms and analytics for PHY objects that interact each other. These systems, or "things", are also clearly identifiable and have to be secured. IoT devices (things) can autonomously acquire then partially process and forward collected data. IoT augments [1-6] the convenient distributed IT and wireless sensor networks (WSN). The common Architecture and Components of a IoT device is given in Fig.1.

The classic communication models like C-S and P2P are extended for IoT via the direct development of the so-called M2M systems. The IoT devices can operate, in general, with different frequency of data acquiring: periodically, non-periodically and event-driven. This means that, depending on the system, communication with a networking participant takes place either via the requests from the network with the corresponding return of the desired information, or on the device side when critical events occur. E.g. messaging (warning), if a temperature sensor detects that a certain limit temperature has been exceeded or fallen below. There is no debt for a permanent transmission of data is due to the energy efficiency requirements for IoT systems.

On the one hand, there is a requirement to use as small telegrams and tiny data rates as possible on a sensor network (WSN) by energy savings. Such restrictions cannot be applied to every IoT device, however, as soon as the use of cable systems and fixed networks for efficient power or data transmission no longer makes economic sense, these requirements become even more important. IoT devices are mainly operated by batteries or have their own type of power supply via so-called energy harvesting, e.g. via solar energy, kinetic and thermo-energy converters or wind power.

The protocol stack with multiplicity of protocols for IoT is depicted in Fig.2. The well-known protocols for enabling networks are used in the PHY and Data Link layers, for example, mobile radio 4G, 5G, BT, RFID, NFC and as well as some IoT-specific protocols.

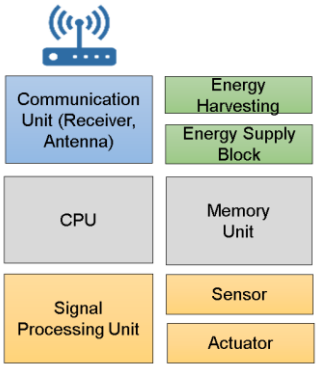


Fig. 1. Architecture and Components of a IoT Device

Most of them use the established combinations with IPv4, IPv6, TCP and UDP for data transfer. An exception is ZigBee, which provides its own protocol for transport and is based on the IEEE 802.15.4 standards (the lowest layers). ZigBee is specified for WSN and corresponds to the basic IoT concept of data economy and energy efficiency. The further development of ZigBee is 6LoWPAN under use of IPv6. The IoT applications have access to the data, provided via various higher-level protocols (HTTP, CoAP, SEP 2.0, MQTT, OneM2M). They are the widely used HTTP can also be used for data queries and data strings transfer. RPL (Routing Protocol for Low power and Lossy Networks) is optimized to use in combination with IPv6 and further WSN. There are several initiatives in industry to promote and spread IoT. The aim is to increase the productivity and flexibility of the producers, while reducing costs and energy consumption. The mostly known are as follows:

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

- Industry 4.0 strategy paper of the Federal Republic of Germany
- Industrial Internet" from General Electric
- Factories of the Future" from the EU.

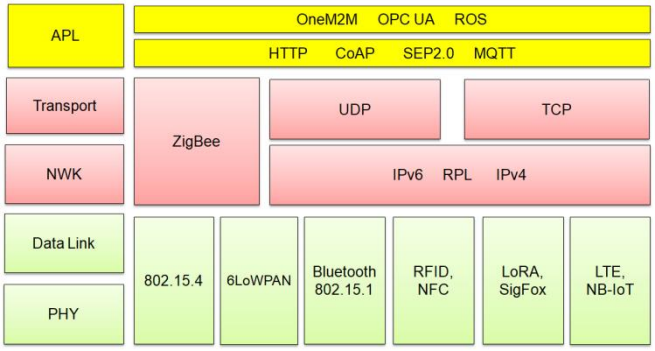


Fig. 2. Protocol Stack for IoT

These approaches are summarized under the term "Industrial IoT" (IIoT), which is sub-area of general IoT. The particular challenge is the integration of two core technologies, the programmable logic controller (PLC) and a system for monitoring and data acquisition (SCADA).

Together with a connected network, they form what is known as operational technology (OT) and LON (Local Operation network). The focus here is on continuous real-time operation and security. The OT is parallel to IT and must be considered separately. At the same time, the challenge is to build interoperability between the two levels. Operational technologies are particularly widespread in so-called critical infrastructure. IIoT can refer to two reference architectures:

- ITU-T Y.2060 of the International Telecommunication Union (ITU) as a general reference architecture for IoT from 2012
- Industrial Internet Reference Architecture V1.7 of the Industrial Internet Consortium (IIC) from 2017 (this reference architecture of the IIC can be seen as a detailed extension of the more generalized architectural description of the ITU).

The available application protocols (refer Fig. 2) for IoT are summarized in Table I.

TABLE I. APPLICATION PROTOCOLS FOR IOT

Protocol	Description
AMQP	Advanced Message Queuing Protocol is an open standard application layer protocol for message-oriented middleware (MOM)
CoAP	REST-based protocol for connecting devices with small QoS. Mainly used for M2M communication
Modbus	Convenient protocol for PLC for coupling of mainly older Legacy devices
MQTT	Light-weight protocol based on OASIS-Standard for message-driven (MQ) data transfer and limited data rate in IoT filed
OPC-UA	Standardized SOA for cross-platform coupling of IoTdevices
REST	Representational State Transfer Protocol, which is used for so-called RESTful Web servicesand based on a set of constraints aimed to max. conformity to HTTP. REST provides interoperability between computers, IoT, robotics,access and processing of textrepresentations of Web resources by using a uniform predefined set of stateless operations. This is in opposite to the other type, so-called SOAP Web services, which use their own sets of operations
SFTP, SMTP	Plain file transfer via FTP or Email
SigFox	Mobile radio service and so-called Low-Power WAN for small data telegrams and wireless coupling of IoT and energy-efficient
SNMPv3	Object-oriented Network Management Protocol based on MIB for message exchange in LAN or correspondingly in IoT scenarios

There is nowadays a large number of platforms from well-known software manufacturers for the management and data processing of IoT devices [4-8, 23,24]:

- IBM Watson IoT Platform
- Microsoft Azure IoT Hub and Win for IoT
- Google Cloud IoT
- Amazon AWS IoT
- SAP Internet of Things, SAP Leonardo IoT, SAP Edge Service.

On the other hand, multiple open-source SW solutions and platforms for IoT device integration can be mentioned like Robot OS, OPC UA, RabbitMQ, MosQuitto, AutomationML tools, which are based the both on the above listed application protocols (refer Table I). These can be taxonomised by their universality and the supported communication protocols: from the mostly simple tools and frameworks up to the whole integration platforms.

A. Case Study 1: OPC UA Platform for IoT Integration

As appropriate examples OPC UA (OPC Unified Architecture) and programming framework, standardized by IEC 62541-2015 [1-8], as well as ROS (Robot OS) and programming framework can be considered. The OPC UA specification as well as the based framework for IoT and Embedded is a multi-part specification and consists of the following parts, layers and tiers as it was depicted in Fig.3: (1) Concepts, Security Model, Address Space Model; (2) Services, Information Model, Mappings; (3) Profiles, DA (Data Access), AC (Alarms and Conditions), Prog (Programs), HA (Historical Access); (4) Discovery, Aggregates, PubSub.

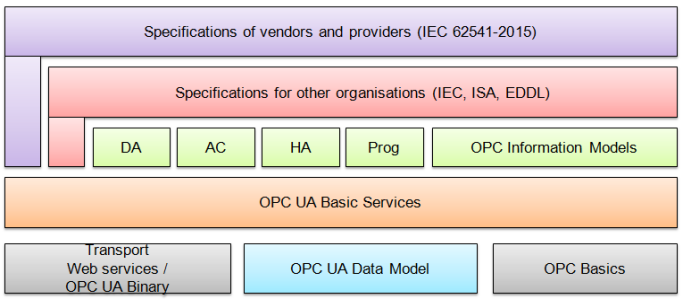


Fig. 3. OPC UA Architecture

The following bus formats can be used: EDDL (Electronic Device Description Language); ISA Bus (Industry Standard Architecture); Profibus as well as further fieldbus formats. The

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

advantages of this OS are as the follows: fault-tolerance, reduced development complexity, availability of the examples on functionality and usage, documents and DB, video streams with tutorials, navigation via way-point navigation and location tracking [1-6].

B. Case Study 2: IoT Application integration via the SAP Platform

The IoT application integration via the SAP platform [7,8] is shown in Fig. 4. The following basic principles are used for such integration:

1. "Out-of-the-box" integration means standardized technologies and applications, as well as preconfigured scenarios for the simple linking of services.

2. "Open integration" means that the connection of non-SAP services is also supported and promoted, based on public APIs.

3. "Holistic integration" means that the type of sender and the receiver is irrelevant, any integration is supported.

4. "AI-driven integration" describes the use of artificial intelligence (AI), which supports the creation of integration scenarios with suggestions based on experience and crowdsourcing.

5. Crowdsourcing is the term for outsourcing internal tasks to external, voluntary participants, in this case the users of the Integration Suite. The integration scenarios created by these users are used anonymously to improve AI.

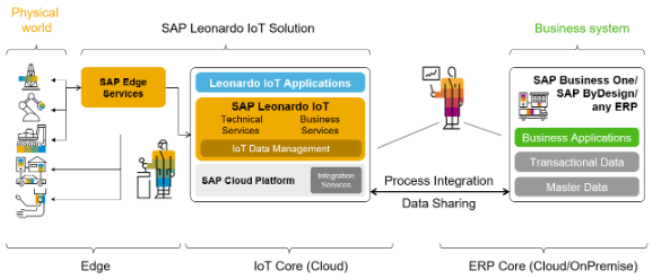


Fig. 4. SAP based IoT Integration (source: https://blogs.sap.com/2019/05/06/sap-leonardo-iot-for-smbs/)

An integration scenario is given in Fig. 5.

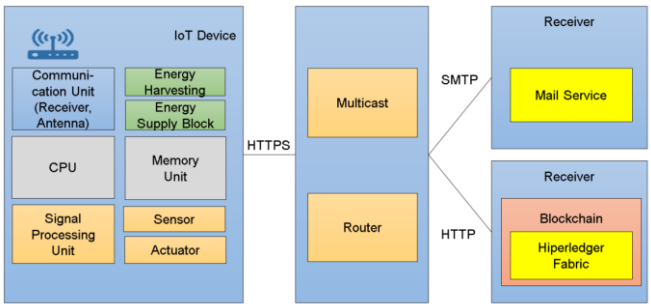


Fig. 5. IoT Integration Scenario [22-24]

An IoT energy efficient device can be represented via DHT11, DHT22, AM2302 sensors for temperature & humidity. The further components like CPU, memory, energy supply, communication units with antenna are provided via a small single-board computer: e.g. Arduino MKR WiFi 1010 or Raspberry Pi 4 Model B; 4 GB, ARM-Cortex-A72 4x, 1,50 GHz, 4 GB RAM, WLAN 11ac, Bluetooth 5, LAN, 4 x USB, 2 x Micro-HDMI.

II. Security Aspects for IoT

The aspects of security can be specially divided into Data Security and Safety, as well as Privacy. According to Gartner Corp. and some further experts [1-5] the robotics and IoT became together more and more "a dangerous technology" because of the standing messaging between the intelligent computing nodes, e.g. usual and ordinary everyday "things" like private cars, walls of apartments and offices, electrical goods, commodity packaging, furniture, valuable papers etc. Any confidential data about the contactor of these things can be immediately transferred in plain text through the available channels: LAN, 4G and 5G mobile, Wi-Fi, WSN, BT, RFID and NFC.

Such technologies can concern the privacy of citizens and, even, harm the state security interests. Therefore, nowadays the above mentioned technologies are thoroughly studied by the leading political and security structures worldwide, in particular, by the EU commissions and in USA. The solution is possible through the use of the corresponding cryptographic protocols, firewalls, intrusion detection systems, that provide reliable protection, but … not only of them.

Therefore, let's discuss the security aspect for up-to-date IoT and platforms more in details. The IoT platforms and apps have to consider the security aspect too and provide the necessary functionality. The following approaches for can be used in this way [1-3, 9-14]:

1. The partition of the interacting robots and "things" to the smaller clusters (piconets).
2. LEACH method for voting for the wireless or mobile cluster heads.
3. Choice of the secured gateways (application level firewalls with antimalware) to linking to the secured clouds for the secured analytic processing.
4. Security risks monitoring (virus, intrusions) on the edges of the robotic clusters and between the clusters.
5. Use for the robotic nodes of the Intrusion Detection Systems (IDS) which operate cooperatively each other.
6. CIDN technology deployment for the interacting robotic nodes.
7. The traffic between the interacting robotic nodes can be encrypted and authenticated via cryptographic protocols like VPN, IPsec, TLS/SSL and/or HTTPS.
8. Use of the private and public Blockchained technologies.

A. *Case Study 3: Hierarchical Security in WSN*

IoT and WSN operate together under considering of energy efficiency (energy harvesting, hierarchy and dynamic clusters, data-aware SW solutions).

Therefore, data security in wireless sensor networks (s. Fig. 6) can also be structured hierarchically (based on foundations of BSI.de). The following techniques such as WPA3, IPv6, TLS and firewalls are used for this multi-level concept. Three hierarchy layers (strata) are provided: SN stratum, CH stratum, GW stratum. The mentioned strata offer the following functionality:

- GW – Gateway, GW implements the protocols DHCP, IPv4, IPv6, as well as firewall filtering and key management and WPA3
- SN – Sensor Nodes, SN provide AES encryption and IPsec
- CH – Cluster Heads, the CH links are implemented by IEEE 802.11 and the SN links by IEEE 802.15.4.

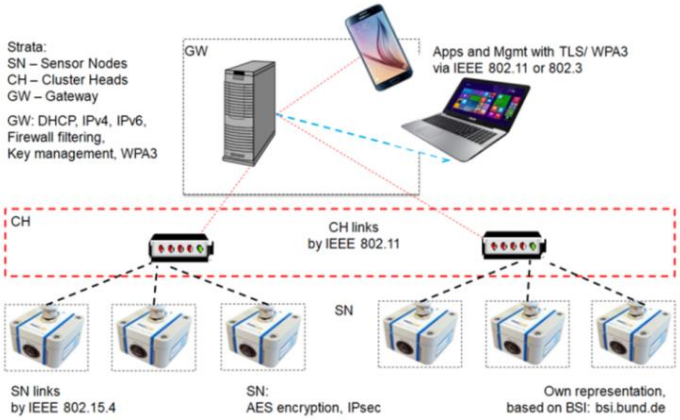


Fig. 6. Security in WSN

The next Multi-Layered Combined Solution for Anti-Drone Defence is discussed below. Drones are becoming increasingly easier and less expensive to access and can thus offer criminals and terrorists a new means of attack (Fig. 7).

Fig. 7. Anti-Drone Defense System AMBOS

Note: "BOS – Behörden und Organisationen mit Sicherheitsaufgaben" (in German). BOS means "Authorities and organizations with security tasks" and is a collective term for institutions that are entrusted with the prevention of dangers for safety and security (emergency services). So-called system "AMBOS" for defence against unmanned aerial objects for BOS (Authorities and organizations with security tasks), developed by the IDMT Fraunhofer Institute for Digital Media Technology, enables the security forces to support the detection of potential threats from drones. The system features possesses hierarchical architecture.

AMBOS sensor system came from the field of audio signal processing and can detect possible threats using four different sensor modalities such as radio, acoustics, electro-optics, infrared and radar. The system AMBOS is developed via the partners from science and industry, which are working since 2017 on information support for the detection and localization as well as an assessment of flying drones [15].

B. *Case Study 4: CIDN for IoT, Sensor Piconets and Robots*

The widespread modern Intrusion Detection Systems (IDS) evaluate and prohibit the potential hackers' attacks that are directed against a computer systems or a network. IDS increase data security significantly in opportune to the classical firewalls which lonely deployment is not satisfying. Intrusion Prevention Systems (IPS) are the enhanced IDS which provide the additional functionality aimed to discovering and avoiding of the potential attacks [1,2]. Nevertheless, as a rule the classical IDS/IPS are operated autonomously. They are not able to detect temporary unknown hackers' threats, which became more sophisticated and complex year by year. Those dangerous threats can serve to disorder the operation of data centres, IoT and robotic clusters round-the-clock in 24/7-mode. Therefore, the cooperation and collaboration of the IDS within a network is of the great meaning [1,2]. A CIDN is a further concept for a collaborative IDS/IPS network intended to bridge over the disadvantage of the standalone defence against the unknown dangerous attacks (Fig.7). The CIDNs allow to the participating IDS as the network peers to share the detected knowledge, experiences and best practices oriented against the hackers' threats [1-3, 9-14].

The main requirements to the construction of a CIDN and the support of such functionality are as follows: efficient communication at short up to middle distance, robustness of the peers (IDS) and links, scalability and mutual compatibility of individual participating peers (single IDS). The typical interoperable networks are as follows: LAN, 3-5G mobile, Wi-Fi, BT and NFC. Collaborative intrusion detection networks (CIDN) consist from multiple IDS-solutions under use of multiple "things", robots, gadgets, PC, end radio-devices and installed firewalls as well of the groups of users which are divided into the clusters – peers (titled as users Alice, Bob, Charlie, Dave etc.).

The coupling between the groups is loosely or tightly. However, the reputation of the users is quietly different (cp. Fig. 8): good, compromised (refer Buddy), malicious (refer Trudy and Mallory). Additionally, the insider attacks to CIDNs are possible (e.g. by Emmy with temporary "good" reputation).The CIDN can efficiently prevent such multiple attacks A1-9 (cp. Fig.7) by peer-to-peer cooperation. This type of networking improves the overall accuracy on the threats

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

assessment. The cooperation among the participating single peers (IDS-collaborators) became more efficient within of a CIDN.

Unfortunately, the CIDN can become itself a target of the attacks and malicious software. Some malicious insiders within the CIDN may compromise the inter-operability and efficiency of the intrusion detection networks internally (Table II). Therefore, the following tasks must be solved: selection of the peers (collaborators), resource and trust management, collaborative decision making [1-3].

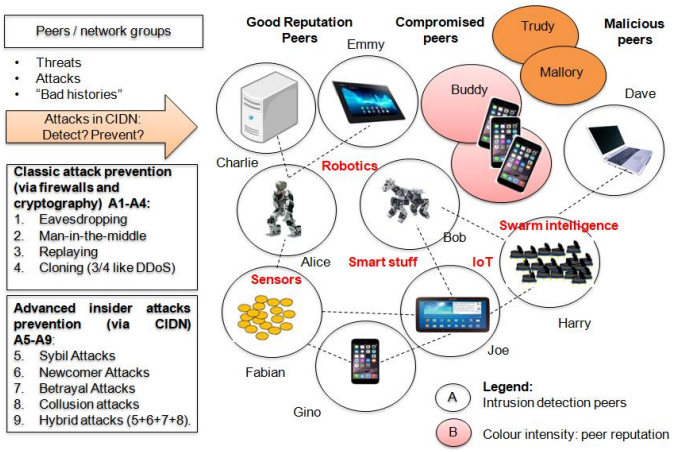


Fig. 8. Example of peer cooperation within a CIDN

TABLE II. CIDN FUNCTIONALITY

Certain CIDN examples	Detection and prevention of the attacks A1-A9	Topology type	Focus	Specialization on the further threats
Indra	+	Distributed	Local	SPAM
Domino	+	Decentralized	Global	Worms
Abdias	+	Centralized	Hybrid	Trojans
Crim	+	Centralized	Hybrid	Social Engineering, web appl. firewalls

C. Case Study 5: Energy-Efficient and Secured Monitoring and Management of Farm Animals via RFID and Wi-Fi

The use of RFID technology in combination with Wi-Fi for secured and energy-efficient monitoring and management of farm animals was depicted in Fig. 9.

III. ON USE OF BLOCKCHAIN FOR SECURED IOT APPLICATIONS

Blockchain (BC) is a cryptographically distributed computer network application [9-22] supporting a decentralized payment system and decentralized financial online transactions in the peer-to-peer (P2P) concept (since 2009). However, the economic success of this crypto-technology will be evident in the next 10 up to 20 years. Potential areas of application can be resulted from the renewed consideration of the main properties of BC:

- Sustainability, general transparency and commitment
- Acceleration of economic workflows, value chains and digitalization processes (so-called "IT in the digital age")
- Completeness and accountability protection in digital Supply-Chain-Management
- Scalability, decentralized network
- Possibility to eliminate the need for an intermediary in financial transactions by making all necessary documents via a BC accessible and fake-proof manner for all involved parties
- Compulsoriness, accountability and liability, especially, for IoT solutions.

The deployment of Blockchain technology speaks for a decentralized financial system. The advantages of such a solution are obvious.

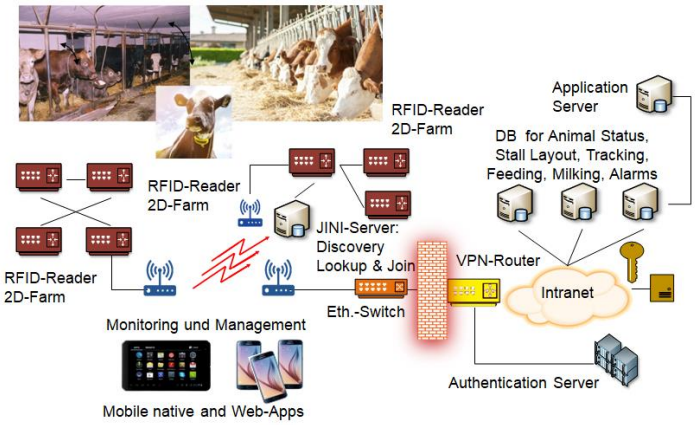


Fig. 9. RFID and Wi-Fi based Monitoring and Management of Farm Animals

BC crypto-technology is also well suited to supporting current crypto-currencies (such as Bitcoin, Ethereum, Ripple, Litecoin, ZCash, Monero, Stellar etc.).

In the Smart Contracting (SC), in the logistics industry, entire supply chains, from the production of raw materials to the ready-on-market products for the end user can be represented via a appropriate BC. Every participants within the logistics chain have insight into the good conditions and WF steps. Subsequent manipulation of data afterwards is completely excluded. Therefore, the trustiness within a supply chain grows between the participants.

There are some pilot projects of a kind of micro-grids, in which the residents can sell on excess self-generated solar power directly to their neighbors. The use of BC technology can also bring some advantages in this area via the SC.

Therefore, there is also multiple BC types, which can be used in public space, by enterprises, for communities or privately (communities, crowdfunding). Table III offers the BC types, ordered by their access model with some provided BC implementation examples.

TABLE III. BC TYPES, ORDERED BY ACCESS MODEL

Blockchain Types		Read OP	Write OP	BC example
Open	no access limitations	everyone	everyone	Bitcoin, Ethereum
	limited access	everyone	authorized	Sovrin
Limited	consortium-limited access	authorized	authorized	Corda, Hyperledger, Quorum
	private limited access	completely private or authorized	network operator	

A flowchart as a decision-making support for BC by Wuest-Gervais [14] is depicted in Fig. 10. Totally, the five following FAQ can help with decision-making to choose the correct BC type: (1) BC with free access; (2) public access limited BC; (3) private access limited BC, or (4) none.

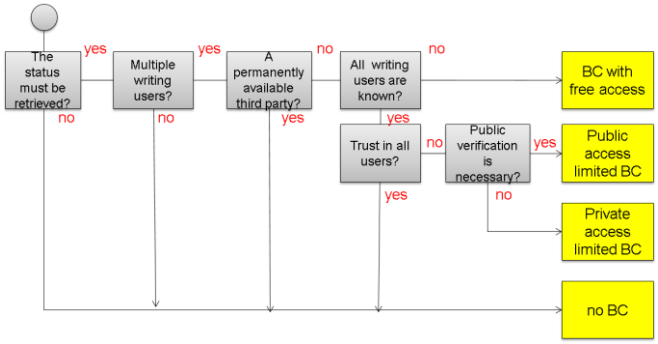


Fig. 10. Wuest-Gervais Flowchart as Decision-Making Support for BC: own representation, based on [14]

A main disadvantage of the BC technology is resource-intensity and energy consumption. Especially, with a view to the large energy and consumption for BC it makes sense for further research and integration consultants to develop further skills in this area. Multiple projects are still in a pilot phase, but with increasing of maturity degree can be expected so that Blockchained IoT can already be provided in mid-term.

A. Constructing A BC

The general "chaining principle" for BC construction is depicted in Fig. 11. Hash and signature algorithms like SHA-256 and RSA are commonly used for chaining, authentication and integrity verification.

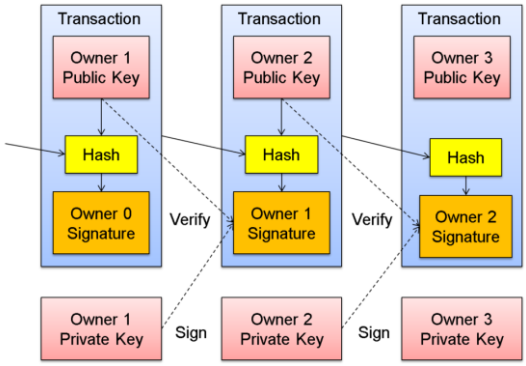


Fig. 11. General principle for a BC

The financial transactions are used as a rule as blocked transactions. Therefore, the BCs possess in the practice the following detailed structures (Fig.12).

The central element of a BC is a block, which contains the transactions that are carried out between the participants within a P2P network. According to Distributed Ledger Technology, the transactions are saved in chronological order. A block has only a limited storage capacity at a time; when the maximum number of transactions is reached, the current block is closed and a new one is created. In addition to the transactions, a block contains a header area with various attributes (see refer Fig. 12). The attributes are usually the block number, a time stamp, the previous block key value (hash) and, depending on the technology, a sample and a nonce optionally.

When a block is completed, a unique key value (hash) is created from the combination of the transactions and the header attributes. This hash is used for validation and is reused in the following block as part of the header area. The individual key values (hashes) of the blocks are interdependent as well as the blocks are cryptographically chained.

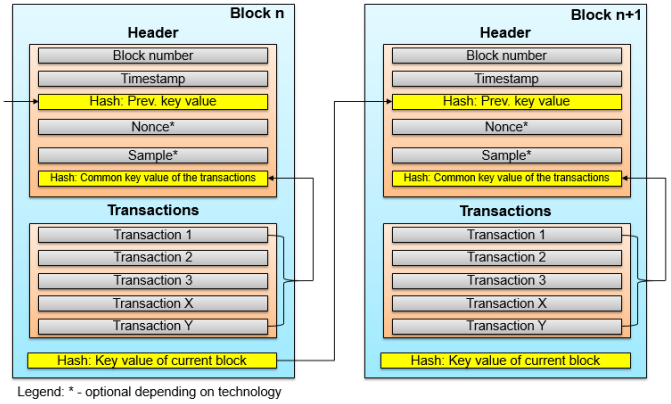


Fig. 12. General BC Structure (detailed representation)

The key values for the block: previous key, common key of the transactions and current key (refer Fig. 12) are generated from the considered block and called the "Hash Values" according to the SHA-256 algorithm from NIST.

By definition, so-called SHA-256 hash algorithm delivers a hash, i.e. an output with a fixed size, from an input of any size and shape. The hash functions are unidirectional, so that conclusions about the origin are not available.

It means that each change for each character in a string in the initial data (i.e. transactions) creates a completely new hash value.

In addition to the SHA-256 hash algorithm, the public key method (Fig. 13) is another important component (e.g. RSA) for integrity control of the data (transactions) within a BC. The method can be used both to verify the sender and to encrypt a transaction. A transaction is the sending of data via the BC network, for example, the digital payment of an invoice under use of cryptocurrency (BTC, ETH, ZEC, XMR).

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

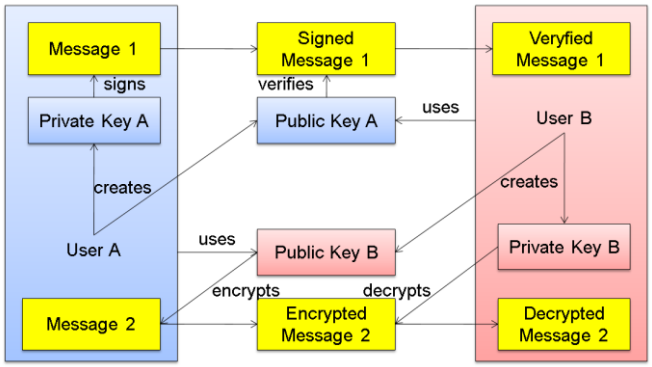


Fig. 13. Public Key Method

Aimed to the distributed transactions several consensus methods and algorithms for the BC block creation must be used. The most important of these methods are listed in Table IV below [9-12, 22].

TABLE IV. DIFFERENT CONSENSUS METHODS AND ALGORITHMS

Method	Description of the algorithm	Dis-advantages	System examples
Proof-of-Work	Participants (so called miners) provide their computing power for a fee in order to calculate the hash values of new blocks. The difficulty of the calculation is adjusted so that blocks are created at regular time intervals.	Very computationally and energy-consuming	Bitcoin
Proof-of-Authority	Only certain nodes within the P2P network can create the blocks, queuing-based choice	Less robust to manipulation	Hyperledger Fabric, Ripple
Proof-of-Stake	The participants possess the P2P network's crypto-currency themselves. The higher the percentage of participation, the more likely it is to choose to create a block for a reward, which is measured in this crypto-currency.	Capital accumulation can put off other participants	Stratis, Reddcoin

The Blockchains can be separated into two big classes. Generally, the public (open) Blockchains are accessible for everyone, although a distinction is made here between:

- a completely open access, for instance, Bitcoin.
- a limited access, to which the participant must first register themselves, for example, Sovrin.

B. Blockchain Architecture

Decentralized, cryptographically secured and unified blocks, their chains and transactions are grouped under a general, global public ledger (account), the structure of which is as follows. The Blockchain, as a networked Public Ledger, consists of participating nodes that represent an efficient P2P communication model. Typical features of the Blockchain are as follows [9-12]:

- Redundancy and synchronization
- Cryptographic hash procedures for integrity assurance and attack safety
- Decentralized management and control of the Blockchain
- Network subscribers are also referred to as Nodes (Full-Nodes, Miners, Validators) and run redundantly with mutual synchronization
- In addition, large block volumes can cause the "Big Data" problem.

Fig. 14 depicts the structure for an exemplary BC.

- The defining block chain (green color) consists of the longest sequence of secured blocks from the origin (genesis, light blue color) to the current block.
- Alternative chains (orange color) became orphan as soon as they are shorter than another chain.

Within the Blockchain architecture between the following basic components can be distinguished: the simple Nodes, the Full-Nodes, and Miner/ Validator [9-14]:

1. Nodes:

Each Blockchain participant (computer, smartphones, tablets, or even clusters) is qualified as Node, if he has installed the corresponding software, which runs based on the Bitcoin protocol or the program code of Bitcoin.

2. Full-Nodes:

- A Node with full local copy of the Blockchain
- Checking for so-called "consensus rules" (consensus method).

3. Miner/ Validator:

- The individual participants or mining pool (high resource requirements regarding hardware and energy consumption)
- Finalising of blocks (Miner – block generation, Validator – proving)
- Externally they act each like a large participant, but in fact, many small blocks are generated for payment in fractions of the crypto-currency units.

However, the following problems occur during the Blockchain operation:

- Enormous energy consumption due to "mining" of crypto-currencies (processing of the hash blocks via its algorithmic complexity).
- Exponential memory growth (including capacity migration between USB media, smartphones, PC, storage media such as SAN / NAS, as well as cloud storages)
- Cryptographic data security is guaranteed, but privacy issues may arise. One-way out is as follows: no processing the complete Blockchain with all the transactions, but only use of excerpts of the Blockchain without a prehistory.

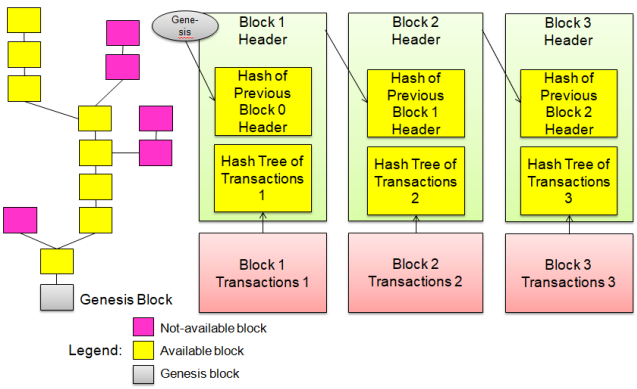


Fig. 14. BC as Distributed Public Ledger: Headers and Blocks within a P2P Hash Tree

C. Recent BC Types

Multiple crypto-platforms can be used for BC integration within different applications: Ethereum Classic, Codius for Ripple, Hyperledger Sawtooth [9-14] etc. The following APIs are used nowadays for these multiple BC types (s. Table V). A detailed comparison between some recent open source BCs [9, 16-22] and their advantages are given in Table VI.

TABLE V. BC APIS IN COMPARISON

Blockchains	Available API
Hyperledger Fabric	REST
Quorum / Ethereum / ETH	JSON-RPC
Corda	RPC
MultiChain / BTC	JSON-RPC
Ripple	REST

TABLE VI. COMPARISON OF RECENT BC TYPES (BASED ON [22-24])

Blockchain Type	Hyperledger Fabric 2016	MultiChain 2015-2019	Quorum 2016
BaaS in a cloud	suitable	suitable	suitable
Private or public P2P network?	Private, participation is only by prior invitation, subject to approval	designed for private networks, subject to approval, both publicly and privately operated	both publicly and privately operated
Open source	open source	open source	open source
Modularity	available	available	available
Suitability as a solution in the corporate environment	suitable	suitable, a modified BC version of BTC	Quorum is the opposite of MultiChain: Like MultiChain is an enterprise version of the public BTC-BC, is Quorum as an enterprise version of the public ETH
Transactions und blocks	All transactions within the BC run via previously defined, encrypted channels Programs are created in the form of "chain code" Chain code management	Public key encryption Since v2.0 SW for so-called "Smart Filters" Two consensus mechanisms: - "Proof-of-Work" (like BTC) - resource optimized Round-Robin	Based on Ethereum/ ETH Multiple alternative consensus mechanisms Programs, which are based on Quorum-BC are called analogously the ETH like "Smart Contracts"
Support via PL	Go, Java, JavaScript	Python, JavaScript, Ruby, PHP, C#	Proprietary PL: Solidity as JavaScript-Dialect
System examples	Collection of container weights by Kuehne+Nagel and Capgemini platform: monitoring and transparency of the container weights for all parties involved in a supply chain	„Air-Quality-Chain" for measured environmental data of the city of Vienna are saved in a MultiChain and made publicly available	This technology has not yet been used productively Bank and Business BC in mid-term

IV. BC BASED APPLICATIONS

However, as the main disadvantages of the BC-based tools and APIs for IoT, the performance reduction by real-time services as well as enormous energy consumption can be mentioned as a critical position.

A further important requirement to IoT is the data security and trust especially by risks of ransomware and other crypto-Trojans like Petya, WannaCry or GandCrab.

In opposite to standard solutions, based on PKI and combined symmetric-asymmetric encryption (RSA) and digital signatures, BC provides its own security decentralized infrastructure, which distinguish from centralized PKI or bilateral Web-of-Trust incorporated in convenient distributed apps.

In Germany, the BC cloud initiatives came from the Telekom.de and Siemens too. So-called augmented BaaS (BC as a Service) BaaS among other XaaS (Everything as a Service) from the cloud can reduce the investment risks for the enterprises in the digital age. IoT can be successfully combined with a correct digitization solution like a BC is.

MindSphere from Siemens as well as further Siemens solutions for Blockchained IoT can be used for food and beverage industry aimed to complete ensure traceability from the farm and field up to the consumers.

Furthermore, Siemens relies on the BC, which secures the tool management during repair of power plants Siemens brings transparency to the supply chain.

The increasing recognition of BC technology in the enterprise cloud environments by "the digital age" has also prompted multiple giants of IT industry, i.e. companies such as MS and SAP to offer their own BaaS among other XaaS. The following kinds of BC are integrated within these software platforms [16-24]: Hyperledger Fabric, MultiChain, Quorum, and Ripple.

A further successful BC example is based on the use of Hyperledger Fabric BC from the Hyperledger Foundation and concerns to the collection of container weights. The worldwide known logistics company Kuehne+Nagel and the consulting company Capgemini have developed a platform for monitoring and transparency of the container weights for all parties involved in a supply chain.

MultiChain from CoinSciences Ltd. is a modified version of the BTC BC, which is specially designed for private P2P networks and can also be adapted to the respective requirements. The proven public key method is used.

Quorum by J. P. Morgan is an open source BC based on Ethereum. It can be operated both publicly and privately and offers several alternative consensus procedures. The software on the Quorum BC offers analogously Ethereum SC network. Thus, the software has some legal restrictions and cannot yet be directly deployed; probably, in mid-term as bank and business BC.

A. Case Study 6: BC with Framework MS Bletchley

MS Bletchley was launched as a specific BaaS (Blockchain as a Service) and as a part of the Azure cloud platform (Fig. 15).

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

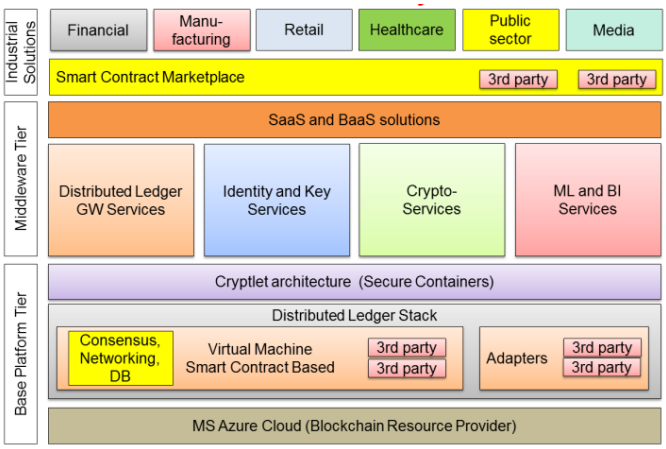


Fig. 15. Blockchain Platform MS Bletchley [16]

The main goal of the framework introduction [16] was the acceleration of the development of the practical Blockchain applications aimed to financial institutions, manufacturing, retail, healthcare, public sector, media on a common platform. The above mentioned platform contains three following layers: Base platform tier; Middleware (MW) tier; Industrial solutions layer.

The framework MS Bletchley introduces a new concept: Blockchain middleware (MW) with so-called cryptlets. To SaaS and BaaS belong so-called ML and BI functionality (ML and Business Intelligence). The Blockchain MW components support the core features of services in the clouds, such as identity management, analytics, or machine learning. Based on the Azure cloud, these components can work with different Blockchain-based technologies aimed to creation of the practical applications. The cryptlets are the building blocks for the Blockchain technology.

They are designed to ensure secure communication between the Azure cloud, the MW ecosystem and each specific customer's technology. The interoperability of the applications with Azure cloud and Azure stack is secured, as well as with the 3rd parties, the clouds like AWS, Google and further private clouds.

B. Case Study 7: BC with SAP Cloud Platform

Three BC frameworks are currently available in SAP Cloud Platform [6-8, 16, 24]: Hyperledger Fabric, MultiChain and Quorum. These services enable the creation and operation of own nodes in the corresponding BC P2P networks as well as integration of external networks. A BC business service is specially designed for cross-company communication via these BCs of the types: Hyperledger Fabric, MultiChain and Quorum. The services support the software, i.a. under IoT, which are based on BCs and provide compulsoriness, accountability and liability for their WF steps and IoT communication. The following BC components can be provided:

- A timestamp service for the BC.
- A proof-of-state service that can be used to write an object in JSON format to the BC.
- A proof-of-history service for displaying the version history of an object in the BC.
- All of these components can be used via standardized REST APIs.

C. Case Study 8: Mining of Cryptocurrency and Resource consumption by Blockchain

At least five factors are necessary and must be considered when calculating the profitability of the mining [1,2,9,10]:

- Investment (basis device costs): hardware investment for a Mining Rig
- Electricity costs: in EUR per kWh
- Energy consumption: electrical power of Mining Rig in KW
- Hash-rate: how much hash values can be computed each second?
- Mining difficulty: factor has always a new actual value!
- Pool fees: how much in % belongs to the joined pool
- Unit „reward“ per Block: Bitcoin amount for each new computed block (or for other units)
- Unit price: exchange course for the crypto-currency BTC (or other units like XMR, ETH, ZEC, LTC).

There are different multipurpose and specialized devices for mining available, so-called Mining Rigs. The old good PCs or plain smartphones can be used too but under considering of the energy consumption problems. The following types of devices can be deployed to generate the hash values for mining process:

- CPU mining: powerful processor is required
- GPU mining: powerful graphics card is required
- ASIC mining (Application-Specific Integrated Circuit, s. Table VII).

TABLE VII. FEATURES OF A MINING DEVICE (SOURCE: AMAZON.DE)

Features of AntMiner S15	**High-Performance Mode**	**Low-Energy Mode**
Hash rate	28 TH/s	18TH/s
El. power	1596 W	900 W
Power Efficiency	57 J / TH	50 J / TH
NW connection	Ethernet	Ethernet
Weight and dimensions	7kg, 240 mm x 178 mm x 296 mm	

The practical experience has shown that in a lot of cases the Mining of the crypto-currencies like BTC, ETH, ZEC, XMR etc. leads unfortunately to “no reward” cases due to a large energy consumptions as well as essential costs (CAPEX + OPEX) [9,10].

V. CONCLUSIONS AND EVALUATION ON USE OF BC WITH IOT SCENARIOS

1. Based on the statistics and predictions of leading consulting firms like Gartner’s, it can be concluded that IoT and BC became the subjects of intensive research and development in recent years. Multiple IT companies have started own pilot projects to test them as new opportunities for the digital age.

2. This work examined the architectures, protocols and SW platforms for IoT devices as well as the important issue for data security and safety.
3. A BC is a digital ledger hosted by a P2P network that stores a digital record of transactions as discrete secured blocks.
4. Secured and Blockchained IoT can be provided only under use of the appropriate combination of the existed approaches possible like crypto-protocols, FW, IDS/ IPS and CIDN. As advanced approach, the BC technology can be step-by-step deployed with the aim to compulsoriness, accountability and liability for IoT solutions.
5. As one of the known BC problems is use of existing computing and resources in a responsible manner. The trade-off between energy-efficient IoT and BC must be found as well as the BC deployment must be appropriate and acceptable in terms of energy consumption.
6. The case studies on the Secured and Blockchained IoT devices are discussed. The focus is shifted on the integration using the platforms. However, it remains to be expected that these integration platforms for IoT and BC will be equipped in the future with additional functions, which improve handling and increase their acceptance.

ACKNOWLEDGMENT

The authors' great acknowledgement goes to the colleagues from BA Dresden (Saxon Academy of Studies), Integration Experts Dresden (Hann & Kropp Consulting GmbH & Co.KG) and Lviv National Polytechnic University, especially to Prof. Dr. habil. A.Haensel, Mrs. Ilona Scherm, Mr. Egon Hann, N.Liebich, M.Stoll, E.Zumpe, Dr. T.Maksymyuk, Dr.D.Guetter for inspiration and challenges by fulfilling of this work.

REFERENCES

[1] A.Luntovskyy, J.Spillner. Architectural Transformations in Network Services and Distributed Systems: Service Vision. Case Studies, Springer Nature, 2017, 344p. (Monograph, ISBN: 9-783-6581-484-09).

[2] Luntovskyy, A., Gütter, D. Moderne Rechnernetze - Lehrbuch, Modern Computer Networks - Handbook, Springer (2020), 481 p. + 265 pict., ISBN: 978-3-658-25616-6 (in German, 11 July 2020), vol.1.

[3] Luntovskyy, A., Gütter, D. Moderne Rechnernetze - Übungsbuch, Modern Computer Networks - Exercises, Springer (2020), 145 p. + 44 pict., ISBN: 978-3-658-25618-0 (in German, 11 July 2020), vol. 2.

[4] Mahmood Zaigham (Ed.): Fog Computing: Concepts, Frameworks and Technologies, Springer 2017, London, ISBN 978-3-319-94890-4.

[5] Jamil Y. Khan, Mehmet R. Yuce (Eds.). Internet of Things (IoT): Systems and Applications, 2019, New York, Jenny Stanford Publishing, ISBN 9780429399084, 366 p.

[6] Lueth, Knud Lasse. The 25 best IoT Platforms 2019 - Based on Customer Review (Online 2020): https://iot-analytics.com/the-25-best-iot-platforms-2019/.

[7] SAP Cloud Integration (Online 2020): https://www.sap.com/.

[8] SAP Leonardo for IoT (Online 2020): https://blogs.sap.com/2019/05/06/sap-leonardo-iot-for-smbs/.

[9] Survey on Crypto-platforms (Online 2002) https://hackernoon.com/top-blockchain-platforms-to-watch-out-in-2019-aa80e336a426/.

[10] A.Luntovskyy, D.Guetter. Cryptographic Technology Blockchain and its Applications, in "Advances in Information and Communication Technologies", Springer (ISBN: 978-3-030-16769-1), LNCS "Processing and Control in Information and Communication Systems (Int. Conf. UkrMiCo-2019)" (eds.:M.Ilchenko, L.Globa et al.), 2019, pp. 14-33 (https://link.springer.com/book/10.1007/978-3-030-16770-7).

[11] MIT Blockchain Course (Online 2020): http://executive-education.mit.edu/MIT-Blockchain/Online-Course/.

[12] A.Antonopoulos, G.Wood. Mastering Ethereum: Building Smart Contracts and Dapps, 2019, O'Reilly Media, 345p., ISBN:978-1491971-949.

[13] Smart Contracts (Online 2020): http://www.icertis.com/.

[14] Karl Wuest, Arthur Gervais. Do you need a Blockchain? ETH Zurich & Imperial College London (Online 2020): https://eprint.iacr.org/2017/375.pdf.

[15] AMBOS Closure – Fraunhofer FKIE (Online 2020): https://www.fkie.fraunhofer.de/.

[16] MS Bletchley (Online 2020): https://github.com/Azure/azure-blockchain-projects/blob/master/bletchley/bletchley-whitepaper.md/.

[17] Leske, Christophe, Göbel, Andreas, Joswig, Steffen (2020): Blockchain mit SAP, 1. Auflage.

[18] Codius: Open-source Hosting Platform for Smart Programs (Online 2020): https://codius.org/.

[19] Hyperledger Sawtooth (Online 2020): https://www.hyperledger.org/projects/sawtooth/.

[20] Blockchain as a Service: Teamwork ohne Daten-Grenzen (Online 2020): https://www.t-systems.com/blockchain/.

[21] Siemens Blockchain (Online 2020): https://www.plm.automation.siemens.com/.

[22] Tim Zobjack, Andriy Luntovskyy. Blockchained IoT: Verbindlichkeit in der dezentralisierten Welt smarter Dinge, BA Magazine "Wissen im Markt", Year 4, Issue 4 Nov. 2020 (in German), 9 p., ISSN 2512-4366, Berufsakademie Sachsen, URL: https://www.ba-sachsen.de/

[23] Siemens Blockchain IoT (Online 2020): https:// new.siemens.com/.

[24] SAP Cloud Platform unterstützt Blockchain-Frameworks, Kap. 2 (in German, Online 2020): https://www.edvbuchversand.de/productinfo.php?replace=false&cnt=productinfo&mode=2&type=2&id=rw-6914&index=2&nr=0&window=edvbv&art=Leseprobe&preload=false

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Impact of Fairness Constraints on Average Service Accessibility in Emergency Medical System

Marek Kvet
University of Žilina, Faculty of Management Science and Informatics
Univerzitná 8215 / 1
010 26 Žilina, Slovakia
marek.kvet@fri.uniza.sk

***Abstract*—Healthcare represents one of the most important disciplines to ensure certain life quality for human beings. Its importance can be fully shown mainly in emergency situations, in which health or life finds itself in direct danger. Emergency medical service systems face significant challenges and they should perform as effectively as possible. From the viewpoint of clients, to whom the associated service is provided, one of the most important and most frequently used criteria in system optimization takes into account service accessibility for clients. Thus, the usual objective can be expressed by minimization of the distance to the nearest source of service from each client location. This paper focuses on additional fairness constraints, which take into account those clients, which are far from any located service center. Presented computational study is aimed at investigation how these fairness constraints affect the computational demands of the problem and how much they may affect the average service accessibility for all clients.**

Keywords—location science, emergency medical system, average service accessibility, fairness constraints, radial approach

I. INTRODUCTION

Healthcare ambulances, fire brigades, police stations, public administration systems and many other forms of public service systems are designed and established to keep, control and manage certain level of safety, health and life quality of served population [1, 5, 6, 15, 16, 22]. In this paper, we put emphasis on those areas of operational research field that find their application in the medical sphere. Special attention is paid to the emergency medical service system in Slovakia.

Emergency medical services (EMS), also known as ambulance services or paramedic services, are operated to provide urgent pre-hospital treatment and stabilization in such cases, in which life or health becomes suddenly deteriorated. If necessary, the patient, who is directly dependent on the provision of urgent medical care, is transferred to a hospital or other specialized medical facility for further treatment or cure [3, 15, 24, 25, 26].

In most places, any person can summon the EMS via an emergency telephone number, which puts them in contact with a control facility. Then, the operating center dispatches a suitable resource to solve the emergency situation. Usually, the mentioned resource used to satisfy the demand for service represents an ambulance, which can be understood as the primary specially equipped vehicle for delivering EMS. Naturally, the EMS agencies may also provide many different services, i.e. non-emergency patient transport, and some have rescue squads and other equipment to provide technical rescue services. The ambulances are located in service centers called EMS stations. From these stations, they are sent into the demand points, their staff (doctors and paramedics) provide necessary treatment and after completion of the service, they return to their common EMS station, from which they were dispatched.

Since the number of ambulances is limited and the staff operating the EMS stations and the individual ambulances is also finite and small, it is not possible to establish an ambulance in each node of the transportation network representing a possible demand point. Therefore, the main effort of the operations researchers and systems designers consists in finding such EMS stations deployment, which would guarantee the highest possible access to the service for served clients. Based on this short explanation, the problem of searching for the optimal service center deployment - not only in EMS systems, but generally - belongs to the family of location problems studied and successfully solved by many professionals in Operations research and Applied Informatics fields. Currently, various exact and approximate approaches to large problem instances are available [7, 8, 9, 11, 13, 19].

As far as the service accessibility is concerned, several facts must be mentioned. First, the number of operated EMS stations is crucial. The higher is this number; the better is the access to service for clients affected by sudden and randomly occurred danger. Furthermore, the number of stations is usually given and the system optimization consists in finding the most suitable locations for them. Let us return to the description of basic performance of an EMS system reported in previous paragraphs. When an emergency occurs, then the nearest ambulance or any other resource is dispatched from its EMS station to the demand point by the operating center to fulfill the basic principles of first aid. Thus, the associated service is provided on the scene. Obviously, the accessibility of the urgent healthcare service can be measured by the distance that the vehicle must travel to get to the demand point of the affected patient. Some studies replace the distance by so-called response time, which elapses until the client location is reached from the EMS station. Without any loss of generality of studied models and principles, we can assume that the response time is proportional to the network distance and that is why we will compute with the network distances in this study. Of course, the shorter is the distance or time, the higher is the rescue service accessibility for clients. Maximization of service accessibility for all clients therefore means minimizing total distance traveled by all ambulances from their stations to the clients' locations. From the point of integer programming, the model of the weighted *p*-median problem can be used to solve the problem of finding the optimal EMS stations deployment [10, 15, 17, 18, 19].

A big disadvantage of the weighted *p*-median problem with a min-sum criterion consists in the fact that the total

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

distance is minimal, but there are some clients, whose distance from the nearest EMS station exceeds the average value several times. It must be realized that those clients pay the same taxes, from which the EMS system is funded, as those clients, who are much nearer to the source of provided service. No wonder such a solution can be considered unfair.

Recently, several fairness approaches to EMS design problem have been developed [4, 12, 14, 21]. Some of them minimize only the distance of the worst situated clients [14], some of them combine the min-max approach with a standard min-sum criterion [21] and other are based on so-called lexicographic approach, which makes the solving process iterative [12, 23].

The main research topic of this paper consists in maintaining a certain level of fairness in service access, when the mathematical model is being formulated. The fairness can be achieved by adding special constraints to the model. Presented computational study is aimed at investigation how these fairness constraints affect the computational demands of the problem and how much they may affect the average service accessibility for all clients. For such parameter settings, which make the original problem infeasible, we make a separate study. In the additional research, we try to find the minimal number of EMS stations, which are necessary to maintain given fairness level. Presented experiments were performed to meet the research objectives. To make the study practice-oriented, we use real-world benchmarks of existing EMS system in Slovakia.

II. Motivation and Research Background

The motivation for the study of fairness and extending the weighted *p*-median model by special constraints comes from the following analysis of existing EMS system in Slovakia, which is summarized in Table I.

The emergency medical service in Slovakia is provided by private EMS agencies, which operate in eight self-governing regions. The regions are depicted in Fig. 1.

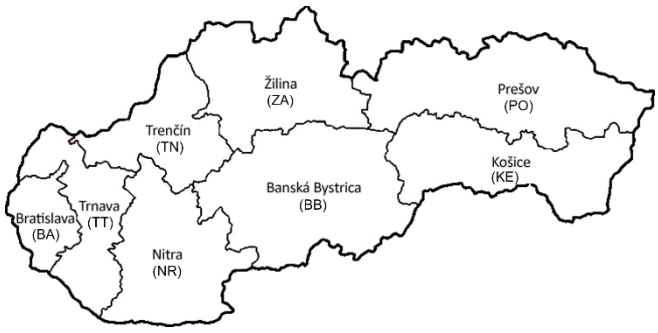


Fig. 1. Self-governing regions of Slovakia

In each self-governing region depicted in Fig. 1, given number *p* of EMS stations is spread over the geographical region to satisfy demands of clients located in *n* nodes of the transportation network. Some of the EMS stations are equipped by more than one ambulance. The total number of vehicles in each region is reported in the column of Table I, which is denoted by *NoV*. It must be noted that within this computational study, we use a macroscopic level of transportation network modelling. It means that each city or village is represented by one node and all inhabitants of individual community are aggregated to one point. Even if the city of Zilina has 4 EMS stations in different city districts and parts of the city, in our models, there is only one station equipped with 4 ambulances. The most interesting part of the current EMS stations deployment analysis is reported in the right part of the table. *MRD* denotes the maximal relevant distance between clients and their nearest located EMS station. The sum of distances from each client locations multiplied by the number of clients sharing these locations to the nearest station is reported in the column denoted by *minSum*. Finally, the last column denoted by *AvgDist* contains the values of average distance.

TABLE I. Analysis of crrent EMS stations deployment for the self-governing regions of Slovakia

Region	*n*	*p*	*NoV*	*MRD*	*minSum*	*AvgDist*
BA	87	14	25	19	21842	3.60
BB	515	36	46	23	32476	4.91
KE	460	32	38	25	36363	4.59
NR	350	27	36	21	38831	5.63
PO	664	32	44	34	42740	5.22
TN	276	21	26	30	26683	4.49
TT	249	18	22	38	31582	5.68
ZA	315	29	36	24	31955	4.62

As we can observe, there are two highlighted columns in Table I to show the problem mentioned in Section I. Current EMS stations deployment can be legitimately considered unfair. If we compare the maximal relevant distance *MRD* to the average value *AvgDist*, we can see that in each region there are same clients' locations very far from any located EMS station and their distance to the nearest service provider exceeds the average value more than five times. Therefore, the necessity of system optimization with fairness consideration does not need to be specially justified.

III. Fair Optimization of Current EMS System

This section is devoted to the formulation and explanation of mathematical models studied in this computational study. The first subsection contains the radial formulation of the weighted *p*-median problem as the core model for the further fair approach. After that, we discuss the extensive fairness constraints and fairness evaluation in general. At the end of this section, we present the concept of additional study, which enables us to find the minimal number of EMS stations, for which it is possible to meet strong fairness requirements, which are not feasible for current number of operated stations.

A. Basic mathematical model

The weighted *p*-median problem can be generally formulated as a task to choose *p* elements from the set *I* of all possible facility locations in order to optimize given quality criterion. In our case to minimize total distance between served clients and their nearest EMS station. Let symbol *J* denote the set of the clients' locations. Each clients' location $j \in J$ is described by the constant b_j, which expresses the number of clients sharing this location. Finally, the last input of the problem consists in the matrix $\{d_{ij}\}$, which represents the distances from the candidates for EMS stations located at $i \in I$ to the clients located at $j \in J$.

The output of the problem solution is defined by the vector **y** of location variables $y_i \in \{0, 1\}$ introduced for each $i \in I$. The decision variable y_i takes the value of one if the EMS station is located at the location $i \in I$ and it takes the value of zero in the opposite case.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

After introducing the input data and the decision variables, the problem can be formulated by the expression (1) in the following way.

$$\min\left\{ f(\mathbf{y}) : i \in I, y_i \in \{0, 1\}, \sum_{i \in I} y_i = p \right\} \quad (1)$$

The minimized quality criterion of the design $f(\mathbf{y})$ can take many forms. For our purpose, the objective expressed as the total distance from the clients' locations to the nearest EMS station can be formulated as follows.

$$f(\mathbf{y}) = \sum_{j \in J} b_j \min\left\{ d_{ij} : i \in I, y_i = 1 \right\} \quad (2)$$

As we can see, the above expression (2) is non-linear. To complete the model with all linear constraints and objective, the radial approach can be applied [8, 11, 13, 19]. The radial model formulation is based on the following assumptions: First, we do not necessarily need to know the assignment of clients to the located EMS stations, if each EMS station has enough capacity to serve all assigned demands. The location-allocation model directly produces the assignment, but it can be easily computed. The most important information consists in the value of distance from a client to the nearest located station and we do not need to know, which EMS station it is. Furthermore, the basic form of the weighted *p*-median problem expects only one EMS station to provide the urgent medical care to the individual client, so it is important to identify this distance. To formulate the radial model, several additional variables and input data need to be introduced.

Let the symbol m denote the maximal distance in the matrix $\{d_{ij}\}$, i.e. $m = \max\{d_{ij}: i \in I, j \in J\}$. In this computational study, we assume that all values in the matrix are integer. Of course, the radial model can be easily adjusted for real values. For each clients' location $j \in J$ and for each integer value $v = 0, 1 \ldots m\text{-}1$ we introduce a binary variable $x_{jv} \in \{0, 1\}$, which takes the value of one, if the distance d_{j*} from the client located at $j \in J$ to the nearest EMS station is greater than the value of v and it takes the value of zero otherwise. Then, the expression (3) holds for each $j \in J$.

$$d_{j*} = \sum_{v=0}^{m-1} x_{jv} \quad (3)$$

Similarly to the set-covering problem, a binary matrix $\{a^s_{ij}\}$ must be computed according to the formula (4).

$$a^v_{ij} = \begin{cases} 1 & \text{if } d_{ij} \le v \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i \in I, j \in J, v = 0,1,\ldots m-1 \quad (4)$$

After these preliminaries, the radial model of the weighted *p*-median problem for obtaining the optimal EMS stations deployment can be formulated by the expressions (5)-(9).

$$\textit{Minimize} \qquad \sum_{j \in J} b_j \sum_{v=0}^{m-1} x_{jv} \quad (5)$$

$$\textit{Subject to}: \quad x_{jv} + \sum_{i \in I} a^v_{ij} y_i \ge 1 \quad \text{for } j \in J, v = 0, 1, \ldots, m-1 \quad (6)$$

$$\sum_{i \in I} y_i = p \quad (7)$$

$$y_i \in \{0, 1\} \quad \text{for } i \in I \quad (8)$$

$$x_{jv} \in \{0, 1\} \quad \text{for } j \in J, v = 0, 1, \ldots, m-1 \quad (9)$$

The quality criterion of the design formulated by the objective function (5) expresses the sum of distances from all clients to their nearest EMS station. The link-up constraints (6) ensure that the variables x_{jv} are allowed to take the value of 0, if there is at least one center located in radius v from the location j and the constraint (7) limits the number of located stations by p. The last two series of obligatory constraints (8) and (9) keep the domain of the decision variables y_i and x_{jv}.

As it was mentioned in previous sections, the radial model of the weighted *p*-median problem itself does not reflect any requirements for fairness at all. It is only the core model, to which the fairness constraints need to be incorporated. Despite the fact that the average distance is minimal, we cannot talk about a fair solution. As it will be reported in the case study, there can be many clients, whose distance to the nearest source of service exceeds the average value many times.

B. Fairness constraints and fairness evaluation

The weighted *p*-median problem itself does not reflect any request for individual fairness except minimizing the average distance from all clients' locations to the nearest EMS station in the objective function. Therefore, it must be extended by additional constraints. In this study, the fairness demands will be expressed in the following way.

Let us introduce a new parameter D (does not necessarily need to be integer), which represents a "critical" distance – see Fig. 2. Furthermore, we set up a parameter B to limit the number of clients or their locations that are allowed to be further than D from any located EMS station.

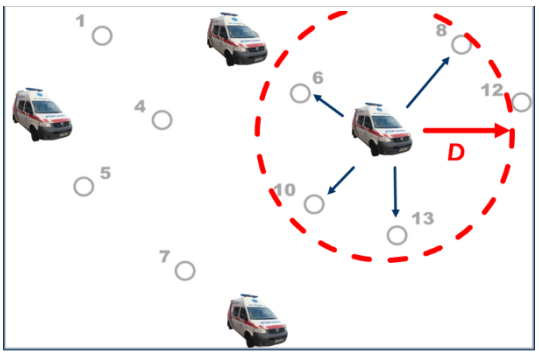


Fig. 2. Visualisation of fairness approach parameters

To formulate mentioned fairness rule by means of mathematical expressions, a set of binary variables z_j must be introduced. The variable $z_j \in \{0, 1\}$ is defined for each clients' location $j \in J$. It takes the value of zero, if the clients' location j is covered by service within the radius D from the nearest EMS station and it is equal to one in the opposite case. As it can be seen in Fig. 2, the variable z_{12} will take the value of one and the variables z_6, z_8, z_{10} and z_{13} will take the value of zero. The mathematical expression can take the form of (10).

$$z_j + \sum_{\substack{i \in I \\ d_{ij} \le D}} y_i \ge 1 \quad \text{for } j \in J \quad (10)$$

The number of clients, whose distance from the nearest stations exceeds D, can be limited by the expression (11). We

have to note, that the parameter B is given in percentage, i.e. it takes the real value from the interval [0, 1].

$$\sum_{j \in J} z_j b_j \leq B \sum_{j \in J} b_j \tag{11}$$

If we wanted to apply fairness demands not so strongly for individual clients, but only for clients' locations regardless the number of clients sharing the locations, we could replace the constraint (11) by a simpler expression (12).

$$\sum_{j \in J} z_j \leq B|J| \tag{12}$$

Based on these two different fairness constraints (11) and (12), we will verify both approaches. We will study the impact of fairness constraints with parameters D and B on average service accessibility first for individual clients (here we will solve the model (5)-(9) with additional constraints (10) and (11)) and then for clients' locations (model (5)-(9) with constraints (10) and (12)). In both cases, we will also add the obligatory constraint (13) for variables z_j.

$$z_j \in \{0, 1\} \quad for\ j \in J \tag{13}$$

If we extend the former model (5)-(9) by additional fairness constraints, then the original set of feasible solutions shrinks. Consequently, the optimal solution of the extended model may report a higher objective function value and thus, the average service accessibility for clients may get worse. To evaluate the difference, we can introduce a special coefficient in the following way.

Let the symbol $\mathbf{y}^{st}$ denote the resulting vector of location variables y_i of the standard model (5)-(9) corresponding to its optimal solution. Analogically, let the symbol $\mathbf{y}^{fair}$ denote the vector obtained as a result of solving the extended model with fairness criteria. For any vector of location variables, the objective function value can be computed according to the formula (2). Then, the coefficient called *Price of Fairness* (*PoF*) can be evaluated in accordance to (14) to measure the relative service accessibility worsening [2, 20].

$$PoF = 100 * \frac{f\left(\mathbf{y}^{fair}\right) - f\left(\mathbf{y}^{st}\right)}{f\left(\mathbf{y}^{st}\right)} \tag{14}$$

C. Infeasible fairness constraints – what to do?

Parameters D and B introduced in previous subsection significantly affect the resulting solution, because their values restrict the set of feasible solutions. If they are set in a too strong way (very small values), the associated fairness constraints can make the problem infeasible. A simple example can be taken from Table I, in which current EMS stations deployment is reported and discussed. Let us consider the benchmark of Žilina (ZA). The average distance from clients to the nearest located EMS station is 4.62 and the maximal relevant distance is 24. If we solved the model (5)-(9) with additional fairness constraints (10) and (11) for $D = 6$ and $B = 0.05$, i.e. at most five percent of all clients could have their nearest EMS station further than six, we would not obtain any feasible solution. Such fairness criteria are not possible to be fulfilled with given number of EMS stations. In such a case, the following question arises: What is the minimal number of facilities, which could meet the fairness requirements?

The minimal number p_{min} of stations, which is sufficient to meet the fairness restrictions, can be obtained as the resulting objective function value of the following model (15)-(19). The model follows the set-covering approach and does not require any further input data except for those, which were introduced in the above model (5)-(9).

$$Minimize \qquad p_{min} = \sum_{i \in I} y_i \tag{15}$$

$$Subject\ to: \quad z_j + \sum_{\substack{i \in I \\ d_{ij} \leq D}} y_i \geq 1 \quad for\ j \in J \tag{16}$$

$$\sum_{j \in J} z_j b_j \leq B \sum_{j \in J} b_j \tag{17}$$

$$z_j \in \{0, 1\} \quad for\ j \in J \tag{18}$$

$$y_i \in \{0, 1\} \quad for\ i \in I \tag{19}$$

The minimized objective function (15) expresses the number of located EMS stations. Structural constraints (16) and (17) are taken from the explanation of fairness extension. Of course, the expression (12) can replace the constraint (17). The obligatory constraints (18) and (19) keep the domain of used decision variables.

Since the problem described by terms (15)-(19) minimizes only the number of located EMS stations and it does not take into account the average distance from clients to their nearest service sources, the model (5)-(9) with fairness constraints can be solved for p_{min} in order to obtain the optimal locations of the EMS stations.

IV. Case Study

The main content of this research paper consists in the following computational study, the aim of which consists in evaluation of the impact of fairness constraints on average service accessibility in emergency medical system.

The numerical experiments reported in this case study were performed in the optimization software FICO Xpress 7.3, 64-bit equipped with the Xpress Mosel 3.4.0 version. The experiments were run on a PC equipped with the Intel® Core™ i5 9300HF CPU 2.4 GHz processor and 16 GB RAM.

The benchmarks used in the numerical experiments were taken from the road network of Slovak self-governing regions, in which the EMS system is operated. The problem sizes and other characteristics of current EMS stations deployment are summarized in Table I, which was discussed in Section II. In all solved instances, the sets I and J are formed by all network nodes corresponding to villages and cities of the region. It means that each node represents both a candidate for locating an EMS station and a demand point with clients. The cardinalities of these sets are equal to the value of n reported in Table I.

To study the impact of fairness constraints on the computational time demands and on the average accessibility of service for clients, we need a referential solution of the weighted p-median problem. Therefore, the model (5)-(9) without any fairness constraints was solved first. The results of this computation are reported in Table II, which has the following structure: The column Region is used to identify the used problem instance. The computational time in seconds is reported in the column denoted by *CT*. The optimal solution of the model (5)-(9) may bring such a vector $\mathbf{y}^{opt}$ of location variables y_i that differs from the current stations deployment

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

described by the vector $\mathbf{y}^{cur}$. To evaluate the difference between these two vectors, the concept of Hamming distance can be used. The Hamming distance *HD* can be evaluated according to the expression (20).

$$HD\left(\mathbf{y}^{opt},\mathbf{y}^{cur}\right)=\sum_{i\in I}\left|y_i^{opt}-y_i^{cur}\right| \qquad (20)$$

The right part of Table II contains a short analysis of the obtained results. Let the symbol *MRD* denote the maximal relevant distance defined by (21).

$$MRD=\max\left\{\min\left\{d_{ij}:i\in I,y_i=1\right\}:j\in J\right\} \qquad (21)$$

The value of objective function (2) computed for the resulting vector $\mathbf{y}^{opt}$ is reported in the column denoted by *minSum*. Finally, let the symbol *AvgDist* denote the average distance from clients to their nearest located EMS stations.

TABLE II. RESULTS OF THE WEIGHTED P-MEDIAN PROBLEM WITHOUT ANY FAIRNESS CONSTRAINTS

Region	*CT*	*HD*	*MRD*	*minSum*	*AvgDist*
BA	0.14	12	25	19325	3.19
BB	16.35	20	26	29873	4.52
KE	12.94	32	25	31200	3.93
NR	5.29	26	24	34041	4.93
PO	39.33	24	42	39073	4.77
TN	3.25	14	30	25099	4.22
TT	2.58	18	24	28206	5.07
ZA	3.65	24	26	28967	4.19

A cursory glance at the results shows that the radial formulation of the weighted *p*-median problem can be solved in a very short time (see the column *CT*).

As far as the average service accessibility for clients is concerned, the model (5)-(9) brings better values that could be observed for current state in Table I. Let us consider the benchmark Žilina (ZA). Current average distance is 4.62 and the model (5)-(9) reached its reduction to 4.19.

From the viewpoint of fairness, the weighted *p*-median problem (5)-(9) worsened the current situation. Even if the average distance got better, the maximal relevant distance has risen from 24 to 26. The reasons that led us to formulate additional fairness constraints, followed from the distance distribution in the optimal solution. Let´s look at the problem solved for the self-governing region of Žilina. The distance distribution is reported in the following Table III.

TABLE III. DISTANCE DISTRIBUTION IN THE OPTIMAL SOLUTION OF THE WEIGHTED P-MEDIAN PROBLEM FOR THE REGION OF ŽILINA

Distance	EMS stations	Percentage of stations	Clients	Percentage of clients
>6	149	47.30	1520	21.99
>7	121	38.41	1170	16.93
>8	96	30.48	889	12.86
>9	70	22.22	598	8.65
>10	52	16.51	382	5.53
>11	39	12.38	265	3.83
>12	29	9.21	177	2.56
>13	24	7.62	124	1.79
>14	17	5.40	70	1.01
>15	13	4.13	50	0.72
>16	9	2.86	34	0.49

The distance distribution of the resulting system design for the self-governing region of Žilina reported in Table III is interesting from different points of view. We know that the average distance from clients to their nearest located EMS stations is 4.19. As we can observe, there are more than 8 percent of clients, whose distance to the nearest service provider exceeds twice the average value. Analogically, there are more than 2.5 percent of clients, who are further than three times the average distance. Such situation may be legitimately considered unfair.

If we look at the clients' locations, the situation seems even worse. We can see more than 22 percent of network nodes, the distance of which from the nearest station exceeds twice the average value and there are still more than 9 percent of clients' locations further than three times the average value.

Other self-governing regions of Slovakia achieved similar results; therefore, they are not reported in details.

This simple analysis confirms the necessity of extending the original model (5)-(9) by additional fairness constraints. Furthermore, the distance distribution can help us to set up suitable values of fairness parameters *D* and *B*. Since the suggested fairness criteria may be formulated either for clients (by constraints (10), (11) and (13)) or for clients' locations regardless the number of clients inside (by constraints (10), (12) and (13)), the impact of fairness criteria on average service accessibility will be studied for both cases separately.

A. Fairness results for individual clients

The following Table IV contains the results of experiments for different settings of parameters *D* and *B* for the benchmark of Žilina. The first two columns are designed to specify the parameters. The symbol *CT* denotes the computational time in seconds. The resulting objective function value can be found in the column denoted by *minSum*. The denotation *AvgDist* is devoted to the average distance. Furthermore, the value of Hamming distance *HD* of the obtained result from the optimal solution of the model (5)-(9) is reported. Finally, we evaluated the price of fairness *PoF* according to the expression (14).

TABLE IV. RESULTS OF NUMERICAL EXPERIMENTS WITH FAIRNESS CRITERIA FOR CLIENTS - SELF-GOVERNING REGION OF ŽILINA

D	*B*	*CT*	*minSum*	*AvgDist*	*HD*	*PoF*
6	≤0.18	*Problem infeasible*				
6	0.19	6.35	32032	4.63	26	9.57
6	0.20	4.98	29542	4.27	12	1.95
6	0.21	4.43	29045	4.20	4	0.27
6	0.22	3.76	28967	4.19	0	0.00
7	≤0.13	*Problem infeasible*				
7	0.14	6.36	32893	4.76	28	11.94
7	0.15	3.98	29477	4.27	14	1.73
7	0.16	5.29	29043	4.20	6	0.26
7	0.17	3.75	28967	4.19	0	0.00
8	≤0.10	*Problem infeasible*				
8	0.11	5.69	30451	4.41	22	4.87
8	0.12	3.69	29076	4.21	10	0.37
8	0.13	3.70	28967	4.19	0	0.00
9	≤0.06	*Problem infeasible*				
9	0.07	4.75	30457	4.41	14	4.89
9	0.08	6.32	29102	4.21	4	0.46
9	0.09	3.81	28967	4.19	2	0.00
10	≤0.03	*Problem infeasible*				
10	0.04	9.16	30691	4.44	16	5.62
10	0.05	3.64	28967	4.19	2	0.00

The results reported in Table IV confirm the following facts: There are many combinations of parameters *D* and *B*, which make the problem described by the model (5)-(9) with fairness extension by constraints (10), (11) and (13) infeasible. Such situations will be discussed later.

Let us focus now on the computational time reported in the column denoted by *CT*. As it can be observed, the extended problem with additional fairness constraints is solved in a higher computational time than the original problem described by (5)-(9). The average time of the extended problem solving is 4.98 seconds and the computational time of the former model (5)-(9) was 3.65 seconds (see the last rows of Table II). The increment of average computational time demands can be explained by a simple fact that the extended problem contains additional variables z_j introduced for each $j \in J$ and the number of structural constraints grows as well. On the other hand, the increase is not too high to affect the solvability of the problem in a considerably negative way. Thus, we can conclude, that the fairness constraints do not have a significant impact on the solving process time requirements.

As far as the resulting EMS system design is concerned, we can observe that if we set up the parameters *D* and *B* in a strong way (if they take small values), the average access to service for individual clients worsens and the price of fairness grows. In some cases, the price of fairness (*PoF*) can be so high that such resulting solution would not be acceptable for authorities responsible for EMS service management. Another trend consists in the fact that finer settings of parameters in the fairness constraints (10) and (12) lead to such a result that approaches the optimal solution of the weighted *p*-median problem formulated by the expressions (5)-(9), what can be demonstrated by decreasing value of Hamming distance *HD* and also by the declining value of *PoF*.

Similar findings can result from the following Table V, in which the EMS system design for the self-governing region of Banská Bystrica (BB) is studied.

TABLE V. RESULTS OF EXPERIMENTS WITH FAIRNESS CRITERIA FOR CLIENTS - SELF-GOVERNING REGION OF BANSKÁ BYSTRICA

D	*B*	*CT*	*minSum*	*AvgDist*	*HD*	*PoF*
6	≤0.21	*Problem infeasible*				
6	0.22	18.93	32768	4.96	28	8.83
6	0.23	21.90	30497	4.61	18	2.05
6	0.24	25.52	29941	4.53	2	0.23
7	≤0.17	*Problem infeasible*				
7	0.18	17.33	32046	4.85	26	6.78
7	0.19	28.83	30643	4.64	18	2.51
7	0.20	15.14	30119	4.56	6	0.82
7	0.21	16.81	29907	4.53	2	0.11
7	0.22	15.41	29873	4.52	0	0.00
8	≤0.13	*Problem infeasible*				
8	0.14	29.65	34079	5.16	36	12.34
8	0.15	61.78	31937	4.83	28	6.46
8	0.16	37.79	30842	4.67	24	3.14
8	0.17	15.56	30096	4.55	12	0.74
8	0.18	15.94	29888	4.52	2	0.05
8	0.19	14.91	29873	4.52	0	0.00
9	≤0.09	*Problem infeasible*				
9	0.10	29.36	32248	4.88	28	7.36
9	0.11	17.75	30447	4.61	16	1.89
9	0.12	37.03	30002	4.54	8	0.43
9	0.13	18.28	29876	4.52	2	0.01
9	0.14	14.98	29873	4.52	0	0.00

Since the results obtained for other self-governing regions show very similar trends, we do not report all of them in this computational study.

Let us return now to those situations, in which the fairness constraints cause infeasibility of the problem. In such cases, the values of parameters *D* and *B* are too strict. It means, that current number *p* of located stations is too small to fulfill the criterion of fairness and a higher number of EMS stations need to be located. Therefore, the covering model (15)-(19) was suggested to get the minimal value p_{min}, for which a feasible solution of the problem (5)-(9) with fairness criteria exists. After obtaining the value of p_{min}, the original model (5)-(9) with fairness criteria can be solved to optimize the service accessibility for clients. The following Table VI contains the results of numerical experiments for the self-governing region of Žilina (ZA). The table is divided into three parts. The left part reports on the values of fairness parameters *D* and *B*. The middle part contains the values of p_{min} and the computational time *CT* in seconds. If the reported value of *CT* is 0.00, then the real computational time of the covering model (15)-(19) was lower than 0.01 second. It is generally known that the covering models are easily solvable. The right part of the table contains the results of the problem (5)-(11) and (13).

TABLE VI. RESULTS OF NUMERICAL EXPERIMENTS WITH ORIGINAL INFEASIBLY STRICT FAIRNESS CRITERIA - REGION OF ŽILINA

D	*B*	model (15)-(19)		model (5)-(11), (13)		
		p_{min}	*CT*	*CT*	*minSum*	*AvgDist*
6	0.01	75	0.01	3.74	20144	2.91
6	0.03	62	0.00	5.65	21899	3.17
6	0.05	55	0.01	7.88	22665	3.28
6	0.07	49	0.01	3.81	23833	3.45
6	0.09	45	0.00	6.40	24177	3.50
6	0.11	41	0.01	4.56	25625	3.71
6	0.13	38	0.00	6.21	26484	3.83
6	0.15	35	0.01	4.40	27595	3.99
6	0.17	32	0.01	4.67	29463	4.26
6	0.19	29	0.01	6.46	32032	4.63
7	0.01	62	0.01	6.98	23499	3.40
7	0.03	51	0.00	6.51	25384	3.67
7	0.05	45	0.01	4.93	25274	3.66
7	0.07	40	0.01	5.79	28095	4.07
7	0.09	37	0.00	7.20	27599	3.99
7	0.11	33	0.01	6.04	31812	4.60
7	0.13	31	0.01	4.38	29406	4.25
7	0.15	28	0.01	6.58	32551	4.71
8	0.01	52	0.01	10.17	25652	3.71
8	0.03	44	0.01	6.75	26224	3.79
8	0.05	38	0.01	3.90	28966	4.19
8	0.07	34	0.02	8.48	31788	4.60
8	0.09	31	0.01	6.79	31977	4.63
8	0.11	28	0.01	7.23	34177	4.95
9	0.01	44	0.01	4.47	30364	4.39
9	0.03	37	0.02	8.16	29829	4.32
9	0.05	32	0.04	7.50	31159	4.51
9	0.07	29	0.02	4.84	30457	4.41
10	0.01	37	0.01	3.78	27167	3.93
10	0.03	31	0.02	8.32	29927	4.33
10	0.05	27	0.01	9.34	32600	4.72

The achieved results reported in Table VI show that the stronger the fairness criteria are (the lower is the values of *D* and *B*), the higher is the minimal number p_{min} of EMS stations that are needed to fulfill the fairness requirements. The next

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

observation consists in the fact that the model (15)-(19) for obtaining the value of p_{min} is easily solvable thanks to a simple structure of the covering model. Another trend resulting from the experiments concerns the average accessibility of service. The softer the fairness parameters D and B are (the lower is the value of p_{min}), the higher is the average distance from individual clients to their nearest located EMS stations.

The performed numerical experiments for other self-governing regions of Slovakia showed similar results as reported for the benchmark Žilina. That is why we do not consider it necessary to report all of them in separate tables.

B. *Fairness results for clients' locations*

This subsection is devoted to the results of numerical experiments, in which the fairness criterion was defined not for individual clients, but only for clients' locations without the knowledge of number of clients inside. Mathematically spoken, the constraint (11) in the model (5)-(9) extension is replaced by the expression (12). Since the nature of numerical experiments is the same, we will report just the results themselves without any further discussion.

The following Table VII contains the results of performed experiments for different settings of parameters D and B for the benchmark of Žilina.

TABLE VII. RESULTS OF EXPERIMENTS WITH FAIRNESS CRITERIA FOR CLIENTS' LOCATIONS - SELF-GOVERNING REGION OF ŽILINA

D	*B*	*CT*	*minSum*	*AvgDist*	*HD*	*PoF*
6	≤0.325	*Problem infeasible*				
6	0.350	6.49	33617	4.86	32	13.83
6	0.375	4.75	31195	4.51	26	7.14
6	0.400	4.65	30137	4.36	22	3.88
6	0.425	4.14	29547	4.28	12	1.96
6	0.450	3.65	29078	4.21	6	0.38
6	0.475	3.66	28967	4.19	0	0.00
7	≤0.225	*Problem infeasible*				
7	0.250	4.82	38348	5.55	34	24.46
7	0.275	3.97	31832	4.61	22	9.00
7	0.300	8.03	30342	4.39	14	4.53
7	0.325	7.74	29668	4.29	10	2.36
7	0.350	6.70	29139	4.22	8	0.59
7	0.375	3.83	28967	4.19	2	0.00
8	≤0.150	*Problem infeasible*				
8	0.175	3.86	37161	5.38	38	22.05
8	0.200	3.86	31592	4.57	30	8.31
8	0.225	3.78	30452	4.41	26	4.88
8	0.250	5.86	29577	4.28	20	2.06
8	0.275	4.25	29044	4.20	6	0.27
8	0.300	3.75	28967	4.19	2	0.00
9	≤0.125	*Problem infeasible*				
9	0.150	3.68	30897	4.47	16	6.25
9	0.175	4.45	29803	4.31	12	2.81
9	0.200	4.37	29133	4.22	4	0.57
9	0.225	3.66	28967	4.19	0	0.00
10	≤0.075	*Problem infeasible*				
10	0.100	3.95	30176	4.37	14	4.01
10	0.125	7.34	29487	4.27	8	1.76
10	0.150	3.75	28978	4.19	4	0.04
10	0.175	3.77	28967	4.19	0	0.00

The last Table VIII summarizes the study, in which the value of p_{min} was obtained first, and then, the weighted p-median problem with fairness extension was solved.

TABLE VIII. RESULTS OF EXPERIMENTS WITH ORIGINAL INFEASIBLY STRICT FAIRNESS CRITERIA - REGION OF ŽILINA

D	*B*	model (15)-(19)		model (5)-(10), (12), (13)		
		p_{min}	*CT*	*CT*	*minSum*	*AvgDist*
6	0.025	82	0.00	3.82	20664	2.99
6	0.050	74	0.01	3.89	21394	3.10
6	0.075	66	0.01	4.88	22900	3.31
6	0.100	58	0.01	4.15	25717	3.72
6	0.125	54	0.01	4.23	25103	3.63
6	0.150	50	0.01	5.46	25995	3.76
6	0.175	46	0.00	4.26	27331	3.95
6	0.200	42	0.02	3.66	30052	4.35
6	0.225	40	0.01	3.77	28706	4.15
6	0.250	37	0.05	3.82	30334	4.39
6	0.275	34	0.02	3.93	33763	4.89
6	0.300	32	0.02	3.95	33504	4.85
6	0.325	30	0.01	4.63	35174	5.09
7	0.025	70	0.01	3.86	22054	3.19
7	0.050	62	0.01	3.89	22860	3.31
7	0.075	54	0.01	3.85	24834	3.59
7	0.100	48	0.02	4.69	26418	3.82
7	0.125	44	0.02	5.78	27494	3.98
7	0.150	40	0.02	7.18	29487	4.27
7	0.175	36	0.01	4.56	33259	4.81
7	0.200	34	0.01	9.55	32752	4.74
7	0.225	31	0.01	4.60	37602	5.44
7	0.250	29	0.02	4.93	38348	5.55
8	0.025	56	0.01	3.89	25320	3.66
8	0.050	48	0.00	4.51	27849	4.03
8	0.075	42	0.00	4.26	29443	4.26
8	0.100	38	0.01	4.42	30620	4.43
8	0.125	34	0.01	4.04	33598	4.86
8	0.150	32	0.00	4.45	33003	4.78
8	0.175	29	0.01	3.98	37161	5.38
9	0.025	49	0.01	3.51	26344	3.81
9	0.050	41	0.01	4.33	29145	4.22
9	0.075	37	0.01	5.27	29123	4.21
9	0.100	33	0.02	4.77	32646	4.72
9	0.125	30	0.02	5.11	35419	5.13
9	0.150	28	0.02	3.94	32421	4.69
10	0.025	39	0.01	3.67	26834	3.88
10	0.050	33	0.02	3.71	29409	4.26
10	0.075	30	0.04	3.96	30870	4.47
10	0.100	27	0.01	4.13	34072	4.93

As before, we can observe the same trends as reported in the previous subsection. Other experiments performed with the benchmarks corresponding to other self-governing regions of Slovakia confirmed our previous findings and therefore, we do not report the complete results set.

V. CONCLUSIONS

This contribution was aimed at application of the Operations research knowledge into the healthcare segment. Special attention was paid to the Emergency Medical Service system in Slovakia and improving its accessibility for clients. From the viewpoint of clients, the most commonly used optimization criterion consists in minimizing the average distance from clients to the nearest source of provided service. Thus, the model of the weighted p-median problem finds its application in service system designing.

This paper was focused on additional fairness constraints, which take into account those clients, which are far from any located service station. Presented computational study was

aimed at investigation how these fairness constraints affect the computational demands and how much they may affect the average service accessibility for clients. In the case study, we have analyzed and discussed the obtained results.

We have shown, that the additional fairness constraints do not significantly affect the computational time, even if they cause its little increase. It must be noted that the model extension raises the size of the solved problem and small computational time increase is natural. On the other hand, we have shown that the fairness parameters directly impact the accessibility of the service for clients. If they are set up in a too strict way, the problem may get infeasible. For such cases we have suggested a simple covering model, which can be used to obtain the minimal number of stations to fulfill the fairness demands.

Future research in this field could be possibly aimed at incorporating fairness rules into more complicated problems, which take into account more than one centers providing the service to clients or to fair optimization of emergency system under uncertainty.

ACKNOWLEDGMENT

This work was supported by the research grants VEGA 1/0689/19 “Optimal design and economically efficient charging infrastructure deployment for electric buses in public transportation of smart cities”, VEGA 1/0216/21 “Design of emergency systems with conflicting criteria using artificial intelligence tools”. This paper was supported by the Slovak Research and Development Agency under the Contract no. APVV-19-0441.

REFERENCES

[1] Avella, P., Sassano, A. and Vasil'ev, I. (2007). Computational study of large scale p-median problems. *Mathematical Programming* 109, pp. 89-114.

[2] Bertsimas, D., Farias, V. F., Trichakis, N. (2011). *The Price of Fairness*. In Operations Research, 59, 2011, pp. 17-31.

[3] Brotcorne, L, Laporte, G, Semet, F. (2003). Ambulance location and relocation models. European Journal of Operational Research, 147, pp. 451-463.

[4] Buzna, Ľ., Kohání, M., Janáček, J. (2013). Proportionally Fairer Public Service Systems Design. In: Communications - Scientific Letters of the University of Žilina 15(1), pp. 14-18.

[5] Current, J., Daskin, M. and Schilling, D. (2002). Discrete network location models, Drezner Z. et al. (ed) Facility location: Applications and theory, Springer, pp. 81-118.

[6] Doerner, K. F., Gutjahr, W. J., Hartl, R. F., Karall, M. and Reimann, M. (2005). Heuristic Solution of an Extended Double-Coverage Ambulance Location Problem for Austria. *Central European Journal of Operations Research*, 13(4), pp. 325-340.

[7] Elloumi, S., Labbé, M. and Pochet, Y. (2004). A new formulation and resolution method for the p-center problem. *INFORMS Journal on Computing*, 16(1), pp. 84-94.

[8] García, S., Labbé, M. and Marín, A. (2011). Solving large p-median problems with a radius formulation. *INFORMS Journal on Computing*, 23(4), pp. 546-556.

[9] Chanta, S., Mayorga, M. E., McLay, L. A. (2014). Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. Annals of Operations Research, 221, pp. 133-159.

[10] Ingolfsson, A., Budge, S., Erkut, E. (2008). Optimal ambulance location with random delays and travel times. Health care management science, 11(3), pp. 262-274.

[11] Janáček, J. (2008). Approximate Covering Models of Location Problems. Lecture Notes in Management Science: Proceedings of the 1st International Conference ICAOR, Yerevan, Armenia, pp. 53-61.

[12] Janáček, J., Kvet, M. (2014). Lexicographic optimal public service system design. In Mathematical methods in economics, Olomouc, Czech Republic, September 10-12, 2014, Olomouc: Palacký university, 2014, ISBN 978-80-244-4209-9, pp. 366-371.

[13] Janáček, J., Kvet, M. (2016). Sequential approximate approach to the p-median problem. In Computers & industrial engineering, Vol. 94, ISSN 0360-8352, 2016, pp. 83-92.

[14] Janáček, J. and Kvet, M. (2016). Min-max Optimization and the Radial Approach to the Public Service System Design with Generalized Utility. In *Croatian Operational Research Review*, Vol. 7, Num. 1, pp. 49-61.

[15] Jánošíková, Ľ. (2007). Emergency Medical Service Planning. In: Communications - Scientific Letters of the University of Žilina 9(2), pp. 64-68.

[16] Jánošíková, Ľ. and Žarnay, M. (2014). Location of emergency stations as the capacitated p-median problem. In: Quantitative Methods in Economics (Multiple Criteria Decision Making XVII). pp. 117-123.

[17] Karatas, M. and Yakıcıa, E. (2019). An analysis of p-median location problem: Effects of backup service level and demand assignment policy. European Journal of Operational Research, 272(1), pp. 207-218.

[18] Kvet, M. (2014). Computational Study of Radial Approach to Public Service System Design with Generalized Utility. In: *Proceedings of International Conference Digital Technologies 2014*, Žilina, Slovakia, pp. 198-208.

[19] Kvet, M. (2015). Advanced Radial Approach to Resource Location Problems. Studies in Computational Intelligence: Developments and Advances in Intelligent Systems and Applications, Springer, pp. 29-48.

[20] Kvet, M., Janáček, J. (2014). *Price of fairness in public service system design*. In Mathematical methods in economics = MME 2014: 32nd international conference: Olomouc, Czech Republic, September 10-12, 2014, Olomouc: Palacký university, 2014, ISBN 978-80-244-4209-9, pp. 554-559.

[21] Kvet, M., Janáček, J. (2017). Composed min-max and min-sum radial approach to the emergency system design. In Operations research proceedings 2015: selected papers of the international conference of the German, Austrian and Swiss operations research society (GOR, ÖGOR, SVOR/ASRO), University of Vienna, Austria, September 1-4, 2015, Springer, 2017, ISBN 978-3-319-42901-4, ISSN 0721-5924, pp. 41-47.

[22] Marianov, V. and Serra, D. (2002). Location problems in the public sector, *Facility location - Applications and theory* (Z. Drezner ed.), Berlin, Springer, pp 119-150.

[23] Ogryczak, W., Sliwinski, T. (2006): *On Direct Methods for Lexicographic Min-Max Optimization*. In Gavrilova M. et al. (Eds.): ICCSA 2006, LNCS 3982, Berlin: Heidelberg: Springer, 2006, pp. 802-811.

[24] Reuter-Oppermann, M., van den Berg, P. L., Vile, J. L. (2017).: Logistics for Emergency Medical Service systems, Health Systems, Vol. 6, No 3, pp. 187-208.

[25] Schneeberger, K. et al. (2016). Ambulance location and relocation models in a crisis. Central European Journal of Operations Research, Vol. 24, No. 1, Springer, pp. 1-27.

[26] Snyder, L. V. and Daskin, M. S. (2005). Reliability models for facility location; The expected failure cost case, *Transport Science* 39 (3), pp. 400-416.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Customization of Uniformly Deployed Set Kit for Path-relinking Method

Jaroslav Janáček, Marek Kvet
University of Žilina, Faculty of Management Science and Informatics
Univerzitná 8215 / 1
010 26 Žilina, Slovakia
{jaroslav.janacek, marek.kvet}@fri.uniza.sk

***Abstract*—The weighted *p*-median problem solving techniques represent basic tools for designing large emergency service systems, which have to provide public of a serviced region with service from a given number of service centers. A specific form of the set of all feasible solutions of the *p*-median problem enables to employ efficient incrementing heuristics to obtain a near-to-optimal solution. The incrementing heuristics proved their efficiency in combination with preliminary inspection of the set of all feasible solutions performed by a uniformly deployed set of *p*-median problem solutions. As obtaining uniformly deployed set for a specific *p*-median problem is very time consuming, an idea of a universal kit of uniformly deployed set has arisen. The idea consists in building up a standard family of uniformly deployed sets for given ranges of the number *p* of located centers and the number *m* of possible service center locations. If an emergency system has to be designed and its sizes *p* and *m* do not correspond with any standard uniformly deployed set, then a suitable standard set of the kit is adjusted to the sizes of the solved problem. This approach works excellently if the neighborhood search incrementing heuristics are applied, but it fails in case of path-relinking method based incrementing algorithms. This defect is caused by the fact that the solutions contained in the adjusted uniformly deployed set do not cover all possible service center locations of the solved problem. In this paper, we suggest an extending adjusted method overcoming the above-mentioned drawback and we study the impact of this improvement on efficiency of incrementing approaches based on the path-relinking inspection.**

Keywords—generalized weighted p-median problem, path-relinking method, uniformly deployed set

I. Introduction

Operations research belongs to the scientific disciplines that find their applications in a wide spectrum of fields covering different areas of real life [14]. Its main contribution consists in the application of advanced analytical methods, which can be used to make better decisions. Sometimes, Operations research is considered a sub-field of mathematical sciences. Besides that, Operations research covers many areas including mathematical modelling, computer simulation, queueing theory, etc. Those are scientific subjects, which can considerably help in the decision-making process. It must be realized that mathematical models usually describe the solved problems. Since almost all optimization techniques and other solving approaches require advanced knowledge of software development and programming, the Operations research field is closely related also to Applied Informatics.

Out of all mentioned study fields and large scale of solved problems, we concentrate only on the locations problems based on the weighted *p*-median problem formulation. Here, the medical segment constitutes the core application area for the results of our research. The emphasis is put on the pre-hospital urgent healthcare [2, 4, 13, 16, 17].

As the optimization problems are widely used in various fields, methods of their fast solution have become a key area of interest for many IT professionals and operations researchers. The available solving approaches can be divided into two groups.

The first one consists of the exact methods [1, 6, 7, 9, 20]. Some of them are based on so-called radial formulation of the problem, which makes the solving process significantly faster [7, 15]. The main disadvantage of the exact methods lies in the fact that the resulting solution is found quickly, but then, large computational time is spent by verification of the solution optimality. Thus, these methods often fail with large problem instances. Furthermore, a next possible complication may occur, when more than one located centers are assumed to provide the associated service to system users. In such a case, the concept of so-called generalized disutility is applied [5, 15, 21] and possible temporal unavailability of a center is taken into account. Obviously, the mathematical model gets more complex and the exact methods are not able to solve it in acceptably short time.

Mentioned weaknesses can be overcome by the second group of solving approaches, which is formed by various approximate algorithms, advanced heuristics and different metaheuristic methods [8, 18, 19]. The difference from the exact methods consists in the fact that these approaches perform faster or their time is limited and they are supposed to bring as good solution as possible.

This contribution focuses on special kinds of the path-relinking method embedded into a metaheuristic called the discrete self-organizing migrating algorithm [3], which we try to combine with usage of the uniformly deployed set (UDS) of *p*-location problem solutions [10, 11]. Special attention is devoted to improve the proposed algorithm performance in the case, when a universal kit of uniformly deployed sets is used to obtain input population for the discrete self-organizing migrating algorithm. The idea of kit usage consists in building up a standard family of uniformly deployed sets for given ranges of the number *p* of located centers and the number *m* of possible service center locations. If a concrete instance of the *p*-location problem is to be solved and its sizes *p* and *m* do not correspond to any standard uniformly deployed set, then a suitable standard set of the kit is adjusted to the sizes of the solved problem. This approach works excellently if a neighborhood search incrementing heuristics are applied, but it usually fails in the case of path-relinking method based incrementing algorithms. This defect is caused by the fact that

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

the solutions contained in the adjusted UDS do not cover all possible service center locations of the solved problem. In this paper, we suggest an extending adjusted method overcoming this drawback and we study the impact of this improvement on efficiency of incrementing approaches based on path-relinking inspection.

II. The Generalized Weighted p-Median Problem and Uniformly Deployed Set of Solutions

The location problems, which we are interested in, are mostly based on the weighted *p*-median problem formulation. Generally, this problem can be understood so that it is necessary to find, for a given set of *n* demand points, the set of *p* supply points so that the given objective takes its optimal value. The standard quality criterion of system design is usually expressed by minimization of total distance from system users concentrated at the demand points to their nearest supply points. The supply points (service centers) can be chosen from *m* possible locations, which correspond to the nodes of an associated transportation network graph. Thus, the discrete problem formulation requires a list of candidate points. The system users, to whom the service is provided, are concentrated in the demand points. We assume that these users' locations are subscribed by integers from 1 to *n*. Each demand point $j = 1, 2 \dots n$ aggregates b_j users. The service providers can explain the coefficient b_j also as a frequency of demands for service or as the exact number of visits.

As the common *p*-median problem follows the assumption that only the nearest located facility can satisfy the user demand [1, 2, 7], the concept of so-called generalized disutility published in [9, 11, 14] represents an important model extension and it considerably expands the possibilities of possible results applications. Mentioned generalization considers *r* nearest facility locations involved in providing the service to each client. In such a case, the user is served by the *k*-th nearest supply point with probability q_k. Mathematical formulation of the problem uses the symbol d_{ij} to denote the distance from a candidate for locating the supply point at the location $i = 1, 2 \dots m$ to the demand point $j = 1, 2 \dots n$. Furthermore, let the operation $min_k\{\}$ denote the *k*-th minimal value of the list of values in the brackets. Based on these preliminaries, all input data of the mathematical model are completely discussed.

As far as the model variables are concerned, all the decisions must be covered. Let the decision on locating a supply point at the location *i* be modelled by a zero-one variable y_i for each $i=1, 2 \dots m$. The variable y_i gets the value of one if a supply point is located at the location *i* and it takes the value of zero otherwise. Then, the expression (1) can describe the problem.

$$\min\left\{\begin{array}{l}\displaystyle\sum_{j=1}^{n} b_j \sum_{k=1}^{r} q_k \min_k\left\{d_{ij} : i = 1,...,m,\ y_i = 1\right\}: \\ \mathbf{y} \in \{0,1\}^m, \displaystyle\sum_{i=1}^{m} y_i = p\end{array}\right\} \quad (1)$$

The problem (1) can be studied as a search in a sub-set of *m*-dimensional hypercube vertices, which have exactly *p* non-zero components each.

A topology on the set of feasible solutions of the problem can be introduced using so-called Hamming distance *H*, which is defined by (2) for a pair of feasible solutions **y** and **x**.

$$H(\mathbf{y},\mathbf{x}) = \sum_{i=1}^{m} |y_i - x_i| \quad (2)$$

The definition of Hamming distance *H* formulated by the expression (2) indicates that the distance of two solutions described by the vectors **y** and **x** can take only an even integer value from the interval $[0, 2p]$. Furthermore, the number of possible supply point locations, in which a service center is located in both solutions **y** and **x** can be evaluated by the expression $p\text{-}H(\mathbf{y}, \mathbf{x})/2$.

It must be realized that the choice of distance measure affects the complexity of the problem (1) as well as the approach needed to find a solution. In this research paper, we concentrate on such solving techniques that make use of the uniformly deployed set (UDS) of feasible solutions.

The uniformly deployed set *S* of feasible solutions can be formed by a subset of (3) such that the inequality $H(\mathbf{y}, \mathbf{x}) \geq h$ holds for each pair of vectors $\mathbf{y}, \mathbf{x} \in S$.

$$\left\{\mathbf{y} : y_i \in \{0, 1\}, \sum_{i=1}^{m} y_i = p\right\} \quad (3)$$

Having established a uniformly deployed set of big enough cardinality, we can obtain a partial "terrain mapping" by computing the optimization criterion for each element of the set. Mentioned approach enables us to identify those areas of interest, which deserve proper exploration. Furthermore, the UDS can represent a population with maximal diversity, what is a welcome property for our metaheuristic implementation.

III. Path-Relinking Method Based Optimization

To explain suggested optimization method based on usage of UDS of solutions, the original problem (1) needs to be reformulated in a little different way. Of course, the output of the model solving process still consists of the set of candidates, in which a supply point is to be located. The novelty of adjusted formulation lies in the fact that the solution does not necessarily need to be described by a vector **y** of location variables y_i for $i = 1, 2 \dots m$ as before. Obviously, the resulting system design can be formed by a list *P* of indexes *i*, which correspond to those possible supply points locations, in which a center is located. It means that the list *P* is formed by such indexes *i*, for which the variable y_i takes the value of one. From previous definition of the problem, it is clear, that the cardinality of the list *P* must equal to the value of *p*. Based on this assumption, the original problem (1) can be reformulated into the form of (4), in which the used objective $f(P)$ takes the form of (5).

$$\min\left\{f(P), P \subset \{1,2,\dots,m\}, |P| = p\right\} \quad (4)$$

$$f(P) = \sum_{j=1}^{n} b_j \sum_{k=1}^{r} q_k \min_k\left\{d_{ij} : i \in P\right\} \quad (5)$$

A simple version of the discrete self-organizing migrating algorithm with UDS usage in the form of an input set *S* can be described as follows.

Let us start from the assumption that the elements of the set $S=\{P^1, \dots, P^{|S|}\}$ are subscripted and ordered by permutation *O* so that $f(P^{O(1)}) \leq f(P^{O(2)}) \leq \dots \leq f(P^{O(|S|)})$ holds. The solutions are described by a list of indexes as it was defined above. After these preliminaries, the algorithm suggested for solving the problem (4)-(5) consists of these steps:

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

0. Initialize P^* by $P^{O(1)}$.
1. For k=2, 3 … $|S|$ perform subsequently P^*=*Path-Inspection*(P^*, $P^{O(k)}$).
2. Terminate and return P^*.

The Path-Inspection in the above algorithm is performed by function *Path-RelinkingMixed*(P, Q). The function returns the best-found solution P^* obtained by inspection of the shortest path connecting the input solutions P and Q. The used *Path-relinkingMixed* function performs according to following description.

Algorithm *Path-relinkingMixed*(P, Q)

0. Initialize P^* by *argmin*{$f(P)$, $f(Q)$}, determine the sets U and V so that U=P-Q and V=Q-P.
1. While the cardinality of the set U exceeds the value of one or the set V contains at least two elements, perform the following steps, otherwise terminate and return P^*.
2. If the cardinality of the set U is higher than one then perform:
   a. Determine locations c^* and d^* by (c^*, d^*)=*argmin*{$f((P-\{c\})\cup\{d\})$: $c\in U$, $d\in V$}.
   b. Update P, U, V, and P^* according to the following expressions:
      P=(P-{c^*})∪{d^*},
      U=U-{c^*},
      V=V-{d^*},
      P^*=*argmin*{$f(P)$, $f(P^*)$}.
3. If the cardinality of the set V exceeds the value of one then perform these operations:
   a. Determine locations c^* and d^* by (c^*, d^*)=*argmin*{$f((Q-\{c\})\cup\{d\})$: $c\in V$, $d\in U$}.
   b. Update Q, U, V, and P^* according to the formulas given below:
      Q=(Q-{c^*})∪{d^*},
      U=U-{d^*},
      V=V-{c^*},
      P^*=*argmin*{$f(Q)$, $f(P^*)$}.

In the above function, the operation *argmin*{$f(P)$, $f(Q)$} returns the solution P or Q, for which the quality criterion of particular system design takes lower value (objective function is assumed to be minimized). The operation *argmin*{$f((P-\{c\})\cup\{d\})$: $c\in U$, $d\in V$} returns such pair (c^*, d^*), for which the expression $f((P-\{c\})\cup\{d\})$ takes the lowest value.

In addition, two sets U and V are necessary to be introduced. Let the set U contain the selected locations of P, which are not contained in Q and analogically, let the set V contain the selected locations of Q, which are not contained in P. Briefly, U=P-Q and V=Q-P.

The function is suggested so that it tests only feasible solutions on the shortest path connecting P and Q.

IV. Kit Usage and Extension Method

The core idea of the UDS kit construction was inspired by the necessity of solving different location problems with their specific sizes and various quality criteria. The main advantage of the uniformly deployed sets consists in the fact that the sets of solutions are suitable for the whole spectrum of problems regardless the used objective. Let us introduce two ranges M and $P_{\underline{m}}$ of parameters $\underline{m}$ and $\underline{p}$ respectively. Recall that the parameter $\underline{m}$ expresses the number of candidates for supply point location and the parameter $\underline{p}$ is used to denote the number of service centers to be chosen from the set of candidates. These two parameters fully describe the size of any solved location problem. For each pair [$\underline{m}$, $\underline{p}$], where $\underline{m}\in M$ and $\underline{p}\in P_{\underline{m}}$, a set $S_{\underline{m}\underline{p}}$ can be constructed in such a way that the cardinality of this set exceeds given limit. This will ensure a sufficient number of feasible solutions for any problem. Then, the UDS kit is formed by the sets $S_{\underline{m}\underline{p}}$. Obviously, the kit can be prepared generally for arbitrary cases long time before some concrete task emerges.

Let us focus now on a concrete example of UDS kit usage. When a new p-location problem with m possible supply point locations occurs, then the combination of parameters m and p does not necessarily have to be covered by any element of the kit. In such a case, the maximal $\underline{m}\in M$ and the minimal $\underline{p}\in P_{\underline{m}}$ must be chosen in order to meet the inequalities $\underline{m} \le m$ and $p \le \underline{p}$. A uniformly deployed set for the p-location problem with m possible candidates for supply point locating can be obtained so that each $\underline{p}$-tuple of $S_{\underline{m}\underline{p}}$ is reduced by removing $\underline{p}$-p locations from the $\underline{p}$-tuple.

Contrary to an ordinary neighborhood search method, the *Path-relinking* method gave usually much worse results. This bad performance can be explained by the way of "kit reduced" set usage. These sets were obtained by reduction of solutions of $\underline{p}$-location problem defined for $\underline{m}$ possible locations and applied to solution of p-location problem defined for m possible locations, where $\underline{m} < m$. Then, the *Path-relinking* method had no chance to inspect solutions containing some of the m-$\underline{m}$ locations. To remove this defect of UDS kit usage, we propose an extension method, which incorporates the surplus possible service center locations into solutions of the chosen kit set.

The following extension algorithm assumes that the solutions of set S are represented by zero-one m-dimensional vectors $\boldsymbol{y}^k$ for k= 1, 2 … $|S|$, where last m-$\underline{m}$ components equal to zeros.

Extension(S, m, $\underline{m}$)

0. Compute Y_i for i=1, 2 … $\underline{m}$ according to (6);
   compute Q according to (7)
   and define Q' = min{m-$\underline{m}$, Q};
   set t=$\underline{m}$.
1. For each k=1, 2 … $|S|$ perform the following cycle
   for each i= 1, 2 … $\underline{m}$ perform the command:
   if (y_i^k=1) and (Y_i>1) and (t<$\underline{m}$+Q')
   then do y_i^k=0; y_t^k=1; Y_i=Y_i -1; t=t+1.

The expressions (6) and (7) mentioned in the algorithm take the following form.

$$Y_i = \sum_{k=1}^{|S|} y_i^k \tag{6}$$

$$Q = \sum_{i=1}^{m} \max\{0, Y_i\} \tag{7}$$

The above extension algorithm inserts Q' unused locations into solutions of the input uniformly deployed set unless to diminished the Hamming distance among the solutions.

V. Numerical Experiments

The main goal of performed computational study was to study impact of the extension method on efficiency of the suggested discrete self-organizing migrating algorithm (DSOMA) in combination with the kit of the uniformly deployed sets, when used for solving of the generalized weighted *p*-median problem. We compare results of the algorithm with and without extension applied to a uniformly deployed set of the kit. The approaches are denoted by "**DSOMA-Ext"** and "**DSOMA-Orig"** respectively.

To verify the efficiency of suggested path-relinking based algorithm in combination with the usage of the UDS kit, a computational study was performed. The aim of the experiments was to study suggested methods from the viewpoint of computational time demands and also from the point of the resulting solution accuracy. The problem instances were taken from our previous research reported in [10, 11]. The individual benchmarks come from the road network of Slovakia and they correspond to the self-governing regions. The sizes of the individual benchmarks are *m* and *p*. These basic benchmarks characteristics are reported in the left part of Table I. Since all solved problems assume more than one supply points that can contribute to satisfy the user demands, the generalized disutility objective function was computed for r=3. The coefficients q_k for k=1, 2 … r were set at: $q_1 = 77.063$, $q_2 = 16.476$ and $q_3 = 100 - q_1 - q_2$ according to [12], which reports the simulation of the Emergency Medical Service system in Slovakia.

The uniformly deployed sets derived by reduction of the kit sets are reported in section "**kit reduced**" in the following Table I. There are reported parameters $|S|$, $\underline{m}$, $\underline{p}$, $\underline{h}$ and the minimal Hamming distance *h* of the reduced sets.

TABLE I. Benchmarks derived from the self-governing regions of Slovakia

Region	*m*	*p*	**"kit reduced"**				
			$\|S\|$	$\underline{m}$	$\underline{p}$	$\underline{h}$	*h*
BB	515	36	206	500	40	72	64
KE	460	32	200	400	40	70	54
NR	350	27	202	300	30	52	46
PO	664	32	147	600	40	74	58
TN	276	21	202	200	30	48	30
TT	249	18	205	200	20	34	30
ZA	315	29	202	300	30	52	50

The main goal of this computational study is to compare both suggested approaches. To achieve this goal, a sufficient set of problems and uniformly deployed sets of solutions must be considered. To make the comparison relevant and robust enough, we followed from a very useful property of any UDS. The mentioned useful feature consists in the fact that any arbitrary permutation of $\underline{m}$ locations subscripts brings a new set with the same parameters. This way, we were able to obtain ten different sets for "**kit reduced**" cases for each solved problem instance.

An individual experiment reported in this computational study was organized so that both optimization approaches were applied to each UDS. To evaluate the quality of the obtained results, the optimal solutions were necessary for us to know. The optimal objective function values for all studied benchmarks were taken from previous research reported in [11]. The obtained results are summarized in Table II, which is divided into two separate sections denoted by "**DSOMA-Ext**" and "**DSOMA-Orig**". Each row of the table corresponds to one solved problem instance. Each section of Table II comprises columns denoted by F^{*avg}, *gap* and *CT* [s]. The column denoted by F^{*avg} reports the average objective function value of the resulting solutions obtained by the approach application to ten uniformly deployed sets. The column denoted by *gap* contains the value, which expresses a relative difference of the obtained result from the optimal solution. Even if the objective function value of the optimal solution is not reported here, its value was used to evaluate the *gap* in percentage of the optimal objective function value. These values were obtained by averaging of gaps for ten uniformly deployed sets. The average computational times in seconds are reported in the columns denoted by *CT* [s].

For completeness, let's add the information that the numerical experiments were run on a PC equipped with the Intel® Core™ i7 3610QM 2.3 GHz processor and 8 GB of RAM. The algorithms were implemented in the Java language making use of the NetBeans IDE 8.2 environment.

TABLE II. Results of numerical experiments for the self-governing regions of Slovakia obtained for "kit reduced" uniformly deployed sets.

Region	**"DSOMA-Ext"**			**"DSOMA-Orig"**		
	F^{*avg}	*gap*	*CT* [s]	F^{*avg}	*gap*	*CT* [s]
BB	44907	0.35	30.3	44923	0.38	30.1
KE	45733	0.32	17.6	46099	1.12	17.5
NR	48996	0.11	8.5	49986	2.14	8.3
PO	56936	0.41	2..8	60476	6.65	20.5
TN	35789	1.46	3.4	49260	39.65	3.2
TT	41432	0.23	2.0	44090	6.66	2.0
ZA	42140	0.07	8.7	42145	0.08	8.7

VI. Conclusions

The paper deals with the possible adjustment of a uniformly deployed set of the *p*-location problem solutions coming from a standard kit of uniformly deployed sets. A uniformly deployed set is used here as an input population for a discrete self-organizing migrating algorithm to find a good solution of the generalized weighted *p*-median problem. The algorithm is based on the path-relinking method usage. The presented research was motivated by a drawback, which appeared, when the path-relinking method was applied to reduced uniformly deployed sets without any further adjustment. We proposed and verified an extension procedure, which improves the family of the standard kit of the *p*-location problems so that the discrete self-organizing migrating algorithm reaches almost a near-to-optimal solution.

The biggest advantage of the uniformly deployed set kit consists in the fact that it can be constructed independently on the solved instances of the *p*-location problems. The kit allows its multiple usage, because the kit construction is not affected by any kind of objective used in the solved problems. The only information for a UDS creation consists in the number of candidates for a supply point and in the number of centers to be located. It follows that the UDS kit represents a very useful tool with a wide range of possible applications.

Further research in this area of the path-relinking method usage may be aimed at some adaptive modifications of the self-organizing migrating algorithm and at possible adaptively

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

controlled reduction of the input uniformly deployed set of solutions.

ACKNOWLEDGMENT

This work was supported by the research grants VEGA 1/0089/19 “Data analysis methods and decisions support tools for service systems supporting electric vehicles”, VEGA 1/0689/19 “Optimal design and economically efficient charging infrastructure deployment for electric buses in public transportation of smart cities”, and VEGA 1/0216/21 “Design of emergency systems with conflicting criteria using artificial intelligence tools”. This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-19-0441.

REFERENCES

[1] Avella, P., Sassano, A., Vasil'ev, I. (2007). Computational study of large scale p-median problems. Mathematical Programming 109, pp. 89-114.

[2] Current, J., Daskin, M., Schilling, D. (2002). Discrete network location models, Drezner Z. et al. (ed) Facility location: Applications and theory, Springer, pp. 81-118.

[3] Davendra, D., Zelinka, I. (2016). Self-Organizing Migrating Algorithm, Methodology and Implementation. Springer, Studies in Computational Intelligence, 289 p.

[4] Doerner, K. F., Gutjahr, W. J., Hartl, R. F., Karall, M., Reimann, M. (2005). Heuristic Solution of an Extended Double-Coverage Ambulance Location Problem for Austria. Central European Journal of Operations Research, 13(4), pp. 325-340.

[5] Drezner, T., Drezner, Z. (2007). The gravity p-median model. European Journal of Operational Research 179, pp. 1239-1251.

[6] Elloumi, S., Labbé, M., Pochet, Y. (2004). A new formulation and resolution method for the p-center problem. INFORMS Journal on Computing, 16(1), pp. 84-94.

[7] García, S., Labbé, M., Marín, A. (2011). Solving large p-median problems with a radius formulation. INFORMS Journal on Computing, 23(4), pp. 546-556.

[8] Gendreau, M., Potvin, J. (2010). Handbook of Metaheuristics, Springer Science & Business Media.

[9] Janáček, J., Kvet, M. (2016). Min-max Optimization and the Radial Approach to the Public Service System Design with Generalized Utility. In Croatian Operational Research Review, Vol. 7, Num. 1, pp. 49-61.

[10] Janáček, J., Kvet, M. (2019). Uniform Deployment of the p-Location Problem Solutions. Operations Research Proceedings 2019, Springer, in print.

[11] Janáček, J., Kvet, M. (2019). Usage of Uniformly Deployed Set for p-Location Min-Sum Problem with Generalized Disutility. In: SOR 2019 proceedings, pp. 494-499.

[12] Jankovič, P. (2016). Calculating Reduction Coefficients for Optimization of Emergency Service System Using Microscopic Simulation Model. In: 17th International Symposium on Computational Intelligence and Informatics, pp. 163-167.

[13] Jánošíková, Ľ., Žarnay, M. (2014). Location of emergency stations as the capacitated p-median problem. In: Quantitative Methods in Economics (Multiple Criteria Decision Making XVII). pp. 117-123.

[14] Kozel, P. Orlíková, L., Pomp, M., Michalcová, Š. (2018). Application of the p-median approach for a basic decomposition of a set of vertices to service vehicles routing design. In Mathematical Methods in Economnics 2018, Praha: MatfyzPress, 2018, pp. 252-257.

[15] Kvet, M. (2014). Computational Study of Radial Approach to Public Service System Design with Generalized Utility. In: Proceedings of International Conference Digital Technologies 2014, Žilina, Slovakia, pp. 198-208.

[16] Marianov, V., Serra, D. (2002). Location problems in the public sector, Facility location - Applications and theory (Z. Drezner ed.), Berlin, Springer, pp 119-150.

[17] Matiaško, K., Kvet, M. (2017). Medical data management. In: Informatics 2017: IEEE International Scientific Conference on Informatics, Danvers: Institute of Electrical and Electronics Engineers, pp. 253-257.

[18] Rybičková, A., Burketová, A., Mocková, D. (2016). Solution to the locating – routing problem using a genetic algorithm. In: SmaRTT Cities Symposiuum Prague (SCSP), pp. 1-6.

[19] Rybičková A., Mocková D., Teichmann D. (2019). Genetic Algorithm for the Continuous Location-Routing Problem, Neural Network World 29(3), pp. 173–187.

[20] Sayah, D., Irnich, S. (2016). A new compact formulation for the discrete p-dispersion problem. European Journal of Operational Research, 256(1), pp. 62-67.

[21] Snyder, L. V., Daskin, M. S. (2005). Reliability models for facility location; The expected failure cost case, Transport Science 39 (3), pp. 400-416.

Some theoretical backgrounds for reinforcement learning model of supply chain management under stochastic demand

Darya Filatova
EPHE CHART
4 Rue Ferrus
75014 Paris, France
Email: daria_filatova@interia.pl

Charles El-Nouty
LAGA
Université Paris XIII, Sorbonne Paris Cité
99 avenue J-B Clément
93430 Villetaneuse, France
Email: elnouty@math.univ-paris13.fr

Roman V. Fedorenko
Department of Research Degree
Samara State University of Economics
Ulitsa Sovetskoy Armii, 141
443090 Samara, Russian Federation
Email: fedorenko083@yandex.ru

***Abstract*—Integrated and collaborative decision-making systems have been increasingly employed by all companies involved in the supply chain. However, steadfast fact-based decision-making tools are still difficult to get as far as these require adapted methodology. This paper studies the sustainable management problem of a two-echelon supply chain in which a supplier has a limited production capacity to satisfy the quality and quantity requirements of a retailer under stochastic demand. We motivate the model selection, which can be used in reinforcement learning, as well as sustainability constraints for a one-product manufacturing system taking into account production, maintenance, quality, and inventory operations and explore the supplier's production behavior with a required quality level. Pursuing the profit maximization, formed as the difference between the sales revenue and the expenditures on the above-mentioned operations, we determine and characterize the optimal conditions for the supply chain management in the presence of the stochastic demand and quality requirements. Our study shows that the stochastic formulation of the optimal control problem and its solution allow taking into account the short-range and long-range dependences exhibited by consumers' behavior. In this context, we characterize the optimal policy for the supplier and the retailer. Finally, we provide managerial insights concerning the sustainable coordination conditions of the supply chain under and quality requirements.**

I. Introduction

Accelerated technological development allows providing consumers with more and more new products resulting in a fast change in consumer flavors, making it difficult to predict the market demand for selected products accurately. The rapid development of technologies and speed-information exchange oblige rummaging for new business models. Beyond the doubt, cost efficiency maximization, as well as all related individualistic actions of business entities, are insufficient. Nowadays, legal frameworks allow for successful collaboration among enterprises, so that the business model selection depends only on the partners and their intention on the information exchange. One highly considered model is the two-echelon supply chain, where each participant is aware of the customer's demands and shares information with a partner causing lowering the transaction-distribution costs as well as optimizing the volume of sales and investment ([1]). Despite its conceptual simplicity, the results of the collaboration are highly uncertain and dependent on many factors.

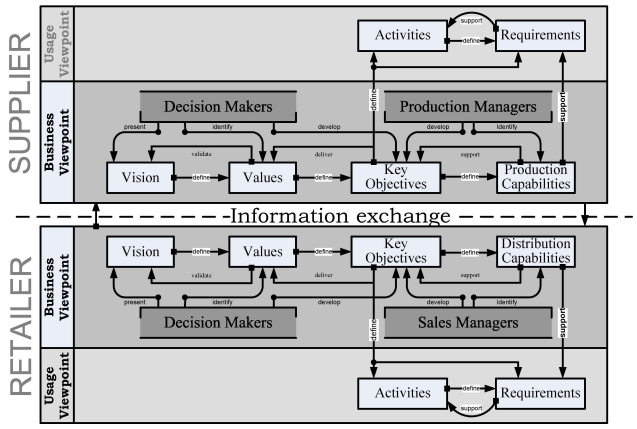


Fig. 1. The business motivation model

In general, to address this issue, one has in mind the partners' key objective defined by production and distribution capacities accorded with values and own visions. These last heavily depend on the competitive market situation, which determines not only volumes and selling prices, but also raw material costs, equipment maintenance, quality control, and inventory charges. Besides, the anticipation of the price and cost shocks requires a sustainable long-term production plan (strategic management or business viewpoint) and optimal coordination program (operation management or usage viewpoint) from the decision-makers (see Fig.1). Meanwhile, with increasing market competition, the required level of service is becoming one of the key factors that affect consumers' purchases. By guaranteeing a certain level of service quality, companies can maintain market share, gain a competitive advantage, and expand into new outlets. It is obvious to strive for the highest level of service to promote the sale of products. However, a higher level of service means higher expenses, which increases the risk of doing business. Therefore, how to set the appropriate level of service is crucial to balance customer satisfaction with expected profit.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Trying to overcome the aforementioned problems as well as the risks and the uncertainties including national and international regulations, pressures of stakeholders, customers and competitors, coercive social and environmental performances, the members of the supply chain are forced to use methods of sustainable management. Thinking about this kind of management as about "development that meets the needs of the present without compromising the ability of future generations to meet their needs" (see the Brundtland Commission – World Commission on Environment and Development, 1987, p. 8) is not enough. This conception is too broad and does not reflect the managerial needs of the supply chains. [2] presented deeper definition stating the sustainable management "as the strategic, transparent integration and achievement of organization's social, environmental, and economic goals in the systemic coordination of key interorganizational business processes for improving the long-term economic performance of the individual company and its supply chains". As it is possible to notice, this definition coincides with the ideas of the business motivation model (see Fig.1). Taking into account the impact of sustainability on environmental, social, and economic performance, [3] provided empirical evidences favoring mixed strategic orientations (sustainability and operations). Despite the progress made in this study, there is a room for the long-term strategic choice-performance issues ([4], [5], [6]).

We limit the object of our study to the two-echelon supply chain. This supply chain is a kind of a relationship between one supplier (or producer) and one retailer (or distributor), where the first one wholesales a product to the second one, who then retails the product to the final consumer. It is not difficult to imagine that Decision Makers as well as Production and Sales Managers want to maximize savings and to enhance profit for the whole supply chain while fixing the demand rate (this is a very common situation for the mass production). The agreement on the optimal production quantity and quality (for the supplier) and the optimal order quantity (for the retailer) has to established before the production circle. Various types of coordination mechanisms or the integrated and collaborative decision-making systems take into account quantity discount, credit option, return policies, quantity flexibility, and commitment of purchase quantity, cooperative advertising. Even assuming that the supplier and the retailer have complete information on each other's operations, each decision is made with limited available data due to uncertain market demand. The managerial problem can be formulated as a following question: "How to achieve key objectives under limited capacities and stochastic demand?" Therefore the primary aim of the paper is to determine the mechanism of sustainable supply chain management leading to the strategical goal. Moreover, we take into account that in the real-world situation the notion of "information" is related to Big Data. This last requires adapted solution mostly based on the methods of artificial intelligence. Hence, the reinforcement learning model could be an optimal choice. The understanding of its theoretic properties would be helpful for its computer implementation.

The rest of the paper is organized in the following manner. First of all by means of system analysis, we indicate the determinants of sustainable business management (Section II). In Section III, we motivate a selection of a mathematical model of the sustainable supply chain management taking into account stochastic market demand and guaranteed quality requirement such that the financial key objectives of the supply chain could be reached. The first-order optimality coordination conditions for the sustainable supply chain management are presented in Section IV.

II. Concepts of sustainable management

Let us consider the problem of supply chain sustainable management. We think on the supply chain as on a complex system with a hierarchical multilevel structure of interconnected homogeneous or heterogeneous elements that constitute an ordered and single whole. This representation is a general one, it can be easily adapted to different international standards (ISA-95, IEC 62264, IEC 61512, etc.), providing consistent information for the supply chain management. The elements of the system constitute the internal environment (each member of the supply chain can be considered as a subsystem or an element), the system itself is surrounded by an uncertain external environment (the supply chain faces competition with hardly forecastable results). Having in mind the theoretical backgrounds, the system must have the following features, including: integrality and emergence, organization (having structure and functions), functionality (showing certain properties when affecting the external environment and pursuing a goal), development (irreversible, targeted, regular change of the internal environment, which results in a new quality), self-regulation (striving to maintain the state of internal balance despite the changes in the external environment), endurance (suppressing harmful effects of the external environment), adaptation (ability to change behavior in order to maintain or improve the existing features or acquire new ones in an unspecified environment), reliability (ability to maintain structure despite the incapacity for work of distinct elements of the system).

The determinants of the supply chain sustainability can be distinguished when analyzing the above-mentioned properties. They include: self-regulation, endurance, adaptation, reliability. These properties indicate the possibility of achieving business objectives not only for the supply chain but also for its each member, as well as long-term cooperation regardless of the harmful activities of the competitors. Therefore, in this work we introduce the definition as it was presented in [7].

Definition 2.1: The supply chain sustainability is self-regulatory and adaptive managerial effort toward remaining on the market, reliable cooperation among the partners, and development aimed at achieving short-term and long-term business objectives.

As it is possible to notice from the Definition 2.1 the sustainability requires the implementation of appropriate methods of the supply chain coordination. It should be mentioned that within the framework of "fair" cooperation the interests of each party are equally important, hierarchy of enterprises and favoring the interests of one partner is not allowed. The coordination program has to be formulated as a joint one with respect of the selected functions (e.g. production, distribution). These functions determine a quick exchange of the most important information that is usually subject of a cooperation agreement. In practice, this exchange is achieved through appropriate decomposition of the management systems of each enterprise and further integration of selected subsystems into a coherent group management (sometimes only

informative). The following items are characteristic of each enterprise: hierarchy of management (management levels - strategic, tactical or operational management), and functions or subsystems of management (planning, controlling, motivation, organization, coordination). The five management subsystems can be considered as the decomposition of management system of each enterprise. Integration demands analyzing the functions of particular subsystems and understanding their hierarchy.

Let us consider this problem exclusively from the perspective of information flows. The elements of the planning subsystem include information system containing a database along with the database management system, the procedures for collecting and analyzing the information acquired and the system of interrelated plan of delivery, production, marketing, etc. at each management tier. The formation of an integrated planning subsystem consists of four stages [8], [9]. The first stage consists in developing the parameters of the information system and thus determining the required amount of information, its scope, direction, content and specificity level, as well as defining the methods of its storage and access. The second and third stage consists respectively in developing the procedures of collecting and analyzing information. The last, fourth stage consist in developing procedures of devising plans, developing entry forms, time frames and appointing individuals responsible for formation, coordination and approval. The elements of the control subsystem as part of the information system include: a system of plans, with a system of target indicators, a system of standards for the operations of entities; "management indicators". On the other hand the procedures of subsystem control involve: monitoring the current operations of an enterprise; assessment/measurement of the business entities perform assessment/measurement of the implementation of plans; transmission and dissemination of information. It should be noted that the parameters of the control subsystem are entirely dependent on the parameters of the planning subsystem, therefore the control subsystem should be formed after the planning subsystem has been created. The formation of this integrated subsystem involves defining a set of indicators and determining their quantitative values, as well as developing the procedures for monitoring, assessment/measurement, transmission and dissemination of information. The staff motivation subsystem depends on the two previous subsystems, and it is created to meet the strategic and operational goals of enterprises. The system of indicators used to assess the performance of various entities and the system of indicators of employee performance are the elements of the information system here. The procedures include assessing and motivating employees. The elements of the subsystem of forming the organizational structure include the description of the functions performed in the enterprise, powers and responsibilities of the executive entities and the organization scheme of the enterprise. The organizational structure itself is created in several stages, namely: defining executive functions in the enterprise and executive entities implementing the functions, distribution of the functions by links; organizational analysis; defining professional responsibilities, devising the "Statute of the organizational structure". The key elements of the coordination subsystem, whose formation is based on the principles of operation of the four previous subsystems, include responsibility and authority understood as a limited right to use the organization's resources and perform delegated tasks. Management procedures imply clear decision-making methods.

It results from the fact that the powers are limited to plans, procedures, rules and verbal orders of managers, as well as environmental factors. In the case of the organizational structure formation subsystem the integration involves the organization of the strategic management process, organization and formalization of the management process as well as the implementation of new management procedures. After a detailed analysis of each management subsystem, allowing for the management tier, subsystems can be incorporated into an integrated management system (IMS) which fully reflects the needs of all stakeholders: employees, customers and the public. This integration enables effective (and consequently, sustainable) management of the region, enterprise, projects, processes, etc. as it is easy to track the performance of individual tasks and information flows. The advantage of the IMS is not only simplicity, transparency, uniformity and explicitness, but also the capability of parallel existence with other business management systems. One can assume that IMS is one of the solutions to sustainable supply chain management. This means that enterprises cooperating in a group can agree on selected a model (a kind of mechanism) to ensure sustainable development of the group without specifying activities that will lead to achieving business objectives. Summing up, the integration of the supply chain management systems can be reduced to the integration of planning subsystems (at the strategic level) on the basis of the selected model using the determinants of sustainable development. [10]

III. Concepts of the mathematical model

A. The two-echelon supply chain: basic concepts and notations

In this work, the research problem is related to the determination of optimal coordination conditions for sustainable two-echelon supply chain management. We suppose that this supply chain consist from one Supplier and one Retailer, who acquire the necessary resources for the manufacturing and the distribution. Supplier has a work center and produces one type of mass products and sells it to Retailer, who commits to sell the product to the target consumers. Moreover, Retailer orders only from Supplier. To increase customer satisfaction, Supplier and Retailer undertake to provide the product of a guaranteed quality. In addition, to protect against the risk of uncertain demand, the Retailer is interested in entering into an agreement containing the terms of purchasing as much the product from the Manufacturer as the existing market demand. The supply chain members are rational and neutral. They symmetrically exchange information concerning production, market situation and etc. For the both partners the optimal operational decisions maximize the expected profits with ensuring the required level of product quality and stochastic market demand. The decision process concerns the manufacturing quantity and sales under fluctuating demand and uncertainty in the production process. In the sequel, we introduce the profit generation model, which permits to achieve the sustainability as it defined by Definition 2.1.

Let $t \in [t_0, t_1] \subset \mathbb{R}_+$ be a time moment of the supply chain members' cooperation, where t_0 stands for the beginning and

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

t_1 stands for the end of planning period. The basic parameters as well as the variables of the model are defined in Table I.

TABLE I. NOTATIONS

Notation	Descriptions
	Supplier
θ_1^S	production growth flexibility coefficient, $\theta_1^S \in \mathbb{R}$
θ_2^S	the maximal production capacity (the maximal goods quantity to be produced by Supplier), $\theta_2^S \in \mathbb{R}_+$
θ_{3j}^S	coefficient related to the defective units production, $\theta_{3j}^S \in \mathbb{R}$, $j = \{1, ..., n_S + 1\}$
$u_1(t)$	preventive maintenance rate, $u_1(t) \in [0, 1]$
$u_2(t)$	specification quality rate, $u_2(t) \in [0, 1]$
$\alpha(t)$	work center obsolescence rate, $\alpha(t) \in [0, 1]$
$S(t)$	quantity of goods produced
$I(t)$	inventory level
γ_1	guaranteed quality of finished goods, $\gamma_1 \in [0, 1]$
	Retailer
θ_1^R	distribution capability flexibility coefficient, $\theta_1^R \in \mathbb{R}$
θ_2^R	maximal distribution capacity (the maximal goods quantity to be promoted by Retailer to the consumers), $\theta_2^R \in \mathbb{R}_+$
θ_{3i}^R	coefficient related to the unforecast market demand, $\theta_{3i}^R \in \mathbb{R}$, $i = \{1, ..., n_R + 1\}$
$R(t)$	goods sold on the market
γ_2	guaranteed quality of sold goods, $\gamma_2 \in [0, 1]$

B. The production and the ordering models

We suppose that one-product manufacturing system production of Supplier consists on a work center, which transforms the raw materials through dedicated operations completed by work units into finished goods. Moreover, maintenance, quality, and inventory operations are under managerial control. These operations are used as self-regulatory and adaptive efforts dedicated to the business objectives achievement. The obsolescence rate of the work center is $\alpha(t)$ (a decreasing function on $[0, 1]$). At any time of planning period, the preventive maintenance of equipment – $u_1(t)$ – allows insuring finished goods quality between lower and upper specification limits ($u_2^{min} \leq u_2(t) \leq u_2^{max}$). As it is possible to notice the quantity of finished goods – $S(t)$ – with desired quality is varying. We assume that dynamics of quantity of goods' produced with a guaranteed quality γ_1 is stochastic. To avoid the situation of shortages due to the unexpected demand and the limited production capacities Supplier also insures the inventory level $I(t)$.

Let $\left(\Omega, \{\mathcal{F}_t\}_{t\geq 0}, \mathbf{P}\right)$ be a complete probability space, where $\{\mathcal{F}_t\}_{t\geq 0}$ is the natural filtration generated some stochastic process, augmented by all the $\mathbf{P}$-null sets in $\{\mathcal{F}_t\}_{t\geq 0}$. We suppose that, $S(t)$, B_t and $B_t^{\mathbb{H}}$ are some stochastic processes defined on this probability space. $B_t^{\mathbb{H}} = \{B(t, \mathbb{H}),\ t \in [t_0, t_1]\}$, $0 \leq t_0 \leq t_1 \leq +\infty$, is a fractional Brownian motion (fBm) with Hurst parameter $\mathbb{H} \in (0, 1)$; for $\mathbb{H} = 0.5$ fractional Brownian motion is a Brownian motion (Bm). For any $t \in [t_0, t_1]$ the system of the differential equations presents the stochastic production-inventory dynamics of Supplier concerning the demand on finished goods claimed by Retailer:

$$\begin{aligned} dS^{\gamma_1}(t) &= S^{\gamma_1}(t)\left(\theta_1^S - \frac{\theta_1^S S^{\gamma_1}(t)}{(1 + u_1(t) - \alpha(t))\,\theta_2^S}\right)dt \\ &\quad + S^{\gamma_1}(t)\left(\sum_{j=1}^{n_S}\theta_{3j}^S dB_t^{\mathbb{H}_j} + \theta_{3n_S+1}^S dB_t\right), \quad (1) \\ dI(t) &= (u_2(t)\,S^{\gamma_1}(t) - R^{\gamma_2}(t))\,dt, \end{aligned}$$

where $S^{\gamma_1}(t_0) = S_0 \leq \theta_2^S$, $I(t_0) = I_0$ are the stocks' initial values ($S_0, I_0 \in \mathbb{R}_+$), and $R^{\gamma_2}(t_0) = R_0$ is the initial demand claimed by Retailer ($R_0 \in \mathbb{R}_+$), $dB_t^{\mathbb{H}_j}$ are the increments of fBm associated with Hurst parameter $\mathbb{H}_j$, dB_t is the increment of Bm, the processes $B_t^{\mathbb{H}_j}$ and B_t are assumed to be mutually independent. The last term of right-side the first equation shows the short-range (for $\mathbb{H}_j \in (0, 0.5)$), long-range (for $\mathbb{H}_j \in (0.5, 1)$) dependences as well as the white noise (for $\mathbb{H}_j = 0.5$) in the quality variation.

The uncertain market demand obligates Retailer to get used to the changing market environment. The use of the option contracts is a typical way of the reducing the degree of uncertainty. It makes possible the flexibility in ordering products as well as in the production schedule. Besides, it is considered as a way to hedge against the risk of unexpected claims. The decision process with options contracts, as it was presented by [6], is not suitable for the long-period sustainable cooperation. To introduce the self-regulation mechanism we use the statement on the markets antimonopoly regulation and competition, which prohibits "domination" and determines the maximal sales for each trader (we denoted it by θ_2^R). Moreover, irregularities, large-amplitude, short- or long-term fluctuations, that could appear in uncertain market environment, are presented by the stochastic terms scaled by coefficients θ_{3i}^R. We let to Retailer the possibility to change flexibly the distribution capability (we denoted this parameter by θ_1^R). In this context, we assume that at any time $t \in [t_0,\ t_1]$ Supplier must deliver Retailer exactly $R^{\gamma_2}(t)$ goods. The ordering model corresponds to the following stochastic differential equation:

$$\begin{aligned} dR^{\gamma_2}(t) &= \theta_1^R R^{\gamma_2}(t)\left(1 - \frac{R^{\gamma_2}(t)}{\theta_2^R}\right)dt \quad (2) \\ &\quad + R^{\gamma_2}(t)\left(\sum_{i=1}^{n_R}\theta_{3i}^R dB_t^{\mathbb{H}_i} + \theta_{3n_R+1}^R dB_t\right), \end{aligned}$$

where $R^{\gamma_2}(t_0) = R_0$ is the initial demand claimed by Retailer ($R_0 \in \mathbb{R}_+$), $dB_t^{\mathbb{H}_i}$ is the increment of fBm associated with Hurst parameter $\mathbb{H}_i$, dB_t is Bm, processes B_t and $B_t^{\mathbb{H}_i}$ are supposed to be mutually independent. The last term of right-hand side of Eq. (2) takes into account the short-range (if $\mathbb{H}_i \in (0, 0.5)$), long-range (if $\mathbb{H}_i \in (0.5, 1)$) dependences as well as the white noise (for $\mathbb{H}_j = 0.5$) in the market demand.

C. The profit generation model

First, the profit model for Supplier is considered. We assume that at any moment of time $t \in [t_0, t_1]$ the profit $\pi_S(t)$ from economics activity related to the associated

product is formed as the difference between the sales revenue $\mathcal{P}_0^S(R^{\gamma_2}(t))$ and the expenditures on production, inventory, maintenance, and quality control operations; we denote these expenditures as $\mathcal{P}_1^S(S^{\gamma_1}(t), u_1(t))$, $\mathcal{P}_2^S(I(t), u_2(t))$, $\zeta_1(u_1(t))$, and $\zeta_2(u_2(t))$. Taking into account the stochastic nature of the production and sales in addition to the discount rate $\rho \in [0,1]$, the total expected discounted profit during the planning period is

$$\int_{t_0}^{t_1} e^{-\rho t} \pi_S(t)\, dt, \tag{3}$$

where

$$\begin{aligned} \pi_S(t) &= \mathbb{E}[\mathcal{P}_0^S(R^{\gamma_2}(t)) - \mathcal{P}_1^S(S^{\gamma_1}(t), u_1(t)) \\ &\quad -\mathcal{P}_2^S(I(t), u_2(t)) - \zeta_1(u_1(t)) - \zeta_2(u_2(t))], \end{aligned} \tag{4}$$

where $\mathbb{E}[\cdot]$ is the mathematical expectation operator, the function π_S assumed to be smooth enough.

Now we describe Retailer's profit model. The profit depends on the revenue from the sales of the goods which meet the desired quality standards and the costs of goods ordering from Supplier, denoted as $\mathcal{P}_0^R(R^{\gamma}(t))$ and $\mathcal{P}_1^R(R(t))$, respectively. Hence, the expected discounted profit during cooperation period is

$$\int_{t_0}^{t_1} e^{-\rho t} \pi_R(t)\, dt, \tag{5}$$

with

$$\pi_R(t) = \mathbb{E}\left[\mathcal{P}_0^R(R^{\gamma_2}(t)) - \mathcal{P}_1^R(R^{\gamma_2}(t))\right], \tag{6}$$

where the function π_R assumed to be smooth enough.

The key objective of the supply chain members during the cooperation period is to generate desired financial results maximizing the expected profit, namely

$$\begin{aligned} &\mathcal{J}(S(t), I(t), R(t)(t), \mathbf{u}(t)) \\ &= \sup_{\mathbf{u}(t)} \int_{t_0}^{t_1} e^{-\rho t}(\pi_S(t) + \pi_R(t))\, dt \\ &\qquad + \Phi_0(t_1, \pi_S(t_1), \pi_R(t_1)), \end{aligned} \tag{7}$$

where $\mathbf{u}(t) = [u_1(t), u_2(t)]$, $\Phi_0(t_1, \pi_S(t_1), \pi_R(t_1))$ stands for the pessimistic financial objective.

Subsequently, we consider the problem of sustainable supply chain management under stochastic demand and guaranteed quality requirement as the optimal control problem. Its solution gives the managerial insights concerning the sustainable coordination conditions for the financial target achievement.

IV. The optimal managerial coordination conditions

A. General model: some required transformations

Let us introduce some required transformations and notations before formulating the optimal control problem [11]. First, we denote state variables $s(t) = \mathbb{E}[S^{\gamma_1}(t)]$ and $r(t) = \mathbb{E}[R^{\gamma_2}(t)]$. Taking into account the rules of fractional Brownian motions approximation by fractional noises (see [12], [13]), we rewrite Eq.2 and Eq.2 as the following system

$$\begin{cases} ds(t) = \gamma_1 f^S(s(t), u_1(t))\, dt \\ \quad + \frac{\gamma_1(\gamma_1-1)}{2} \sum_{j=1}^{n_S} g_j^S\left(\sqrt{s(t)}\right) (dt)^{2\mathbb{H}_j}, \ s(t_0) = S_0, \\ dI(t) = f^I(s(t), r(t), u_2(t))\, dt, \ I(t_0) = I_0, \\ dr(t) = \gamma_2 f^R(r(t))\, dt \\ \quad + \frac{\gamma_2(\gamma_2-1)}{2} \sum_{i=1}^{n_R} g_i^R\left(\sqrt{r(t)}\right) (dt)^{2\mathbb{H}_i}, \ r(t_0) = R_0, \end{cases} \tag{8}$$

where

$$\begin{aligned} f^S(s(t), u_1(t)) &:= \theta_1^S(t)\, s(t) \\ &\times \left(1 - \tfrac{s(t)}{(1+u_1(t)-\alpha(t))\theta_2^S(t)}\right), \\ f^I(s(t), r(t), u_2(t)) &:= (u_2(t)\, s(t) - r(t)), \\ f^R(r(t)) &:= \theta_1^R(t)\, r(t) \left(1 - \frac{r(t)}{\theta_2^R(t)}\right), \\ g_j^S\left(\sqrt{s(t)}\right) &:= \sqrt{s(t)}\theta_{3j}^S, \ j \in \{1, 2, ..., n_S\} \\ g_i^R\left(\sqrt{r(t)}\right) &:= \sqrt{r(t)}\theta_{3i}^R, \ i \in \{1, 2, ..., n_R\}. \end{aligned}$$

The system (8) can be considered as the fractional differential system. To formulate the optimal control problem we rewrite each equation of (8) in the integral form

$$\begin{aligned} s(t) &= S_0 + \int_{t_0}^{t} f^S(s(\tau), u_1(\tau))\, d\tau \\ &+ \gamma_1(1-\gamma_1) \sum_{\Lambda_1^S} \mathbb{H}_j \int_{t_0}^{t} \frac{g_j^s(s(\tau))}{(t-\tau)^{1-2\mathbb{H}_j}} d\tau \\ &+ \sum_{\Lambda_2^S} \left[\mathbb{H}_j \int_{t_0}^{t} \frac{\sqrt{\frac{1}{2}\gamma_1(1-\gamma_1) g_j^s(s(\tau))}}{(t-\tau)^{1-\mathbb{H}_j}} d\tau\right]^2, \end{aligned} \tag{9}$$

$$I(t) = I_0 + \int_{t_0}^{t} f^I(s(\tau), r(\tau), u_2(\tau))\, d\tau, \tag{10}$$

$$\begin{aligned} r(t) &= R_0 + \int_{t_0}^{t} f^R(r(\tau))\, d\tau \\ &+ \gamma_2(1-\gamma_2) \sum_{\Lambda_1^R} \mathbb{H}_i \int_{t_0}^{t} \frac{g_i^R(r(\tau))}{(t-\tau)^{1-2\mathbb{H}_i}} d\tau \\ &+ \sum_{\Lambda_2^R} \left[\mathbb{H}_i \int_{t_0}^{t} \frac{\sqrt{\frac{1}{2}\gamma_1(1-\gamma_1) g_i^R(r(\tau))}}{(t-\tau)^{1-\mathbb{H}_i}} d\tau\right]^2, \end{aligned} \tag{11}$$

where $g_j^S(s(t)) := \left(\theta_{3j}^S\right)^2 \ s(t)$, $g_i^R(r(t)) := \left(\theta_{3i}^R\right)^2 \ r(t)$,

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

$\Lambda_1^S = \{j \in \{1, ..., n_S + 1\} : \mathbb{H}_j < 0.5\}$ and $\Lambda_2^S = \{j \in \{1, ..., n_S + 1\} : \mathbb{H}_j > 0.5\}$, $\Lambda_1^R = \{i \in \{1, ..., n_R + 1\} : \mathbb{H}_i < 0.5\}$ and $\Lambda_2^R = \{i \in \{1, ..., n_R + 1\} : \mathbb{H}_i > 0.5\}$.

We consider Eq.(9)–Eq.(11) as the object equation, where $s(t)$, $I(t)$, and $r(t)$ stand for the state variables, $u_1(t)$ and $u_2(t)$ are the control variables ($t \in [t_0,\ t_1]$). For the sake of simplicity, we assume that $n_R = 1$ and $n_S = 1$ (so that we omit summation index i in (9) and index j in 11). Moreover, to uniform the notations we introduce new notations $x_1 := s$, $x_2 := I$, $x_3 := r$. Thus $\mathbf{x} = [x_1, x_2, x_3]$ will be a state vector. We also uniform the "shape" of the equations (9) – (11), so that each component of state vector takes a form:

$$x_m(t) - a_m = \int_{t_0}^{t} \mathbf{x}(\tau)\, \Phi_{1m}(\tau, \mathbf{u}(\tau))\, d\tau \tag{12}$$
$$+ \int_{t_0}^{t} \frac{\beta_m}{(t-\tau)^{1-\beta_m}} \mathbf{x}(\tau)\, \Phi_{2m}(\tau, \mathbf{u}(\tau))\, d\tau$$
$$+ \mathbb{H}_{2m}^2 \left[\int_{t_0}^{t} \frac{\sqrt{\Phi_{3m}(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))}}{(t-\tau)^{1-\mathbb{H}_{2m}}} d\tau \right]^2,$$

where $\beta_m = 2\mathbb{H}_{1m}$, $\mathbb{H}_{1m} < 0.5$, $\mathbb{H}_{2m} > 0.5$, $a_m = x_m(t_0)$, Φ_{1m}, Φ_{2m}, and Φ_{1m} are some functions, $m = \{1, 2, 3\}$.

At the same time, taking into account the concepts discussed at length in Section III for all $t \in [t_0, t_1]$ the following constraints on the parameters and variables are fulfilled:

- the state constraints: $\mathcal{G}(\mathbf{x})$ is some vector function of the dimension ℓ_1, such that

$$\mathcal{G}(\mathbf{x}(t)) \leq 0, \tag{13}$$

e.g. for $\ell_1 = 1$

$$x_3(t) - x_1(t) - x_2(t) \leq 0, \tag{14}$$

- the control constraints: $\varphi(\mathbf{u})$ is some vector function of the dimension ℓ_2, such that

$$\varphi(\mathbf{u}(t)) \leq 0, \tag{15}$$

e.g. for $\ell_2 = 3$

$$u_1(t) - \alpha(t) \leq 0, \tag{16}$$
$$u_2^{min} - u_2(t) \leq 0, \tag{17}$$
$$u_2(t) - u_2^{max} \leq 0, \tag{18}$$

(the control constraints form the set $\mathcal{U} = \{u : u_1(t) - \alpha(t) \leq 0, u_2^{\min} \leq u_2(t) \leq u_2^{\max},\ a.e.\ t \in [t_0, t_1]\}$)

- the mixed constraints: $\xi(\mathbf{x}, \mathbf{u})$ is some vector function of the dimension ℓ_3, such that

$$\xi(\mathbf{x}(t), \mathbf{u}(t)) \leq 0, \tag{19}$$

e.g. for $\ell_3 = 1$

$$(1 + u_1(t) - \alpha(t))\, \theta_2^S - x_1(t) \leq 0. \tag{20}$$

The constraint (16) prohibits the maximization of the production capacity. The constraints (17) and (18) show the target quality levels u_2^{min} and u_2^{max}. The constraints (14) and (20) mean that the shortages are not allowed.

Finally, we put $F(\mathbf{x}(t), \mathbf{u}(t)) := e^{-\rho\ t}(\pi_S(t) + \pi_R(t))$ and rewrite the goal function (7)

$$\mathcal{J}(\mathbf{x}(t), \mathbf{u}(t)) = \sup_{\mathbf{u}(t) \in \mathcal{U}} \int_{t_0}^{t_1} F(\mathbf{x}(t), \mathbf{u}(t))\, dt + \Phi_0(t_1, \pi_S(t_1), \pi_R(t_1)), \tag{21}$$

where $\mathbf{x}(t)$ is the state variable, $\mathbf{u}(t)$ is the control variable, F and Φ_0 are some functions.

The problem (21) under constraints (13), (13), (15), and (19) will be called *Problem*. The solution of *Problem* can be considered as the optimal coordination conditions for the sustainable supply chain management.

B. Necessary optimality conditions

Problem can be classified as the optimal control problem with integral equations on a variable time interval. The necessary optimality conditions for a weak minimum in this class of problems with homogeneous integral equations were studied by [14] and generalized by [15]. In our case the system (13) contains the inhomogeneous integral equations. Therefore, in our case the methodology of the solution of *Problem* has to be adapted. One can use the classical calculus of variation obtaining the Euler-Lagrange equation and transversality conditions for Lagrange problem (see for the details [16], [17]). We will omit the details of the calculus and present only first-order necessary optimality conditions for weak maximum in *Problem*.

For this purpose, we denote the time interval $\Delta = [t_0, t_1]$ and make some assumptions on the functions used in the problem formulation, namely: $\mathbf{x}(\cdot) \in \mathcal{AC}(\Delta, \mathbb{R}^n)$, $\mathbf{u}(\cdot) \in \mathcal{L}^\infty\ (\Delta, \mathbb{R}^{\ell_1})$, $F \in \mathcal{C}^1$ in $\mathbf{x}$ and continuous in $\mathbf{u}$, the set $\mathcal{U} \subset \mathbb{R}^{\ell_1}$ is defined by the control constraints, $\mathcal{G} \in \mathcal{C}^1$, $\varphi \in \mathcal{C}^1$, and $\xi \in \mathcal{C}^1$ are vector functions of dimension ℓ_1, ℓ_2, and ℓ_3 respectively. Introduce the notation: $\Delta[t_0, t_1] = \{(t, \tau) : t_0 \leq \tau \leq t \leq t_1\}$.

Definition 4.1: A pair of functions $(\mathbf{x}(t), \mathbf{u}(t))$ defined on an interval Δ will be called a process in Problem if its "extended graph" $\{(t, \tau, (x)(\tau), (u)(\tau) | (t, \tau) \in \Delta[t_0, t_1]\}$ lies in the set $\mathcal{R}$ with some "margin", i.e. $dist((t, \tau, \mathbf{x}(\tau), \mathbf{u}(\tau)), \partial\mathcal{R}) \geq const > 0$ for a.a. $(t, \tau) \in \Delta[t_0, t_1]$.

Definition 4.2: Any process will be called an admissible one, if the conditions of *Problem* are fulfilled.

Definition 4.3: A process $(\widehat{\mathbf{x}}(t), \widehat{\mathbf{u}}(t))$ will be called an optimal one if there exists an $\varepsilon > 0$ such that, for any admissible process $(\mathbf{x}(t), \mathbf{u}(t))$ satisfying the restriction

$$\|\mathbf{x}(\cdot) - \widehat{\mathbf{x}}(\cdot)\|_{\mathcal{C}(\Delta, \mathbb{R}^n)} < \varepsilon,$$

one has

$$\mathcal{J}(\mathbf{x}(\cdot), \mathbf{u}(\cdot)) \leq \mathcal{J}(\widehat{\mathbf{x}}(\cdot), \widehat{\mathbf{u}}(\cdot)).$$

Let us introduce a new state variable

$$y_m(t) = \int_{t_0}^{t} \frac{g_m(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))}{(t-\tau)^{1-H_m}} d\tau,$$

where $g_m(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau)) = H_{2m}\sqrt{\Phi_{3m}(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))}$. Then Eq. (13) reduces to the form

$$\begin{aligned} x_m(t) - a_m &= \int_{t_0}^{t} \mathbf{x}(\tau)\,\Phi_{1m}(\tau, \mathbf{u}(\tau))\,d\tau \\ &+ \int_{t_0}^{t} \frac{\beta_m \mathbf{x}(\tau)}{(t-\tau)^{1-\beta_m}} \Phi_{2m}(\tau, \mathbf{u}(\tau))\,d\tau \\ &+ y_m^2(t), \end{aligned}$$

where $\beta_m = 2\mathbb{H}_{1m}$, $\mathbb{H}_{1m} < 0.5$, $H_{2m} > 0.5$, $m = \{1, 2, 3\}$. This last equation allows to consider a more general system of integral equations:

$$\begin{aligned} x_m(t) &= \Theta(y_m(t)) + \int_{t_0}^{t} \mathbf{x}(\tau)\,\Phi_{1m}(\tau, \mathbf{u}(\tau))\,d\tau \\ &+ \int_{t_0}^{t} \frac{\beta_m \mathbf{x}(\tau)}{(t-\tau)^{1-\beta_m}} \Phi_{2m}(\tau, \mathbf{u}(\tau))\,d\tau, t \in \Delta, \\ y_m(t) &= y_m(t_0) + \int_{t_0}^{t} \frac{g_m(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))}{(t-\tau)^{1-H_m}} d\tau, \\ &\quad t \in \Delta, y_m(t_0) = b_m, \end{aligned}$$

where $\Theta(y_m(t))$ is an arbitrary smooth ($\mathcal{C}^1$) function.

By similar arguments, as in [15], we get the first-order necessary optimality conditions for a given process $(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t))$, $t \in \Delta$, $\Theta(y_m(t_0)) = a_m$.

Theorem 4.1: Let $(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t))$ be the optimal process on the interval Δ, where $\mathbf{x}(\cdot), \mathbf{y}(\cdot) \in \mathcal{AC}(\Delta, \mathbb{R}^n)$, $\mathbf{u}(\cdot) \in \mathcal{L}^\infty(\Delta, \mathbb{R}^{\ell_1})$. Then there exist a number α_0 and functions $\mu_i(t)$, $1 \le i \le \ell_1$, $h_k(t)$, $1 \le k \le \ell_3$, an absolutely continuous vector-function of bounded variation $\psi(t)$ (which defines the measure $d\psi$), a function of bounded variation $\lambda_s(t)$, $1 \le s \le \ell_2$, (which defines the Radon measure $d\lambda_s$) such that the following conditions hold:

- nontriviality: $|\alpha_0| + \|\lambda\| + \|\mu\| + \|\xi\| > 0$,
- nonnegativity: $\alpha_0 \ge 0$, $d\mu_i(t) \ge 0$, $d\lambda_s(t) \ge 0$, $dh_k(t) \ge 0$ for all i, s, k;
- complementarity: $\mathcal{G}_i(\mathbf{x}(t))d\mu_{ji}(t) = 0$, $\varphi_s(\mathbf{u}(t))\,d\lambda_s(t) = 0$, $h_k(t)\,\xi_k(\mathbf{x}(t), \mathbf{u}(t)) = 0$ for all i, s, k;
- transversality: $\psi(t_1) = \alpha_0 \Phi'(t_1, \pi_S(t_1), \pi_R(t_1))$;
- adjoint equation:

$$\begin{aligned} -\psi'(t) &= \psi(t)\Phi_1(t, \mathbf{u}(t)) \\ &+\beta\Big[\frac{\psi(t_1)}{(t_1-t)^{1-\beta}} - \int_t^{t_1} \frac{\psi'(\tau)}{(\tau-t)^{1-\beta}} d\tau\Big] \\ &\times \Phi_2(t, \mathbf{u}(t)) \\ &+\Big[\frac{\psi(t_1)\,\Theta'(y(t_1))}{(t_1-t)^{1-\mathbb{H}_2}} \\ &- \int_t^{t_1} \frac{\psi'(\tau)\,\Theta'(y(\tau))}{(\tau-t)^{1-\mathbb{H}_2}} d\tau\Big] g(t, \mathbf{x}(t), \mathbf{u}(t)) \\ &+\alpha_0 F_x(t, \mathbf{x}(t), \mathbf{u}(t)); \end{aligned}$$

- local maximum principle:

$$\begin{aligned} &-\psi'(t) \\ &= \psi(t)\mathbf{x}(t)\,\Phi_{1u}(t, \mathbf{u}(t)) + \beta\Big[\frac{\psi(t_1)}{(t_1-t)^{1-\beta}} \\ &- \int_t^{t_1} \frac{\psi'(\tau)}{(\tau-t)^{1-\beta}} d\tau\Big] \times \mathbf{x}(t)\,\Phi_{2u}(t, \mathbf{u}(t)) \\ &+ \Big[\frac{\psi(t_1)\,\Theta'(y(t_1))}{(t_1-t)^{1-\mathbb{H}_2}} \\ &- \int_t^{t_1} \frac{\psi'(\tau)\,\Theta'(y(\tau))}{(\tau-t)^{1-\mathbb{H}_2}} d\tau\Big] g_u(t, \mathbf{x}(t), \mathbf{u}(t)) \\ &+ \alpha_0 F_x(t, \mathbf{x}(t), \mathbf{u}(t)) \\ &- \lambda^a(t)\,\varphi'(\mathbf{u}(t)) + h^b(t)\,\xi'(\mathbf{x}(t), \mathbf{u}(t)), \end{aligned}$$

where $\lambda^a(t) \ge 0$ and $\lambda^a(t)\,\varphi(\mathbf{u}(t)) = 0$, in addition $h^b(t) \ge 0$ and $h^b(t)\,\xi(\mathbf{x}(t), \mathbf{u}(t))$.

If the conditions of the Theorem 4.1 are fulfilled, the coordination conditions of the sustainable supply chain management are optimal.

V. Conclusions

The successful collaboration among enterprises requires the special methods of management. The managerial decisions are always data-driven. Therefore to solve the problem of two-echelon supply chain sustainable management we proposed the control model and proved that it can be controlled in some optimal manner. In its present mathematical formulation this model can be implemented as reinforcement learning model. The main advantage of this solution consists in the theoretical prove of desired properties. In upcoming works we will concentrate on reinforcement learning algorithm development.

References

[1] S. A. Salehi Amir, A. Zahedi, M. Kazemi, J. Soroor, and M. Hajiaghaei-Keshteli, "Determination of the optimal sales level of perishable goods in a two-echelon supply chain network," *Computers and Industrial Engineering*, vol. 139, p. 106156, 2020.

[2] C. Carter and D. Rogers, "A framework of sustainable supply chain management: moving toward new theory," *International Journal of Physical Distribution and Logistics Management*, vol. 38, no. 5, pp. 360–387, 2008.

[3] Y. Shou, J. Shao, K. hung Lai, M. Kang, and Y. Park, "The impact of sustainability and operations orientations on sustainable supply management and the triple bottom line," *Journal of Cleaner Production*, vol. 240, p. 118280, 2019.

[4] M. Zheng, K. Wu, C. Sun, and E. Pan, "Optimal decisions for a two-echelon supply chain with capacity and demand information," *Advanced Engineering Informatics*, vol. 39, pp. 248 – 258, 2019.

[5] C. Canyakmaz, S. Özekici, and F. Karaesmen, "An inventory model where customer demand is dependent on a stochastic price process," *International Journal of Production Economics*, vol. 212, pp. 139 – 152, 2019.

[6] X. Chen, N. Wan, and X. Wang, "Flexibility and coordination in a supply chain with bidirectional option contracts and service requirement," *International Journal of Production Economics*, vol. 193, pp. 183 – 192, 2017.

[7] D. Filatova and R. Fedorenko, "Sustainable development concepts as a response to expectations of modern business management," *SHS Web of Conferences*, vol. 71, no. 5, 2019.

[8] D. Bochnacka and D. Filatova, "Necessary optimality conditions for enterprises production programs," in *2017 International Conference on Information and Digital Technologies (IDT)*, pp. 61–65, July 2017.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

[9] D. Filatova and C. El-Nouty, "Production process balancing: A two-level optimization approach," in *2019 International Conference on Information and Digital Technologies (IDT)*, pp. 133–141, June 2019.

[10] A. P. V. Acosta, C. Mascle, and P. Baptiste, "Applicability of demand-driven mrp in a complex manufacturing environment," *International Journal of Production Research*, vol. 0, no. 0, pp. 1–13, 2019.

[11] D. V. Filatova, A. Orłowski, and V. Dicoussar, "Estimating the time-varying parameters of sde models by maximum principle," in *2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pp. 401–406, Sep. 2014.

[12] B. Mandelbrot and J. van Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.

[13] H. Sheng, Y. Chen, and T. Qiu, *Fractional Processes and Fractional-Order Signal Processing: Techniques and Applications*. Springer, 2012.

[14] A. Dmitruk and N. Osmolovskii, "Necessary conditions for a week minimum in optimal control problems with integral equations on a variable interval," *Discrete and continuous dynamic systems*, vol. 35, no. 9, pp. 4323–4343, 2015.

[15] A. Dmitruk and N. Osmolovski, "Necessary conditions for a weak minimum in a general optimal control problem with integral equations on a variable time interval," *Mathematical Control & Related Fields*, vol. 7, no. 4, pp. 507–535, 2017.

[16] A. Milyutin, A. Dmitruk, and N. Osmolovskii, *Maximum principle in optimal control*. Moscow State University, 2004.

[17] D. Filatova, M. Grzywaczewski, and N. Osmolovskii, "Optimal control problem with an integral equation as the control object," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 72, no. 3–4, pp. 1235–1246, 2010.

Object Position Estimation from a Single Moving Camera

1st Milan Ondrašovič
Faculty of Management Science and Informatics
University of Žilina
010 26, Žilina, Slovakia
milan.ondrasovic@fri.uniza.sk

2nd Peter Tarábek
Faculty of Management Science and Informatics
University of Žilina
010 26, Žilina, Slovakia
peter.tarabek@fri.uniza.sk

3rd Ondrej Šuch
Mathematical Institude SAV
Ďumbierska 1, 974 01 Banská Bystrica
ondrej.such@fri.uniza.sk

***Abstract*—This paper deals with the position estimation of a road sign from a single camera attached to a vehicle. We developed, implemented, and tested two mathematical approaches based on triangulation when the object annotation in the form of a bounding box is provided. We created a synthetic dataset (a simulation of a car passing by a road sign) to test the methods in a controlled environment. Additionally, the real dataset was created by recording a car trip within a town. Results on the synthetic dataset showed that the position could be estimated within 1 m accuracy. In the case of the real dataset, we measured the accuracy to be up to 4.3 m depending on the distance from the object. We performed experiments with artificial noise on synthetic data to evaluate the impact of different types of noise. Our contribution consists of two computationally inexpensive methods for object position estimation that are easy to use as they do not require calibration of parameters.**

***Index Terms*—position estimation, static object, single moving camera, triangulation, traffic sign**

I. Introduction

There are many vision-related operations that humans perform well at, one of which is the perception of depth. This task is non-trivial, and thus despite a plethora of prior research, it still is not completely solved [19]. A typical camera has a single lens and a single sensor, so it produces a transformation from 3D into 2D (image plane) [7]. This transformation does not preserve depth information, therefore, computing the inverse transformation is an ill-posed problem. The problem of depth estimation is on-going research with multiple approaches proposed [9], [16]. There are numerous techniques available, and the most widely used methods are based on determining depth using stereo vision [11]. Stereo vision is a challenging problem in computer vision [3]. The human visual system solves the problem with two eyes that enable a person to achieve a binocular, stereoscopic vision. Depth perception crucially depends on the ability of the nervous system to register a presence of small differences in object position from the projections onto the left and right retina [17]. In this paper, we estimate object position using a single moving camera. Instead of computing distances based on observed differences in images obtained from a stereo vision setup, we use differences in images caused by the movement of a single camera instead.

This work was motivated by the real-world task of localizing road signs to automatically document their positions from a recording using an ordinary dashboard camera in a car. With this in mind, we proposed two methods to estimate the position of a distant object using just the embedded camera. We are concerned only with 2D position estimation, thus, through the paper, only x and y coordinates will be computed.

The task of depth estimation from a single image is delicate and even more so if the camera is uncalibrated [13]. In such a setup, the parallax effect (achieved using stereovision) cannot be exploited, thus leaving the problem with too many unknowns [6]. Our methods require only a single camera, so the user does not need any stereovision-based cameras. The goal was to derive, implement, and test methods capable of emulating a stereovision environment for the estimation of the object position solely by the use of the single moving camera.

We assume that the input to our system is the object bounding box (BBOX) together with the camera position at the time of capturing the frame. Our goal will be to estimate the position of road signs. We use only a single uncalibrated camera attached to a moving vehicle. Our approach is computationally very efficient and easy to implement. Moreover, it does not require a complicated setup of parameters, which makes it usable in its default settings. Considering this, we would summarize our **contribution** as follows:

- We propose two simple methods for position estimation of a static object from a single camera based on triangulation. Their implementation is straightforward, and the utilization is vast since only a single camera with no special hardware is required.
- We provide an experimental evaluation of our models on synthetic data (including noise) as well as real-world data to measure the accuracy in various scenarios.

The rest of the paper is organized as follows. The upcoming Section II contains an overview of related work. Then, in

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Section III, we describe our proposed methods. Section IV is devoted to experiments and their evaluation. We experiment on synthetic and real data and evaluate the influence of noise as well as the influence of object distance from the camera. We end this paper with Section V by summarizing our conclusions.

II. Related Work

A prevailing method to tackle the problem of object position estimation is to use cameras supporting stereo vision [11]. This approach uses two cameras with known lateral displacement allowing the exploitation of the parallax effect to compute the distance of the visible object on both cameras [12]. One of our methods (III-C) is similar to the one described in [20]. Here, two different views of the same object from the Google Street View interface allowed us to estimate the object position using triangulation. The object can also be localized when coordinates of multiple visible points in space are known as shown in [2]. The authors showed that 4 points with known ground-truth positions in the scene were sufficient to derive the position information about other objects. In our case, we assume that only the camera position is known. Object localization is also possible when the camera is moving in a predictable path, as demonstrated in [14] with a drone flying in a specific pattern while observing the object of interest.

Han et al. [10] tackled the problem of traffic sign positioning for traffic facility maintenance. Their work included the object detection phase and object tracking, too. What we have in common is the use of Global Positioning System (GPS) measurements and the position estimation using the space intersection of image rays at different epochs. Lee et al. [15] installed the GPS, inertial measurement unit (IMU), distance measuring instrument (DMI), camera, and laser sensors on the van for road sign positioning. The authors could estimate the position within the 1.5 m. Camera calibration, i.e. estimating intrinsic and distortion camera coefficients, is not necessary for localization, as shown in [4]. We assume no camera calibration, too. In this work, triangulation combined with the least-squares optimization was exploited to localize the object.

The notion of a depth map is also related to our problem. Such a map represents depth by adjusting the intensity of each pixel according to the depth of the corresponding object. We can obtain the depth map using conventional pixel matching in stereoscopic pair of images [1], or machine learning [18]. In recent years, depth estimation using deep learning has also shown a very good performance [9], [21]. Very similar is the disparity map containing the difference in the positions of the two corresponding pixels in their respective images [3].

III. Proposed Methods

A. Preliminaries

Here we introduce prerequisites and notation common to both methods. A general requirement is the presence of the two frames containing the object of interest to execute the calculation. Since both of our methods rely on triangulation, we demand a sufficient camera displacement that causes the position of the object of interest in the image (frame) to be distinctive enough.

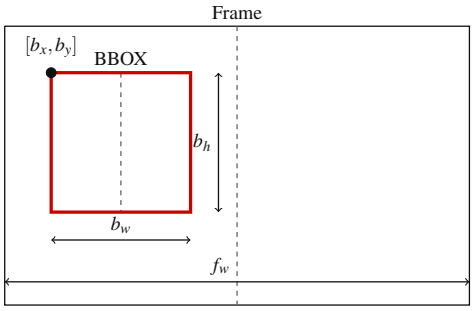


Fig. 1. Schematic illustration of quantities b_x, b_w, and f_w used for calculation of a relative angle under which the center of BBOX is visible on the frame.

Let $\mathcal{F}_1$ and $\mathcal{F}_2$ denote the two input frames used for the position estimation and let t_1 and t_2 be their respective timestamps. Assume $0 \leq t_1 < t_2$. Let $b_x^{(1)}$ be the x coordinate of the top-left corner, $b_w^{(1)}$ be the width of the BBOX of the object of interest, and let $f_w^{(1)}$ be the frame width for the $\mathcal{F}_1$. Denotations $b_x^{(2)}$, $b_w^{(2)}$, and $f_w^{(2)}$ are defined analogically for $\mathcal{F}_2$ (see Fig. 1). The horizontal field of view (HFoV) of the camera is given by v_h. The angles α_1 and α_2, such that $\alpha_1, \alpha_2 \in \left\langle -\frac{v_h}{2}, \frac{v_h}{2} \right\rangle$, represent angles under which the object of interest is visible on both frames. We compute them as

$$\alpha_1 = \varphi\left(b_x^{(1)}, b_w^{(1)}, f_w^{(1)}\right), \quad \alpha_2 = \varphi\left(b_x^{(2)}, b_w^{(2)}, f_w^{(2)}\right),$$

where

$$\varphi\left(b_x, b_w, f_w\right) = \frac{v_h}{2}\left(\frac{\left(b_x + \frac{b_w}{2}\right) - \frac{f_w}{2}}{\frac{f_w}{2}}\right).$$

The angles α_1 and α_2 need to be converted into new, "global", angles $\overline{\alpha_1}$ and $\overline{\alpha_2}$. These new angles bear information about the current camera azimuth (vehicle heading) on the frames $\mathcal{F}_1$ and $\mathcal{F}_2$ denoted as θ_1 and θ_2, respectively. We thus define

$$\overline{\alpha_1} = \theta_1 + \alpha_1, \quad \overline{\alpha_2} = \theta_2 + \alpha_2.$$

Points $\mathbf{p}_1^T = [x_1, y_1]$ and $\mathbf{p}_2^T = [x_2, y_2]$ represented by column vectors specify the camera position corresponding to $\mathcal{F}_1$ and $\mathcal{F}_2$. The final estimated (target) position of our object of interest is denoted by $\mathbf{p}_t^T = [x_t, y_t]$. Fig. 2 covers the general principle of triangulation upon which we based both of the further discussed methods. We provide brief derivations of the equations presented in III-B and III-C in the Appendix.

B. Lines Intersection Method

This method (abbreviated as LIM) is analogous to tracing a ray at the object of interest through the optical center of the camera from two different positions $\mathbf{p}_1$ and $\mathbf{p}_2$ and subsequently computing the point of intersection. The target object position $\mathbf{p}_t^T = [x_t, y_t]$ is estimated as

$$x_t = \frac{y_2 - y_1 + \tan\left(\overline{\alpha_1}\right) x_1 - \tan\left(\overline{\alpha_2}\right) x_2}{\tan\left(\overline{\alpha_1}\right) - \tan\left(\overline{\alpha_2}\right)}, \tag{1}$$

$$y_t = \tan\left(\overline{\alpha_1}\right) x_t + y_1 - \tan\left(\overline{\alpha_1}\right) x_1. \tag{2}$$

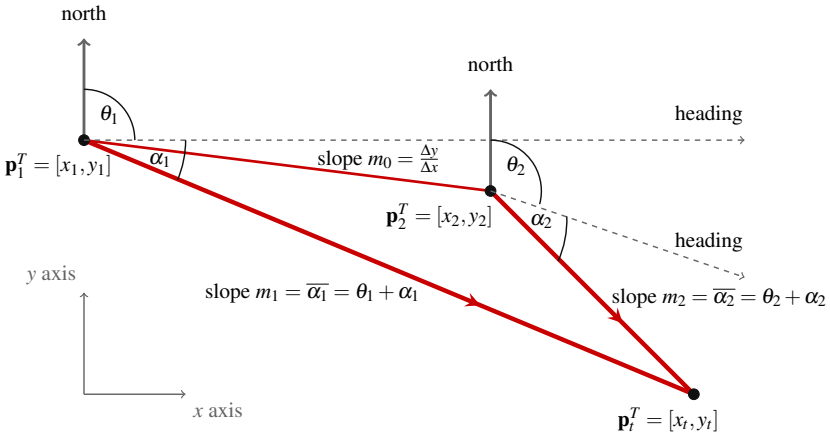


Fig. 2. Illustration of the established denotations in a general scene where the object position is estimated from two different views. Triangulation is based on two different angles produced by a change in camera position. The object of interest is visible from both positions.

C. Law of Sines Method

This method (abbreviated as LSM) is based on the construction of a triangle in the scene and then using the law of sines. This mathematical rule is adopted to compute the length of one side given two other angles (see Fig. 2). The slope of the line connecting the two camera positions is required, so let m_0 be the slope of the line passing through points $\mathbf{p}_1$ and $\mathbf{p}_2$, therefore $m_0 = \frac{y_2 - y_1}{x_2 - x_1}$. Slopes corresponding to the lines connecting camera positions with the position of the object of interest are given by $\overline{\alpha_1}$ and $\overline{\alpha_2}$, but for the sake of clarity, we define $m_1 = \overline{\alpha_1}$ and $m_2 = \overline{\alpha_2}$. Let the function ρ be the angle between two lines specified by their respective slopes a and b. We compute the function ρ as

$$\rho(a, b) = \arctan\left(\frac{a - b}{1 + ab}\right). \tag{3}$$

To simplify the equations, we make substitutions

$$q = \rho(m_0, m_1), \quad r = \rho(m_0, m_2),$$
$$\omega = \sin(\pi - r)\left(\frac{\|\mathbf{p}_1 - \mathbf{p}_2\|_2}{\sin(r - q)}\right).$$

Then, the target object position $\mathbf{p}_t^T = [x_t, y_t]$ is estimated as

$$x_t = x_1 + \omega \cos(m_1), \tag{4}$$
$$y_t = y_1 + \omega \sin(m_1). \tag{5}$$

D. Computing Final Position Estimate

Both aforementioned methods operate on only two frames at a time. We can improve the accuracy of estimation using the fact that the camera is moving while the object is static. From a sequence of n frames, we can create at most $\binom{n}{2}$ distinct pairs of frames for localization. In this section, we describe simple filtering conditions to potentially filter outliers. Additionally, we propose a method to combine multiple estimates into one.

1) Filtering Outliers: In case the object is visible under the same relative angle ($\alpha_1 = \alpha_2$) combined with the same vehicle heading ($\theta_1 = \theta_2$), the expression in (1) results in division by zero (since $\overline{\alpha_1} = \overline{\alpha_2}$). Triangle construction also fails when the two lines with slopes $\overline{\alpha_1}$ and $\overline{\alpha_2}$ are parallel, making the expression evaluating the angle between two lines in (3) too small. Therefore, to circumvent the two mentioned obstacles, we recommend filtering cases where

$$|\overline{\alpha_1} - \overline{\alpha_2}| < \varepsilon, \tag{6}$$

such that ε is a small positive threshold (we used 10^{-5}).

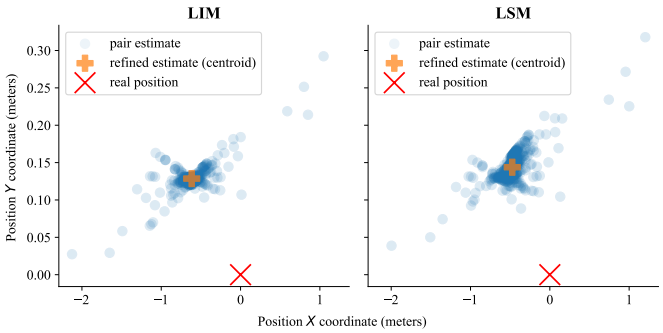


Fig. 3. Scatter plot of pair position estimates in of both methods using synthetic dataset `hill`. Other datasets produce similar patterns.

2) Combining Multiple Position Estimates: We mentioned that from a sequence of n frames we can create at most $\binom{n}{2}$ distinct pairs of frames to estimate the position. Nevertheless, the number is usually considerably lower due to the filtering conditions described in III-D1. Assuming there are multiple position estimates, then we can exploit their quantity to derive the final position estimate. For clarity, we will refer to a single estimate based on two frames as a *pair estimate* while the improved estimate computed from multiple pair estimates will be a *refined estimate*. Let $\mathcal{E}$ be a list of n pair estimates, such that $\mathcal{E} = \left(\mathbf{p}_p^{(1)}, \mathbf{p}_p^{(2)}, \ldots, \mathbf{p}_p^{(n)}\right)$, where $\mathbf{p}_p^{(i)^T} = \left[x_p^{(i)}, y_p^{(i)}\right]$, for $i = 1, \ldots, n$.

The approach we used was a computation of centroid of all pair estimates. We chose this to be the standard method for the refined estimate of the object position because it is simple and efficient. However, we do emphasize that there is

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

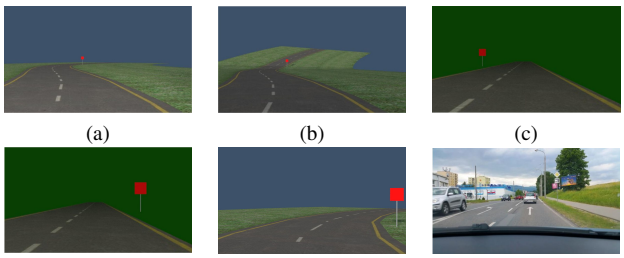


Fig. 4. Examples of different curvature, elevation and road sign position in synthetic datasets (a) `flat`, (b) `hill`, (c) `s-left`, (d) `s-right`, (e) `up`, and a real dataset (f) `town`.

Table I
DESCRIPTION OF THE DATASET ROAD STRUCTURE AND FRAME PROPERTIES.

Dataset	Hills	Elevation change	Curvature	Resolution	Frames no.
`flat`	no	yes	yes	1280×720	254
`hill`	yes	yes	yes	1280×720	130
`up`	no	no	yes	1280×720	255
`s-left`	no	yes	no	1280×720	32
`s-right`	no	yes	no	1280×720	33
`town`	no	yes	no	1920×1080	1474

a potential for improvement (see IV-B3). The refined estimate $\mathbf{p}_r$ represented by the global centroid is computed based on all pair estimates as

$$\mathbf{p}_r = \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_i. \tag{7}$$

Given this additional step, we can describe our entire object estimation pipeline as follows. The input consists of several frames on which the object of interest is visible. Then, all possible combinations of frame pairs are checked to whether they meet conditions for computations. All valid frame pairs are subsequently fed into one of the proposed position estimation methods, which generates multiple pair position estimates. These pair position estimates are then combined into a refined (final) position estimate.

IV. EXPERIMENTS

In this section, we describe the performed experiments. The goal was to measure the accuracy of position estimation of both proposed methods.

A. Dataset Creation

We created a synthetic as well as a real dataset to evaluate our models. The synthetic dataset served the purpose of an environment with ideal conditions. We also employed artificial noise to measure its influence in isolation on synthetic data (IV-B5). In the end, to assess a real-world performance, we created a real dataset as a car trip to a town.

1) Synthetic Dataset: To experiment with the developed methods in a controlled environment we created a synthetic dataset (see Fig. 4). Each scene contains a different type of road in terms of elevation or curvature. The model also contains scenes with a variable road sign position placed on different parts of the road. We created dataset scenes using a 3D modeling/simulation software called *Blender* [5]. Names of datasets in this category are `flat`, `hill`, `up`, `s-left` and `s-right` (with letter "s" abbreviating the word "straight"). Each synthetic dataset contains one object (road sign). We refer to the horizontal and vertical angular field of view of the camera as HFoV and VFoV, respectively. For the synthetic datasets, we used HFoV $\approx 49.1°$ and VFoV $\approx 28.8°$

2) Real Dataset: To test the developed methods in a real environment we also created a real dataset called `town` (see Fig. 4). The real dataset contains 19 objects (road signs). The camera we used to record the car trip had HFoV $\approx 59.5°$ and VFoV $\approx 49.6°$. Since the GPS signal samples were obtained only once or twice per second, while the camera captured 24 frames per second, we used a linear interpolation described in (8) to estimate the GPS position for every frame. We manually measured the GPS coordinates of the road signs using a mobile phone application and the Google Street View interface.

Let t_i and t_j, such that $0 \leq t_i < t_j$, be timestamps of two frames for which a known GPS measurement is denoted as $\mathbf{g}_i^T = [x_i, y_i]$ and $\mathbf{g}_j^T = [x_j, y_j]$, respectively. Let t_k be the timestamp for a frame that has no GPS measurement assigned, where $t_i < t_k < t_j$. For $\mathbf{g}_k^T = [x_k, y_k]$ the linear interpolation [8] is computed as

$$\mathbf{g}_k = \frac{t_j - t_k}{t_j - t_i}\mathbf{g}_i + \frac{t_k - t_i}{t_j - t_i}\mathbf{g}_j. \tag{8}$$

We converted the latitude ϕ and the longitude λ to earth-centered, earth-fixed (ECEF) coordinates as

$$x = r\cos(\phi)\cos(\lambda), \quad y = r\cos(\phi)\sin(\lambda),$$
$$z = r\sin(\phi),$$

with r representing the radius of the Earth equal to 6 370 km. The backward conversion is

$$\phi = \arcsin\left(\frac{z}{r}\right), \quad \lambda = \operatorname{atan2}(y, x).$$

B. Evaluation of Methods

1) Error Computation: We used the l_2 distance to quantify the deviation of the estimated position from the real object position. As mentioned in Section I, we calculate only x and y coordinates. We are not concerned with the object elevation. For pair position estimate $\mathbf{p}_p$ and ground truth position $\mathbf{p}_{gt}$ the error of position estimation is calculated as

$$e(\mathbf{p}_p, \mathbf{p}_{gt}) = \|\mathbf{p}_p - \mathbf{p}_{gt}\|_2. \tag{9}$$

2) Centroid-Based Position Estimation: We evaluated the position estimation using the localization error as defined in (9). Statistics on synthetic data show that the LSM method (III-C) outperforms the method LIM (III-B) on average by 4.93% as measured by the localization error in (9). However, on real data, the outcome is reversed, and the LIM method produces a better estimate of 0.67% on average. Table II shows more results of pair estimates on individual datasets, i.e., statistics of each estimate based on unique frame pairs.

Table II
STATISTICS OF PAIR POSITION ESTIMATES ERRORS ON ALL DATASETS. VALUES ARE IN METERS.

Dataset	Method	Min	Max	Mean	Median	Stdev
flat	LIM	0.185	2.488	0.789	0.761	0.215
	LSM	0.185	2.488	0.790	0.762	0.216
hill	LIM	0.108	2.125	0.646	0.640	0.173
	LSM	0.151	1.996	0.519	0.499	0.174
s-left	LIM	0.493	3.296	1.364	1.359	0.228
	LSM	0.400	3.283	1.329	1.324	0.233
s-right	LIM	0.005	4.408	0.458	0.389	0.405
	LSM	0.011	4.408	0.459	0.390	0.404
up	LIM	0.083	2.828	0.637	0.614	0.212
	LSM	0.076	2.794	0.606	0.580	0.211
town	LIM	0.149	12.039	4.382	4.282	1.813
	LSM	0.072	16.542	4.411	4.238	2.062

As long as centroid is used to produce the refined position estimate, Table II can be used to assess its performance, too. The "Mean" column indirectly shows the accuracy achieved by the proposed refining method from Section III-D2.

3) Cluster-Based Position Estimate: In this section, we will discuss our attempt to improve the global position estimation. Based on the fact that multiple position estimates are available, we could exploit their statistical properties to derive an even better position estimate. We believe that there is a potential to improve the proposed centroid-based method from (7). To further improve the precision of position estimation we tried our custom clustering-based approach. The purpose was to find a cluster of concentrated points with a minimal variance of l_2 distance between each pair of points. The assumption was that the majority of position estimates land within the vicinity of the real object position while the outliers land further away (Fig. 3). We tried to find such a cluster and then use its centroid to derive the refined estimate.

We found that as long as the filtering condition defined in (6) is employed, our methods are not excessively susceptible to the presence of outliers. We can see this effect in Fig. 3. The computations are more prone to have a high bias than a high variance. We conclude that our preliminary attempts generally either did not improve the centroid-based position estimate substantially or made it worse. The reason for deteriorating performance is the presence of systematic errors, such as the influence of low resolution. The filtering condition from (6) provides a simple yet sufficient approach to eliminate the majority of outliers. We believe that to obtain a significant improvement even more sophisticated approaches would have to be employed. This is left to future research.

4) Distance Influence: We tested the influence of distance by estimating the position using only those frames where the distance of the camera from the object of interest was within a specific distance in meters. Fig. 5 shows that the localization error depends on the distance. The effect of distance on the real dataset is expected, given the presence of noise. The mean and median are approximately the same which means that not many outliers are present.

Table III
STATISTICS OF ERRORS OF PAIR ESTIMATES ON SYNTHETIC DATA USING DIFFERENT METHODS IN PRESENCE OF NOISE. NOISE CAN BE PRESENT IN EITHER CAMERA AZIMUTH ("AZ"), OBJECT DETECTION BBOX ("BBOX") OR CAMERA POSITION ("CAM"). VALUES ARE IN METERS.

Noise	Method	Min	Max	Mean	Median	Stdev
az	LIM	0.029	13721.448	29.588	5.471	275.617
	LSM	0.030	45880.039	49.217	5.581	814.591
bbox	LIM	0.020	20336.768	16.758	2.118	330.448
	LSM	0.025	1501.049	9.769	2.310	46.313
cam	LIM	0.009	403.366	9.837	3.781	20.318
	LSM	0.0222	501.444	11.046	4.136	24.756

5) Noise Influence: We utilized a noise with a normal probability distribution. We emphasize that we studied the effect of noise in general and did not focus on a specific combination of road sign detection and localization methods. When we applied the noise, only one of the parameters (either object BBOX, camera position, or azimuth) was modified. The remaining ones were fixed. We wanted to measure the influence of noise in isolation. The mean value of the distribution was always set to the value currently being modified. The standard deviation changed regarding which parameter the noise was applied to. For example, when we applied the noise to the BBOX, the shift was 10 pixels for x and y coordinates (width and height remained unchanged). In the case of camera azimuth, the change was 1°. Camera position deviated by 0.25 m in x and y direction. We subjectively chose these values as we considered them a reasonable deviation. We are aware that more extensive noise evaluation could be performed to assess the absolute influence of different types of noise.

Table III shows that a deviation of the camera azimuth impairs the accuracy the most. The reason is that even a slight angle deviation has a considerable effect on the triangulation. Based on these results, the LIM method seems to be more robust to fluctuations in the camera azimuth than the LSM method. Conversely, the LSM method seems to handle variation in the BBOX position better than the LIM method. In the case of a noisy camera position, the impact is relatively similar for both methods.

V. CONCLUSION AND FUTURE WORK

In this work, we developed, implemented, and tested two different methods for object position estimation that can operate using just a single moving camera. We tested our methods in the laboratory as well as real-world conditions on a video sequence containing our object of interest, the road sign. We measured the position estimation error as the distance between the estimated and the real position. Tests on synthetic, ideal data, showed that the minimal error of position estimation is 30 cm, average 70 cm, and median 60 cm for both of our methods. However, the accuracy is directly influenced by the distance between the object and the camera together with the presence of noise. Our experiments with the change in the distance on the real dataset (see Fig. 5) show that the minimal error is around 60 cm with average and median being around

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

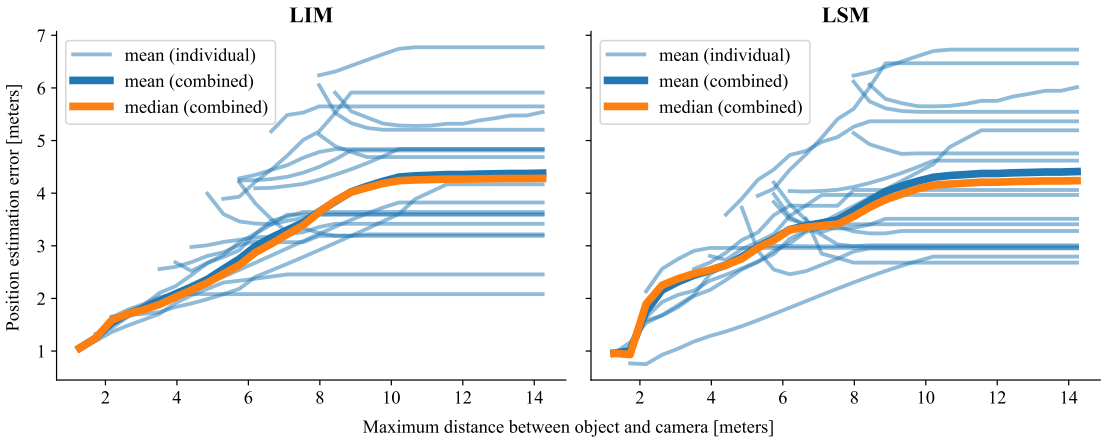


Fig. 5. The influence of the change in the distance of the camera from the object of interest on localization error for the real dataset. Both developed methods are practically identical in terms of their position estimation accuracy. We showed estimates for individual objects by transparent thin lines, and the combined performance with thick lines. The plots are cumulative so all the pair estimates up to the given distance threshold are taken into account.

4 m. As far as the noise is concerned, we also showed that our methods are sensitive to noise, but not equally (Table III). We did not aim to measure the exact influence of each type of noise. Our goal was to emphasize that the presence of noise in different forms may have different effects on our methods.

Experiments with artificial noise pointed to relatively high sensitivity to noise, primarily in the azimuth angle and object detection BBOX. Real-world application on multiple objects placed in terrain under various conditions showed that minimal error is around 60 cm with average and median being around 4 m. Assume a video recording using an average mobile phone device attached to a vehicle dashboard. Additionally, assume that the GPS information about the camera position is obtained with a time delay of approximately 1 second. A potential improvement is to use more frequent and more accurate information about the camera position. If it is not possible, then at least support its occasional absence with an interpolation by a polynomial of a higher degree than just one.

We managed, to a certain extent, to circumvent the obstacles of missing depth information from just one view. The approach was to emulate the stereo vision process using two views of the same object from two different camera positions. With this in mind, our computationally inexpensive solution is, therefore, applicable under financial and performance constraints. Furthermore, no setting up of parameters is required. The only parameter for our methods is epsilon that is used in (6). We fixed this value on the synthetic data and used it for the evaluation of the real data. Thus we avoided potential overfitting. This substantially broadens the potential for utilization, since an average mobile phone nowadays provides video/image information of sufficient quality.

A. Future Work

Our models are generally applicable, but if we specialize in a certain task, we can exploit its unique properties. For example, for traffic sign positioning, we could expect traffic signs to be on the right side of the road. Additionally, as time goes by, the camera should come closer to the road sign. This could aid in the filtering of estimates that do not meet these criteria.

Even though we did not compute the elevation of the object (z-axis), we remark that our preliminary experiments showed that the LIM method (III-B) could be used to compute a 3D position. The intersection of the two imaginary rays would provide the third coordinate, too.

We found that both of the proposed methods perform similarly well. However, each method has its strengths and weaknesses that do not necessarily have an identical influence on the performance, as shown in experiments with noise in Table III. A method to join the two estimates could be devised.

We admit that the real data contain inaccuracies in the ground truth position despite our endeavors to avoid them. To better assess the performance of our methods, an evaluation using a dataset measured more accurately would be ideal.

Acknowledgment

This work was financially supported by the grant VEGA 1/0689/19.

References

[1] Asra Aslam, Mohd Ansari, et al. Depth-map generation using pixel matching in stereoscopic pair of images. *arXiv preprint arXiv:1902.03471*, 2019.

[2] Mohamed Bénallal and Jean Meunier. A simple algorithm for object location from a single image without camera calibration. In *International Conference on Computational Science and Its Applications*, pages 99–104. Springer, 2003.

[3] Viral H Borisagar and Mukesh A Zaveri. Disparity map generation from illumination variant stereo images using efficient hierarchical dynamic programming. *The Scientific World Journal*, 2014, 2014.

[4] Hemang Chawla, Matti Jukola, Elahe Arani, and Bahram Zonooz. Monocular vision based crowdsourced 3d traffic sign positioning with unknown camera intrinsics and distortion coefficients. *arXiv preprint arXiv:2007.04592*, 2020.

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[6] Pavel Davidson, Mostafa Mansour, Oleg Stepanov, and Robert Piché. Depth estimation from motion parallax: Experimental evaluation. In *2019 26th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS)*, pages 1–5. IEEE, 2019.

[7] Kenneth M Dawson-Howe and David Vernon. Simple pinhole camera calibration. *International Journal of Imaging Systems and Technology*, 5(1):1–6, 1994.

[8] Pedro Gil-Jiménez, Hilario Gómez-Moreno, Roberto López-Sastre, and Saturnino Maldonado-Bascón. Geometric bounding box interpolation: an alternative for efficient video annotation. *EURASIP Journal on Image and Video Processing*, 2016(1):8, 2016.

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[10] Jen-Yu Han, Tsung-Hsien Juan, and Tzu-Yi Chuang. Traffic sign detection and positioning based on monocular camera. *Journal of the Chinese Institute of Engineers*, 42(8):757–769, 2019.

[11] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005.

[12] Clemens Holzmann and Matthias Hochgatterer. Measuring distance with mobile phones using single-camera stereo vision. In *2012 32nd International Conference on Distributed Computing Systems Workshops*, pages 88–93. IEEE, 2012.

[13] Ibrar Ullah Jan and Naeem Iqbal. A new technique for geometry based visual depth estimation for uncalibrated camera. In *2009 International Conference on Emerging Technologies*, pages 213–218. IEEE, 2009.

[14] Insu Kim and Kin Choong Yow. Object location estimation from a single flying camera. *UBICOMM 2015*, page 95, 2015.

[15] Jun Seok Lee and Duk Geun Yun. The road traffic sign recognition and automatic positioning for road facility management. *International Journal of Highway Engineering*, 15(1):155–161, 2013.

[16] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5162–5170, 2015.

[17] Andrew J Parker, Jackson ET Smith, and Kristine Krug. Neural architectures for stereo vision. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1697):20150261, 2016.

[18] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[19] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Depth perception from a single still image. In *AAAI*, volume 3, pages 1571–1576, 2008.

[20] Victor JD Tsai and Chun-Ting Chang. Three-dimensional positioning from google street view panoramas. *IET Image Processing*, 7(3):229–239, 2013.

[21] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

VI. Appendix

A. Derivation of the Lines Intersection Method

See Section III-B for details. We assumed the two sought lines have to pass through the target point $\mathbf{p}_t^T = [x_t, y_t]$, thus

$$y_t = m_1 x_t + b_1, \quad y_t = m_2 x_t + b_2.$$

Given the two known points $\mathbf{p}_1^T = [x_1, y_1]$ and $\mathbf{p}_2^T = [x_2, y_2]$, our lines are then expressed as

$$\begin{aligned} y_1 = m_1 x_1 + b_1 \quad &\rightarrow \quad b_1 = y_1 - m_1 x_1, \\ y_2 = m_2 x_2 + b_2 \quad &\rightarrow \quad b_2 = y_2 - m_2 x_2. \end{aligned}$$

The core relationship (the intersection at the same point y_t) is

$$m_1 x_t + b_1 = m_2 x_t + b_2,$$

from which we derive equation (1) as follows

$$\begin{aligned} m_1 x_t + (y_1 - m_1 x_1) &= m_2 x_t + (y_2 - m_2 x_2), \\ x_t &= \frac{y_2 - y_1 + m_1 x_1 - m_2 x_2}{m_1 - m_2}. \end{aligned}$$

The slopes m_1 and m_2 are approximated by the tangens of pixel azimuths given by $\overline{\alpha_1}$ and $\overline{\alpha_2}$, respectively, so

$$x_t = \frac{y_2 - y_1 + \tan(\overline{\alpha_1})\, x_1 - \tan(\overline{\alpha_2})\, x_2}{\tan(\overline{\alpha_1}) - \tan(\overline{\alpha_2})}.$$

The y_t coordinate given by equation (2) can be expressed as

$$y_t = m_1 x_t + (y_1 - m_1 x_1) = \tan(\overline{\alpha_1})\, x_t + y_1 - \tan(\overline{\alpha_1})\, x_1.$$

B. Derivation of the Law of Sines Method

We advise to see Section III-C and Fig. 2 for details. We substitute $A = \mathbf{p}_1$, $B = \mathbf{p}_2$ and $C = \mathbf{p}_t$. Let the inner angles of the triangle be $\alpha = \angle BAC$, $\beta = \angle ABC$, $\gamma = \angle ACB$, and let a, b, c be their opposite sides, respectively. These angles are computed as the angles between two lines using (3) as

$$\begin{aligned} \alpha &= \rho(m_0, m_1), \\ \beta &= \pi - \rho(m_0, m_2), \\ \gamma &= \pi - (\alpha + \beta) = \rho(m_0, m_2) - \rho(m_0, m_1). \end{aligned}$$

Now we have to compute the distance we need to move from the point $\mathbf{p}_1$ in the direction of $\overline{\alpha_1}$, denoted by Δs, to obtain the target object position $\mathbf{p}_t$. Let the moved distance between frames be $\Delta d = \|\mathbf{p}_1, \mathbf{p}_2\|_2$. Using the law of sines, we get

$$\frac{b}{\sin(\beta)} = \frac{c}{\sin(\gamma)} \quad \rightarrow \quad \frac{\Delta s}{\sin(\beta)} = \frac{\Delta d}{\sin(\gamma)},$$

from which we express Δs as

$$\Delta s = \frac{\sin(\beta) \cdot \Delta d}{\sin(\gamma)} = \frac{\sin(\pi - \rho(m_0, m_2)) \cdot \|\mathbf{p}_1, \mathbf{p}_2\|_2}{\sin(\rho(m_0, m_2) - \rho(m_0, m_1))}.$$

Finally, we compute the target object position $\mathbf{p}_t$ (equations (4) and (5)) using

$$\begin{aligned} x_t &= x_1 + \Delta s \cdot \cos(m_1), \\ y_t &= y_1 + \Delta s \cdot \sin(m_1). \end{aligned}$$

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Analysis of counterfeits using color models

Definition of the color of painting in digital image

Irena Drofova
Tomas Bata University in Zlin
Faculty of Applied Informatics
Department of Security Engineering
Zlin, Czech Republic
drofova@utb.cz

Milan Adamek
Tomas Bata University in Zlin
Faculty of Applied Informatics
Department of Security Engineering
Zlin, Czech Republic
adamek@utb.cz

***Abstract*— The black market for counterfeit works of art is the third most lucrative in the world. The digitalization of art and introduction into the virtual environment allows for a high rate of these counterfeit works. The issue of verifying original works in the online environment currently brings new challenges in the field of forensic science, in particular fields of color vision and working with light, colors, and color models, thus opening new approaches of art verification and presentation. This article deals with image processing and digitalization in color fields and their definition in color spaces. A real artwork is analyzed by graphic software to determine a color model of a digitally processed image and for a specific color used. This text aims to find and define the same color tone with the actual color tone of the original artwork in the digital environment and color spaces.**

Keywords—forensic art; counterfeits; digital image; color model; virtual environment

I. INTRODUCTION

Forensic science is a multidisciplinary field that is an integral part of criminology, and it plays an important role in investigating crime and detecting potential perpetrators, primarily because the information and evidence that individual fields of forensic science can provide are based on detailed research. Crimes are reconstructed; crimes committed are detected; the personalities of suspects are analyzed; the evidence is presented in court proceedings. [1] One of the disciplines of forensic science is forensic art, which significantly contributes, among other things, to the possible disclosure of a perpetrator's identity. The appearance of a crime can be detected through the reconstruction of evidence fragments, evidence demonstration, victim identification (in case of an accident), or body decay. At last but not least forensic art can be used in the detection of counterfeit products, counterfeit artworks, and other valuables. Especially in the art field, the gradual advent of digital technologies has created new challenges in this area. New methodologies and procedures for the detection of counterfeit works have been developed. Digital image diagnostics are used to evaluate the attributes of artworks. Computer imaging technologies can obtain evidence of possible manipulation or various information about the origin of the artwork. The possibility of applying modern forensic algorithms to the original painter's work for a diagnostic purpose can help detect and document the potential forgery of an artwork. [2] The use of the theories of light, colors, and color models can complement the detection and documentation of some attributes of artwork. Achieving the most exact representation of colors and structures of artwork depends on the fidelity of the digital reproduction of the image. This article deals with the possibilities of classification and documentation of the original artwork in terms of colors used in combination with color models and areas for the most accurate reproduction of the image in a digital environment. [3] Experimental scanning and digitalization of an object are performed in order to identify the physical color in a digital environment and define the values of the color. A clearly defined color tone value is necessary for other color applications, especially in other color spaces or in subsequent artistic reproduction printing techniques, for digital archiving of artworks and reuse of color values in further digital image processing. In the case of precise capture of artworks for further processing, evaluation of a definite attribute, and digital archives, it is desirable to have a quality device with an appropriate optical system for capturing the whole and partial specific parts.

II. METHODS

Counterfeiting artworks are simple, and often even the author of an artwork cannot distinguish between counterfeit and original. Counterfeits often appear in the depositories of museums and galleries. Currently, there are modern and complementary methods for distinguishing counterfeits from the original. Options and methods for analyzing the original image may be combined. The choice of the methods also depends on the type of the painting and the purpose of reproduction.

A. Visual Cues Method

The initial *macroscopic examination* of the original work evaluates the features that are visible to the human eye. Differences in color tone and painting style may be evident in an artwork. We can also see, for example, the effort to obtain a credible patina that damaged the layer of painting, as well as the canvas to obtain the necessary old look of the artwork. *This stratigraphic method* can be used to evaluate used color pigment as well as the age of drawing. Some pigment can be identified by *polarizing microscopy* and scanning *electron microscopy* (SEM) or *electron microscopy* in conjunction with

International Grand Agency of Thomas Bata University in Zlin, IGA/CebiaTech/2021/004, and the Department of Security Engineering, Faculty of Applied Informatics.

an *electron dispersion spectrometer* (EDS). The pigment can also be identified by a *microchemical test*.

The style of visual art is usually determined after the creation of the artwork. The style should be significantly different from other styles. Styles can overlap or complement each other. An artist can use more styles in his artwork. There is no specific definition of a visual style. However, we can apply some visual cues from paintings, such as color palette, scene, composition, lighting, contour, and brush strokes. [4]

B. RGB to HSL Color Model Conversion

The RGB and CMYK color models are two models used in computer graphics and are directly related. Their abbreviation represents the primary colors; by mixing, we achieve other colors of the color spectrum. These are the example of the primary and secondary colors and their complementary. The RGB and CMYK color models are shown in Figure 1.

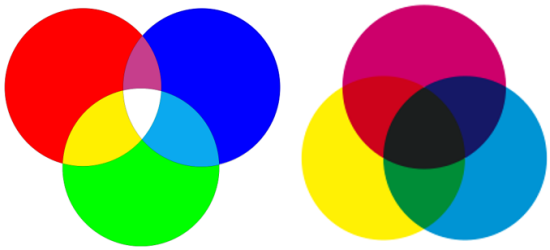


Fig. 1 RGB color model / CMYK color model

All-digital display devices (such as scanners and monitors) using the RGB (red/green/blue) model, while the CMYK (cyan/magenta/yellow/contrast) color space works with real colors and outputs those colors to different materials. For this reason, it is necessary to recalculate the colors of the individual rooms according to the nature of the input and output information to minimize the loss of color information or to make the resulting display as faithful as possible to the original input information. [4]

The HSL color model is very similar to the RGB model but works with different values. This model describes color relationships more accurately and is easier to work with. HSL color model works separately with tones and corresponds better to the color perception of the human eye. The HSL color model can provide other valuable information such as color grade, color saturation, and color brightness. [5] The advantage is a smooth change of tone. This color model is a favorite in digital graphic art. The color perceived by the human eye depends on the external environment and reaction to changes in it. Mainly lighting conditions significantly affect the resulting color perception.

C. Color Measuring Equipment

There are more or less accurate methods of measuring color. The following instruments are used for the measurement of color.

- *The photometer* is mainly used for display devices such as monitors.
- *The fluorometer* (fluorescence spectrometer) is used to measure photoluminescence.
- *The colorimeter* is used for direct measurement of color coordinates or *xy* from *trichromatic values*. This instrument is more accurate than the densitometer below and checks precisely specified and specific colors.
- *The spectrophotometer* is practically the most accurate of all the above instruments. It measures the spectral characteristics of direct or reflected light. With this device, it is also possible to find out the change in hue color depending on the type of lighting used.
- *The densitometer* measures the optical density (density) both in reflection and in the passage of light. These devices are especially suitable for determining the color thickness of layers. The densitometer determines:
  - Transparency (T), i.e., material permeability
  - Opacity (O), i.e., is impermeability rate. It is inverted transparency [6]

Density calculation for individual CMYK process colors using image analysis:

$$D_{cyan} = \log_{10} \frac{I_W}{I_R} \tag{1}$$

$$D_{magenta} = \log_{10} \frac{I_W}{I_G} \tag{2}$$

$$D_{yellow} = \log_{10} \frac{I_W}{I_B} \tag{3}$$

$$D_{black} = \log_{10} \frac{I_W}{(I_R + I_G + I_B)/3} \tag{4}$$

where D_{cyan}, $D_{magenta}$, D_{yellow}, D_{black} are the optical densities of the individual process colors, I_W is the intensity of the comparative white, I_R is the intensity of the pixel red, I_G the intensity of the pixel green, I_B the intensity of the pixel blue.

The experimental study focused on specific artwork and worked with colors and color models to obtain plausible color reproduction in a digital environment. The artistic image belongs to a serial of paintings called Contrasts. The author of the picture, Michal Pasma, likes to use many different styles in his artworks.

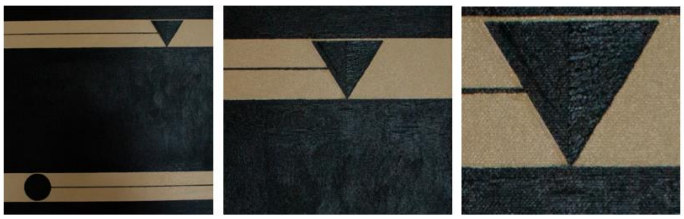


Fig. 2 Structure of painting acrylic on canvas / structure detail enlarged +100% / structure in detail enlarged +200%

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The artwork was a painting with the acrylic technique on canvas, where one color is known. The material structure, stroke direction, and length are captured, as shown by the enlargement of the image in Figure 2. The color measuring devices are not used in the experimental study.

III. IMAGE CAPTURE AN ORIGINAL ARTWORK

An artistic painting chosen for the experimental work was an artwork painted with acrylic paints on canvas. Three direct acrylic paints from Koh-i-Noor have been used: black 0700, white 0100, ocher 0600. The focus was only on the black paint, shown in Figure 3. [7]

Fig. 3 Artwork on canvas painted with Koh-i-Noor acrylic paint black 0700 [7]

At the time, it was not possible to precisely define the values and characteristics of the light in the room. This artistic acrylic paint for painting is dilutable with water and fixed to the substrate. It is a high-quality dispersion paint with a high number of pigments and colorfastness. The paints can be used on any non-greasy substrate and, after drying, create an age-resistant color film. From the three acrylic colors mentioned, only black was used for the experiments because the other two colors were mixed from other color shades in the artwork. The black color can be theoretically related to the CMYK color model as a direct color.

The art object was photographed in not very good conditions, at dusk, with the spotlight at the space, without the proper lighting of the object. There was no color reference scale for the use of a DNG profile or unification. As expected, the scanned image had to be calibrated manually in a graphic program. The subject was captured with a Pentax K-50 mirror camera with the basic lens without the use of automatic or external flash. The properties of the reference image and the scanning device are shown in Table I.

TABLE I. PROPERTIES OF THE REFERENCE IMAGE

Image size	19,3 MB
Image dimensions	4928 × 3264 px
Resolution	96 dpi
Bit depth	24 bit
Color range	sRGB
Aperture shuter	f/5
Exposure distance	1/200 s
Focal distance	35 mm
ISO	177

An important aspect when working with color and image digitization is the color calibration of input and output devices. This was used to define them and create their color scales to further refine colors and tones similar to original images. For these purposes, it is easiest to use a standardized color combination, and each sub-photo with the DNG (digital negative) profile needs to be created. It is reusable between sets of photos and between cameras. RAW files used to archive images may generate digital cameras differently, and specifications are not always available. DNG format is suitable for editing, simple archiving, and providing easy access to archived files. The DNG format is supported by software vendors and used by manufacturers of cameras and 3D imaging devices. The classic standardized reference scale with 24 color fields for creating a DNG profile is used nowadays. The standardized reference color scale is a part of the chromatic triangle (chromaticity diagram) *xy* CIE 1931, as shown in Figure 4. However, this color scale was not available when capturing the image. It was incorporated by subsequent image processing and onerous editing. Image processing, calibration, and conversion of colors and profiles were thus made only in graphics programs.

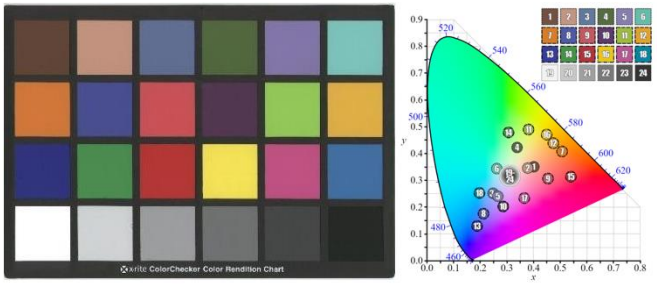


Fig. 4 Scale of 24 color fields and their position in the chromatic diagram

This article is focused on specific works of art and artworks with colors, color models, and spaces to obtain reliable reproduction and definition of colors in a digital environment. When an image of an object is captured under less favorable conditions, without proper illumination, a calibrated scanning device, or a basic standardized color gamut, it provokes several other questions about the image in terms of color processing. Above all, how to define black color paint in a digital environment. The following chapter describes color calibration, working with color models and spaces to approximate the tone of color spots and define the color and its properties in digital form.

IV. DIGITAL PROCESSING OF THE SCANNED IMAGE

The selected artwork contains contrasting transitions and geometric elements. It was possible to work better with the image and the primary color. The light color on the paintings was mixed with ocher and white for the final shade. The painter used three colors. The color black was chosen for the experiment. This color could be theoretically related to the CMYK model as a primary color.

The original RAW format was loaded into the Camera Raw 13.1 [8] software workspace and then made basic adjustments to the digital image. Basic adjustment and color calibrations

were adjusted in the RGB model, as well as the hue and saturation. Subsequently, the curve of the light and shadow values was corrected. This process is shown in Figure 5.

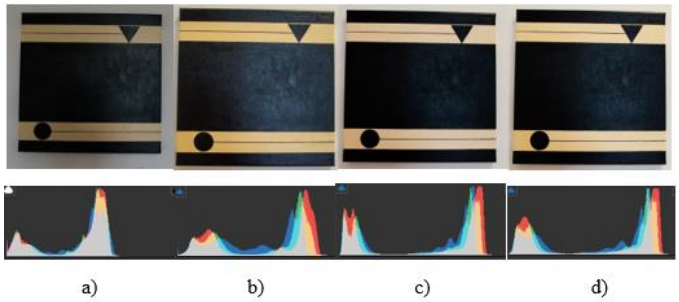


Fig. 5 Images and histograms after individual adjustments of the digitized image: a) of the original image in the RAW format before the adjustments; b) basic adjustments of the digital image. The basic parameters values adjusted from the initial values: hue; exposure; contrast, lights; shadows; and saturation; c) color calibration R, G, B; d) adjustment of the image curve values: light; shadows; white and dark values.

After basic adjustments, there were further image adjustments with emphasis on color transitions and color mixers in connection with the color model HSL. This color model is the most widely used in the digital graphics environment. Its advantage is the smooth transition of colors in tones compared to a similar model HSV. Further image adjustments are shown in Figure 6.

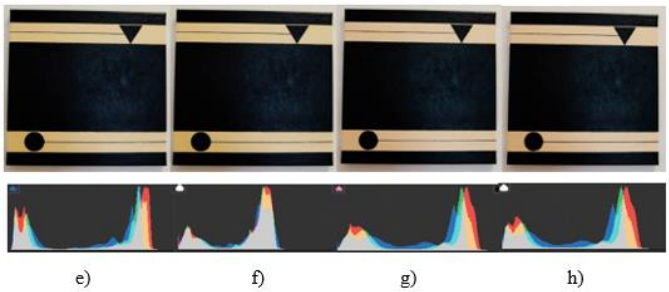


Fig. 6 Images and histograms after further adjustments of the digital image: e) adjustment of the amount and hue of the purple and green components in digital image reproduction; f) adjustment of color transitions in mid-tones, shadows, and light; g) color mixing and adjustment of individual colors RGB, CMYK, purple and orange to the HSL model; h) adjustment of color transitions in the HSL model.

After these adjustments, the digital image was evaluated and was recalculated as the object into the Adobe Photoshop [8] work environment. This image was center by rotating the canvas 2 degrees to the left. Then the image was cropped to clean the background. In this case, it was walls. Further calibration of the RGB color values was applied, and the area with the highest possible black value was selected. The goal was to find the same shade of black that matches the color from this place. There were defined the color value of the painting's black color in the CMYK color model. The hue of the corresponding color was select from many color scales and spaces that the graphic program offers.

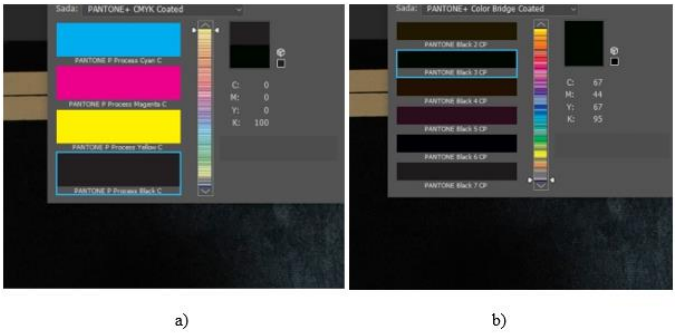


Fig. 7 Comparison of the color Pantone scale with the color in the image area: a) the color Pantone Process Black C; b) the color Pantone Black 3 CP

Figure 7 shows a comparison of two colors from the Pantone scales [9]. The first is the color Pantone Process Black C. This color is a process color from the colors of the CMYK Coated swatch. The second color is Pantone Black 3 CP from the Pantone Color Bridge Coated color range. This color was the same as the black tone on the image area. In the CMYK color model, it corresponds to the color values C = 67; M = 44; Y = 67; K = 95. Figure 8 shows both shades of black from the Pantone scales. The difference between them is obvious to the human eye.

Fig. 8 Comparison of the color Pantone Process Black C and the color Pantone Black 3 CP [9]

Table II gives the exact numerical definition of the black tone on artwork in color models.

TABLE II. NUMERICAL DEFINITION OF THE BLACK TONE

Color model	Black color numerical definition
Hex	#262d26
RGB	rgb(38, 45, 38)
CMYK	cmyk(67, 44, 67, 95)
HSL	hsl(120,8.43%, 16, 27%)
RGBa	rgba(38, 45, 38, 1.00)
Lab Xyz	xyz(2.0876, 2.4288, 2.1927)
HSV	hsv(120,15.56%, 17, 65%)
HSVa	hsva(120,15.56%, 17.65%, 1)

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

V. CONCLUSION

Colors and color models are very narrow and have specific issues. In the case of artworks and counterfeits, colors play an important role. Their chemical and optical properties, as well as other properties, significantly varies not only in terms of time but also environmental influence on color sustainability. In the case of original artwork, its quality in digital reproduction can not only be preserved but its artistic value can also be transferred to other presentation channels and thus reach people all over the world. In terms of color reproduction, this is the most faithful approximation to a reused dye in a physical environment. There are many ways to achieve this, not only in terms of new uses but also for defining individual colors and outputs in different channels. In this article, an experiment was demonstrated concerning a process of color definition for its further use in a digital graphic environment. The subject of the research was the black acrylic paint Koh-I-Noor 007. The original painting was painted using the acrylic technique on canvas. This paint is characterized by a high number of pigments, and its coating is fixed. The experiment included transferring artwork to an online environment, adjusting and calibrating colors, and further eliminating other aspects that have arisen in image capture, where there are no other high-level professional scanning techniques, no other accessories for working with colors and their properties. As can be observed, working with color and color spaces, and mixing channels thus requires considerable patience and time. The above-mentioned color was defined specifically in the form of a shade in the Pantone color swatch as Pantone Black 3 CP, and then it was specified for use in other color spaces. The color defined in this way can serve as a basis for a faithful printed reproduction or further reproduction of the original in an online environment or in virtual reality with a clear specificity. In the case of two more colors, ocher and white, from the same manufacturer, the most appropriate methodology would probably be to perform direct measurements by instruments designed for this purpose and presented in this article. The main focus of this research was on using spectrophotometer or densitometer as the two most important methods concerning debated issue.

ACKNOWLEDGMENT

This research was based on the support of the author of the painting, Mr. Michal Pasma, and of the International Grand Agency of Thomas Bata University in Zlin, IGA/CebiaTech/2021/004, and the Department of Security Engineering, Faculty of Applied Informatics.

REFERENCES

[1] J. A. Siegel and P. J. Saukko, Encyclopedia of Forensic Sciences, Academic Press, 2013, ISBN 978-0-12-382165-2.

[2] A. Pelagotti, A. Piva, F. Uccheddu, D. Shullani, M. F. Alberghina, S. Schiavone, E. Massa and C. M. Menchetti, Forensic Imaging for Art Diagnostics. What Evidence Should We Trust?, IOP Publishing Ltd, 2020, IOP Conf. Ser.: Mater. Sci. Eng. 949 012076, ISSN: 17578981, DOI: 10.1088/1757-899X/949/1/012076

[3] E. Gultebpe, E Thomas, and M. M. Conturo, Predicting and grouping digitized paintings by style using unsupervised feature learning, Journal of Cultural Heritage., 2018, vol. 31, 13-23., ISSN 1296-2074., DOI:10.1016/j.culher.2017.11.008

[4] Dohnal M., Barevne videni. Kolorimetrie, Univerzita Pardubice, 2019, ISBN: 978-80-7560-246-6.

[5] P. Ehkan, S. V. Siew, F. F. Zakaria, M. N. M. Warip and M. Z. Ilyas, Comparative Study of Parallelismand Pipelining of RGB to HSL Colour Space Conversion Architectureon FPGA, IOP Conf. Series: Materials Science and Engineering., 2020 DOI:10.1088/1757-899X/767/1/012054.

[6] M. Dohnal, Fyzikalni zaklady reprodukce obrazu, Univerzita Pardubice, 2007, ISBN: 978-80-7194-945-9.

[7] Koh-i-Noor, the official website 2021, https://www.koh-i-noor.cz/

[8] Adobe, the official website, 2021, https://www.adobe.com/

[9] Pantone, the official website, 2021, https://www.pantone.com/eu/en/color-finder

Time-Current Tripping Characteristics of RCDs for Sinusoidal Testing Current

Stanislaw Czapp
Faculty of Electrical and Control Engineering
Gdańsk University of Technology
Gdańsk, Poland
stanislaw.czapp@pg.edu.pl

***Abstract*—Low-voltage electrical installations are verified initially – before being put into operation, as well as periodically – during their utilization. According to the IEC standards, the scope of the verification includes measurements of both the tripping current and the disconnection time of residual current devices (RCDs). Experiences in RCDs testing show that disconnection times of two or more similar RCDs can be quite different. Significant differences in disconnection times are also noticed for the same RCD in the consecutive trials. This paper presents the result of the test of twenty-four RCDs. Their real tripping current, as well as disconnection time, have been verified. Differences in the obtained values of disconnection times are commented, and their possible sources are indicated.**

Keywords—protection against electric shock; RCDs; verification

I. INTRODUCTION

Effectiveness of protection against electric shock in low-voltage electrical installations depends, among others, on the proper operation of protection devices installed at the origin of circuits. One of the types of protective devices is a residual current device (RCD) which plays the role of a disconnecting device in case of an insulation fault or in case of direct contact with live conductor [1]. The most common RCDs have a rated residual operating current $I_{\Delta n}$ = 30 mA. Such a current is required by multi-part standard IEC/HD 60364 "Low-voltage electrical installations" for the purpose of additional protection in selected circuits (where a higher risk of electrocution exists).

Reliability of RCDs is relatively low, significantly lower than circuit-breakers as well as fuses. This is the effect of the RCDs' principle of operation, and properties of the components, which are used to perform their detecting-tripping internal system [2-6]. If an RCD is obliged to detect residual currents around 30 mA, and its operation type is voltage-independent, the detecting-tripping system must operate under extremely low power (derived from only 30 mA) [7]. It makes that the first tripping of an RCD, after its working in the closed position for a long time, may give the tripping current and the disconnection time higher than right after its first installation in a distribution board. Moreover, utilization of RCDs in areas of high humidity, dusty atmosphere or their exposure to vibrations, also influences this detecting-tripping system and may cause that RCDs real tripping current and disconnection time are different, mainly higher, from those set by manufacturers. If the above-mentioned tripping thresholds are higher than provided by the standards, it may result in making a decision about replacing the particular RCD by a new one.

On the other hand, in some cases, tripping current of the particular RCD, and especially its disconnection time, may vary during consecutive trials performed within the frame of verification of the electrical installation. It should not exclude the RCD from its further utilization, in spite of the relatively wide variation of the measurement results. This is due to the properties of the detecting-tripping system installed inside the RCD. Thus, it is very important to recognize and explain the behaviour of RCDs during their tests, especially in terms of the tripping current threshold and disconnection time.

In this paper, the results of the test of 24 RCDs, under sinusoidal testing current, are presented. Their real time-current tripping curves have been determined. Differences in values of their disconnection time, for consecutive trials, are commented and explained.

II. THE SCOPE AND RESULTS OF THE TEST

A. The Scope

International standard IEC 60364-6 [8] delivers the scope and rules for verifications of low-voltage electrical installations. This scope provides, among others, verification of the tripping current and disconnection times of RCDs. The requirement referring to the tripping current is fulfilled if the disconnection of the RCD occurs with a test current lower than or equal to the rated residual operating current $I_{\Delta n}$ (but higher than $0.5I_{\Delta n}$). Regarding the disconnection times – it is recommended that the times required by HD 60364-4-41 [1] are verified. In order to verify the disconnection times of RCDs in details, provisions of the standard IEC 61008-1 [9] are helpful. This standard delivers required time-current tripping characteristics of RCDs. Fig. 1 presents such a characteristic for general type (no-delayed) RCDs, with marked time-current points, required by the standard [9]. The disconnection times cannot exceed:

- 300 ms for the testing current equal to $I_{\Delta n}$,
- 150 ms for the testing current equal to $2I_{\Delta n}$,
- 40 ms for the testing current equal to $5I_{\Delta n}$.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

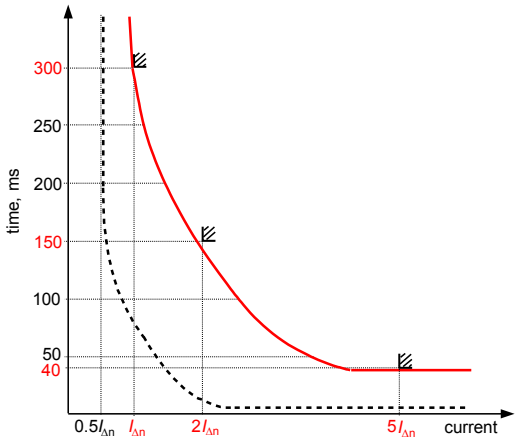


Figure 1. Time-current tripping curve (solid trace) of general type RCDs, according to [9].

The aim of the laboratory test is to check (for a group of 24 RCDs):

- real tripping current – it should be within the range (0.5–1.0)$I_{\Delta n}$,
- disconnection times for the testing currents: $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ – they should not exceed values represented by the red (solid) curve in Fig. 1,

and to compare the results with the required values, as well as to comment some discrepancies in these results. Testing of the above-mentioned parameters has been performed at the output terminals of the RCD, with the use of the professional RCD tester (Fig. 2).

Disconnection times were verified for two types of sinusoidal waveforms (Table I) because the initial phase of the testing waveform influences the threshold of the tripping values.

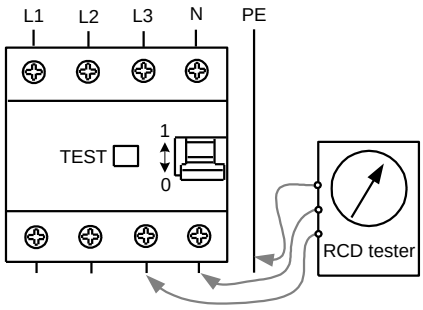
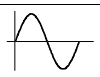
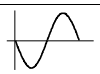


Figure 2. Connection of the RCD tester to the tested RCD.

TABLE I. TYPES OF THE TESTING WAVEFORMS

No.	Waveform	
	Math description	*Graphics*
1.	$i_1(t) = I_{m1}\cdot\sin(\omega t)$	
2.	$i_2(t) = I_{m2}\cdot\sin(\omega t+180^o)$	

B. Results of the Test

The laboratory test has been performed for a group of 24 RCDs from 8 manufactures. All the tested RCDs are of general type (no-delayed) and their rated residual operating current is equal to 30 mA. However, in terms of their sensitivity to the residual current waveform shapes, four types of RCDs have been tested:

1) AC-type (10 RCDs) – provided for sinusoidal (50/60 Hz) residual current only,
2) A-type (10 RCDs) – provided for both sinusoidal (50/60 Hz) and DC pulsating residual currents,
3) F-type (2 RCDs) – provided for sinusoidal (50/60 Hz) residual currents, DC pulsating residual currents, and mixed-frequency residual currents resulting from control equipment supplied from a single-phase,
4) B-type (2 RCDs) – provided for sinusoidal residual currents of frequency up to 1000 Hz, DC pulsating residual currents, smooth DC residual currents, and mixed-frequency residual currents.

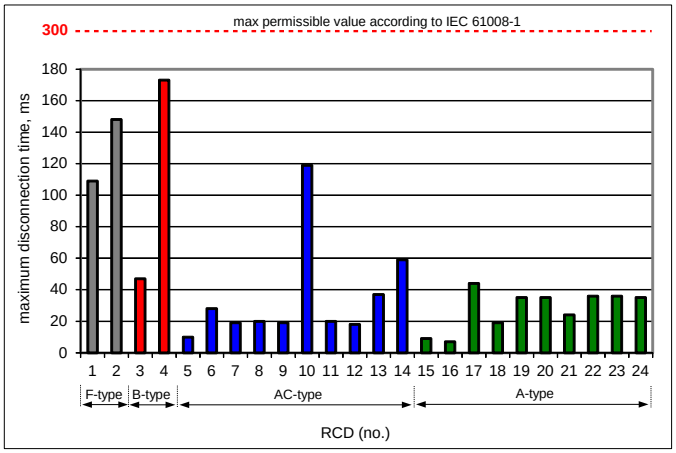
Table II presents a list of the tested RCDs and their basic parameters. In this table, real tripping current of the RCDs is included as well – all the results were repeatable. For sinusoidal testing waveform, this current should be within the aforementioned range (0.5–1.0)$I_{\Delta n}$, but one can see that RCD no. R16 does not fulfil this requirement – this RCD should not be further utilized. Its real tripping current is equal to 11.8 mA what is less than the lower threshold: 15 mA (0.5$I_{\Delta n}$).

TABLE II. BASIC PARAMETERS OF THE TESTED 30 MA RCDS

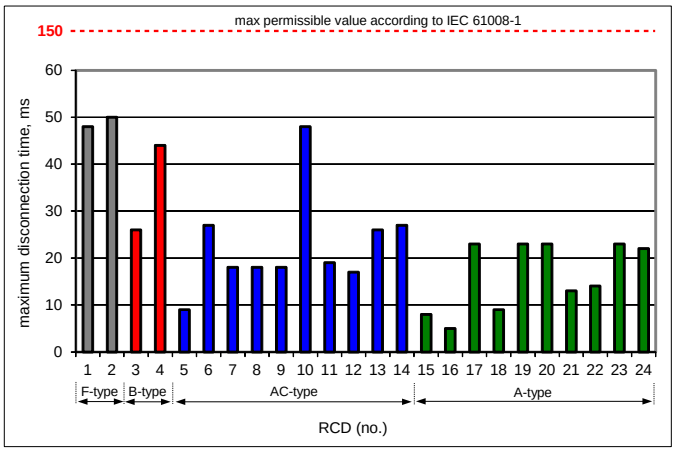
RCD no.	RCD type	Rated current	Number of poles	Real tripping current	Manu-facturer
		A	-	*mA*	
R1	F	40	2	20.4	M1
R2	F	40	2	21.9	M2
R3	B	63	4	21.9	M3
R4	B	40	4	23.3	M1
R5	AC	25	2	16.1	M4
R6	AC	16	2	24.8	M1
R7	AC	25	2	20.4	M5
R8	AC	40	4	21.9	M5
R9	AC	40	2	23.3	M5
R10	AC	40	4	21.9	M6
R11	AC	25	2	23.3	M5
R12	AC	25	2	23.3	M7
R13	AC	10	2	21.9	M8
R14	AC	16	2	29.1	M8
R15	A	25	2	21.9	M4
R16	A	16	2	*11.8*	M5
R17	A	25	2	20.4	M7
R18	A	25	4	23.3	M5
R19	A	25	2	23.3	M8
R20	A	25	2	26.2	M8
R21	A	6	2	16.1	M8
R22	A	6	2	24.8	M8
R23	A	6	2	24.8	M8
R24	A	25	2	26.2	M8

A more detailed analysis has been performed with reference to the verification of the disconnection times of RCDs. For each type of the testing waveform presented in Table I and for each value of the testing current ($I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$), every RCD has been tested 4 times (within intervals 1–2 s).

a)



b)



c)

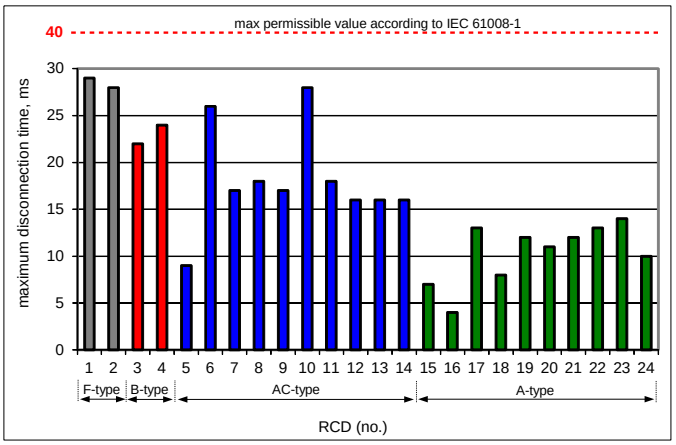


Figure 3. Maximum disconnection times (from 8 trials for each RCD) of the tested 24 RCDs of $I_{\Delta n}$ = 30 mA for testing current: a) $I_{\Delta n}$, b) $2I_{\Delta n}$, c) $5I_{\Delta n}$.

Fig. 3 presents maximum disconnection times (from 8 trials for each RCD and the testing current value: 4 for testing waveform no. 1 and 4 for testing waveform no. 2) of the tested RCDs. All the RCDs fulfil the requirements of the standard IEC 61008-1 [9] – their disconnection times do not exceed the permissible values. However, one can see that the spread of the measured values is very wide, in spite of the fact that all tested RCDs are the no-delayed type.

For the testing current equal to $I_{\Delta n}$ (Fig. 3a), the max measured value is 173 mA (RCD no. R4) and the min measured value is only 9 mA (RCD no. R15[1]). In case of the testing current $2I_{\Delta n}$ (Fig. 3b), it is 50 mA (RCD no. R2) and 8 mA (RCD no. R15[1]) respectively. The testing current $5I_{\Delta n}$ (Fig. 3c) gives the following max-min values: 29 mA (RCD no. R1) and 7 mA (RCD no. R15[1]) respectively. Note, that if the disconnection time of 30 mA RCDs is in compliance with IEC 61008-1 [9], the threshold of ventricular fibrillation in case of direct or indirect contact will not be exceeded (Fig. 4).

Very interesting results have been obtained in case of the comparison of the consecutive 4 trials for each type of the testing waveform and value of the testing current. Selected results – mainly the most varied – are presented in Fig. 5 to Fig. 12.

RCD no. R2 (Fig. 5) has very balanced values of the disconnection times for a specific value of the testing current – they reach the same value in each consecutive trial. Moreover, the time does not depend on the type of testing sinusoidal waveform. However, these times are significantly different if compare trails for $I_{\Delta n}$ with $2I_{\Delta n}$ and $5I_{\Delta n}$. Disconnection times are relatively high (even around 150 mA for testing current equal to $I_{\Delta n}$) in comparison to the other tested RCDs. Similar results have been obtained for RCD no. R4 (Fig. 6). Disconnection times are also relatively high and other conclusions are the same as for RCD no. R2.

RCD no. R6 (Fig. 7) has different behaviour than R2 and R4. Disconnection times are significantly lower than for RCD no. R2 and R4, but results between tests for $I_{\Delta n}$, $2I_{\Delta n}$ and $5I_{\Delta}$ differ only slightly.

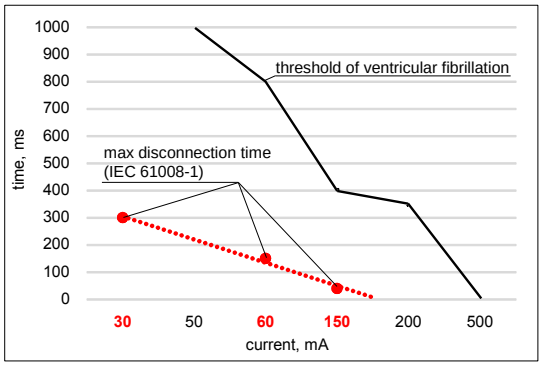


Figure 4. Comparison of the time-current tripping curve for 30 mA RCDs, required by standard IEC 61008-1 [9], with the threshold of ventricular fibrillation according to IEC 60479-1 [10].

[1] In fact, the lowest disconnection time has been achieved for RCD no. R16, however this RCD is not taken into account because its real tripping current (11.8 mA) is not within the normative range.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

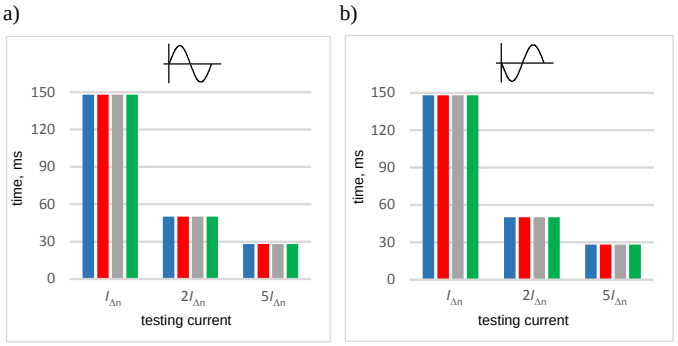


Figure 5. Disconnection times of RCD no. R2 (F-type, from manufacturer M2); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

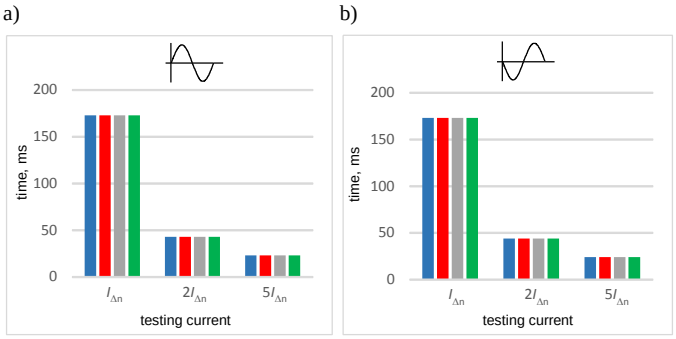


Figure 6. Disconnection times of RCD no. R4 (B-type, from manufacturer M1); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

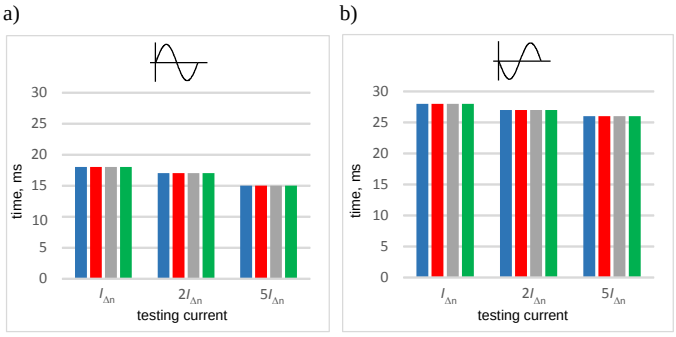


Figure 7. Disconnection times of RCD no. R6 (AC-type, from manufacturer M1); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

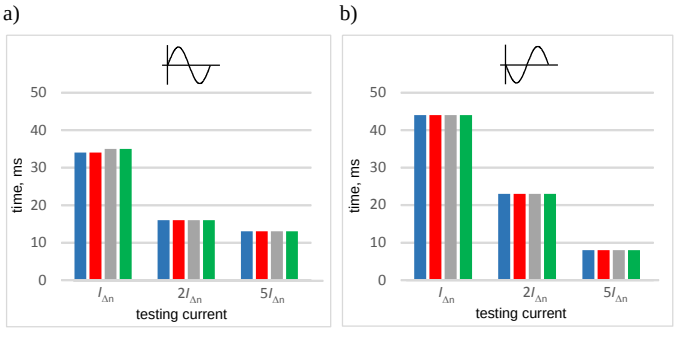


Figure 8. Disconnection times of RCD no. R17 (A-type, from manufacturer M7); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

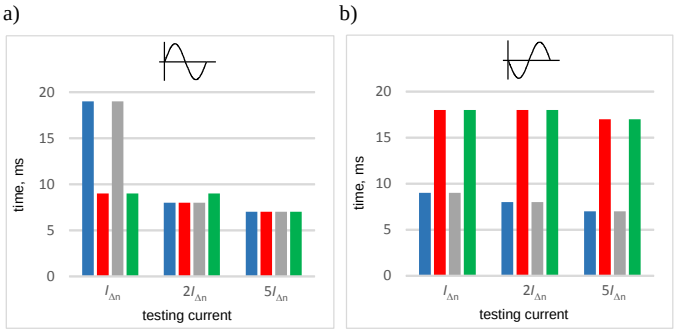


Figure 9. Disconnection times of RCD no. R7 (AC-type, from manufacturer M5); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

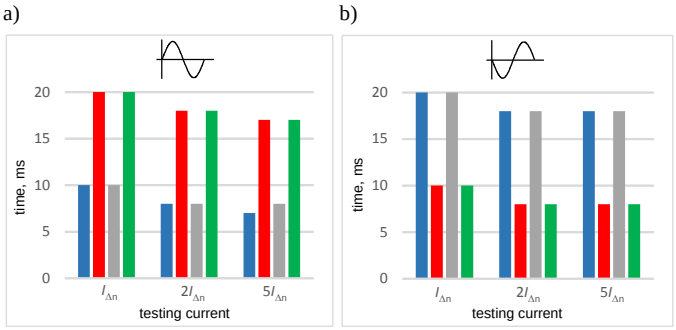


Figure 10. Disconnection times of RCD no. R8 (AC-type, from manufacturer M5); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

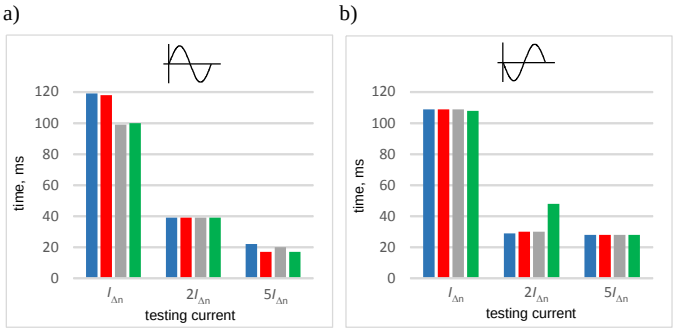


Figure 11. Disconnection times of RCD no. R10 (AC-type, from manufacturer M6); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

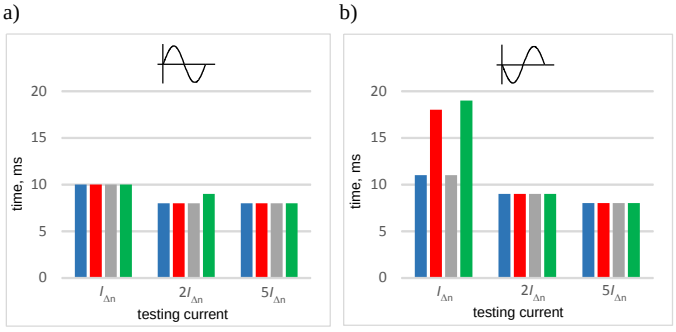


Figure 12. Disconnection times of RCD no. R18 (A-type, from manufacturer M5); results of the consecutive 4 trials for $I_{\Delta n}$, $2I_{\Delta n}$, $5I_{\Delta n}$ and waveforms: a) no. 1, b) no. 2.

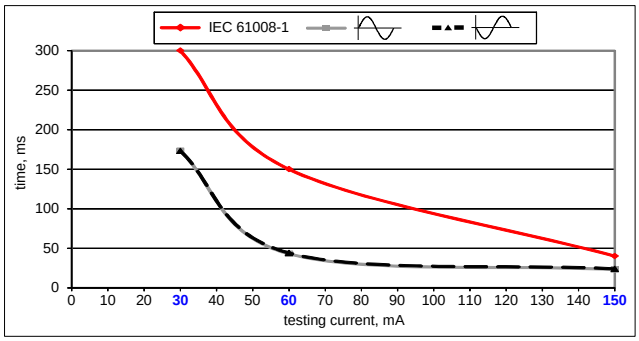
However, the most noticeable differences in results are given by the type of the testing waveform. Disconnection times for waveform no. 1 are clearly lower than for waveform no. 2 (Fig. 7).

Analysis of the results presented in Fig. 8 (RCD no. R17) enables to conclude that behaviour of this RCD is similar to RCD no. R6 (Fig. 7), but higher differences in disconnection times are observed, when compare results $I_{\Delta n}$ vs. $2I_{\Delta n}$ vs. $5I_{\Delta n}$.

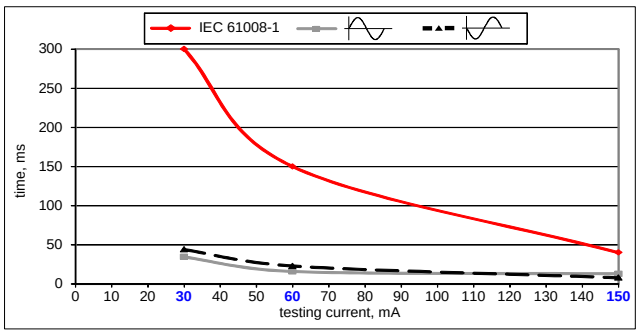
Properties of RCD no. R7 (Fig. 9) and RCD no. R8 (Fig. 10) are quite opposite to the priorly described. Their disconnection times are significantly different between consecutive trials. Consecutive disconnection times vary even over two-times, but this is not a negative result as well as the reason to eliminate these RCDs from further utilization.

RCD no. R10 (Fig. 11) and RCD no. R18 (Fig. 12) have also varied disconnection times within the given type of the test. Moreover, the times of these RCDs (R10 vs. R18) are very divergent, but do not exceed the permissible normative limit.

a)



b)



c)

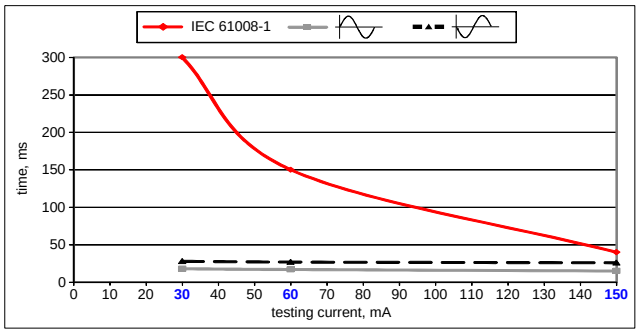


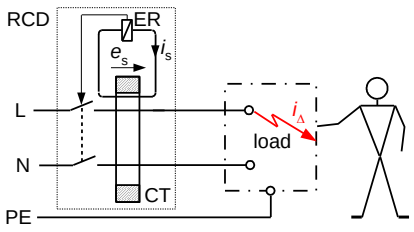
Figure 13. Time-current curves (average times) vs. normative curve IEC 61008-1 [9], for selected RCDs: a) R4 (B-type), b) R17 (A-type), c) R6 (AC-type).

Fig. 13 presents a comparison of the time-current curves for selected RCDs with the permissible curve required by standard IEC 61008-1 – the most different RCDs' curves have been selected. One can see that behaviour of RCDs – in terms of the disconnection times – can be various, and this is their normal property.

III. Discussion – Why the Consecutive Trials May Give Different Disconnection Times?

Tripping threshold of RCDs, including disconnection times, mainly depends on the properties of the current transformer of the RCD, as well as the type of the differential electromechanical relay applied in its secondary circuit (Fig. 14). Residual current i_Δ is transformed to the secondary side of the current transformer (CT) and it is a source of the secondary current i_s flowing through the electromechanical relay (ER). In the polarized electromechanical relay (Fig. 14b), the permanent magnet keeps the relay in the closed position, overcoming the force generated by the spring. When the residual current i_Δ occurs, the secondary current i_s, producing magnetic flux ϕ_{is}, in the first half-wave amplifies magnetic flux $\phi_{N\text{-}S}$ generated by the permanent magnet, but in the second half-wave reduces the flux $\phi_{N\text{-}S}$. If the reducing flux is high enough, tripping of the RCD occurs. Interaction between magnetic fluxes in such a relay is depicted in Fig. 15a. The second type of the ER – non-polarized relay – operates on a slightly different principle. Magnetic flux ϕ_{is} from the secondary current i_s saturates the narrowings of the yoke (pink areas in Fig. 14c). This phenomenon blocks the magnetic path of the magnetic flux $\phi_{N\text{-}S}$ derived from the permanent magnet. As a result, there is no possibility to keep the ER in the closed position – it leads to the tripping of the RCD. Magnetic fluxes interaction for the non-polarized relay is presented in Fig. 15b.

a)



b) c)

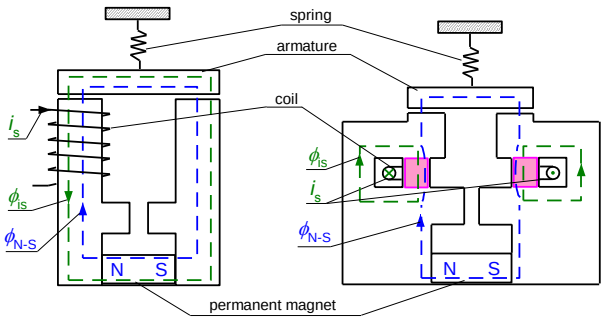


Figure 14. Structure of the RCD (a) and types of differential electromechanical relays of the RCD: b) polarized, c) non-polarized; i_Δ – residual current, i_s – secondary current, e_s – induced voltage, ϕ_{is} – magnetic flux from the secondary current i_s, $\phi_{N\text{-}S}$ – magnetic flux from the permanent magnet, ER – electromechanical relay, CT – current transformer.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

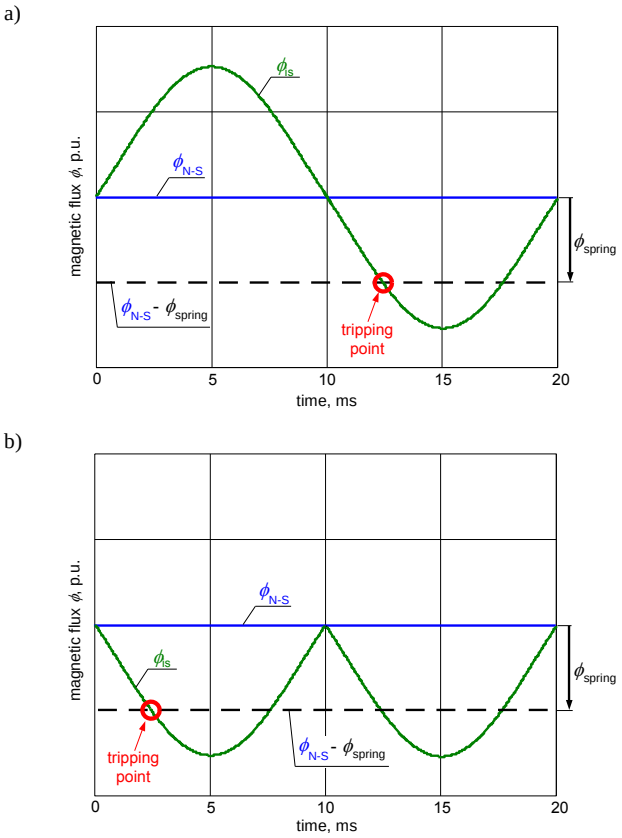


Figure 15. Magnetic fluxes in the electromechanical relay of RCDs: a) polarized, b) non-polarized; $\phi_{N\text{-}S}$ – magnetic flux from the permanent magnet, ϕ_{is} – magnetic flux from the secondary current i_s, ϕ_{spring} – theoretical/equivalent magnetic flux from the spring.

Thus, the non-polarized relay has a shorter average disconnection time because in each half-wave it reduces the keeping force of the permanent magnet. When the polarized relay is used, the initial phase angle (0° vs. 180°) of the secondary current i_s may affect the disconnection time. Differences in the disconnection times for consecutive trials can also be caused by various configurations of the secondary circuit of the RCD, which is presented in Fig. 16.

In the analysis of the operation of the electromechanical relay ER, properties of the current transformers should also be taken into account. The shape of the induced voltage (e_s in Fig. 14a) in the secondary winding of the CT depends on its iron core magnetic properties. Fig. 17 presents e_s waveforms for two types of the CT – for A-type RCDs (Fig. 17a) and for AC-type RCDs (Fig. 17b). The primary (residual) current of the current transformer was sinusoidal during the test. In case of the CT for A-type RCDs, the induced voltage is almost sinusoidal, but in case of the CT for AC-type RCDs, the induced voltage is strongly distorted – it has a shape in the form of impulses, what may influence the disconnection times.

Taking the above considerations into account, varied disconnection times in the consecutive trails and significantly different disconnection times between two or more compared RCDs are acceptable as long as they are within the normative range.

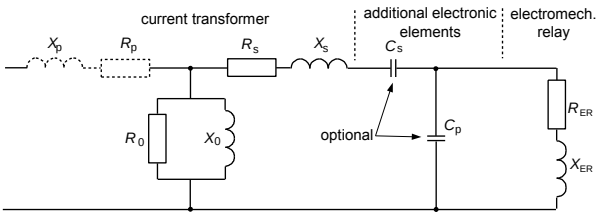


Figure 16. Equivalent circuit of the RCD with indicated example optional electronic elements (C_s – serial capacitor, C_p – parallel capacitor).

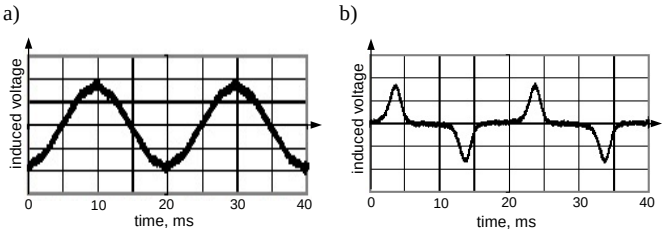


Figure 17. Oscillograms of the induced secondary voltage e_s in the secondary winding of the current transformer of the RCD: a) A-type, b) AC-type. Primary current equal to $I_{\Delta n}$.

IV. Conclusions

Experiences in initial and periodical testing of the RCDs show that the disconnection times of RCDs may vary between two or more general type RCDs of the same rated residual operating current. Differences in these times can be high even in case of the consecutive trials of the same RCD. However, it is normal behaviour, and should not be the basis for excluding the tested RCD from further utilization, as long as the disconnection times and real tripping current are within the normative range.

References

[1] Low-voltage electrical installations – Part 4-41: Protection for safety – Protection against electric shock, HD 60364-4-41, 2017.

[2] S. Czapp, "The effect of PWM frequency on the effectiveness of protection against electric shock using residual current devices," 10th Conf.-Seminar Int. School on Nonsinusoidal Currents and Compensation (ISNCC), 2010, Lagow, Poland, DOI: 10.1109/ISNCC.2010.5524515.

[3] S. Czapp, "The impact of DC earth fault current shape on tripping of residual current devices," Elektronika ir Elektrotechnika, vol. 84, no. 4, pp. 9-12, 2008.

[4] G. Rongyan and Z. Honghui, "Study on the residual current protection device technology," The Open Electrical & Electronic Engineering Journal, vol. 8, pp. 404-411, 2014.

[5] Y. Han, Ch. Ding, and X. Shou, "Design & implementation of an A-type residual current circuit breaker IC," IEEE Int. Symp. on Industrial Electronics, 2012, pp. 280-285, DOI: 10.1109/ISIE.2012.6237098.

[6] The RCD handbook. BEAMA guide to the selection and application of residual current devices (RCDs), 2018.

[7] S. Czapp and J. Horiszny, "Simulation of residual current devices operation under high frequency residual current," Przeglad Elektrotechniczny, vol. 88, no. 2, pp. 242-247, 2012.

[8] Low-voltage electrical installations – Part 6: Verification, IEC 60364-6, 2016.

[9] Residual current operated circuit-breakers without integral overcurrent protection for household and similar uses (RCCBs) – Part 1: General rules, IEC 61008-1, 2010.

[10] Effects of current on human beings and livestock – Part 1: General aspects, IEC 60479-1, 2018.

The efficiency of the Temporal Medical Data Retrieval

Michal Kvet, Karol Matiaško
Department of Informatics, Faculty of Management Science and Informatics
University of Žilina
Žilina, Slovakia
Michal.Kvet@fri.uniza.sk

***Abstract*—One of the significant aspects influencing patient treatment is associated with decision-making correctness based on reliable data inputs. Data should be managed, treated, and provided in a robust and performance-effective manner. Commonly, the patient needs to be monitored over time reflecting the evolution. This paper deals with the temporal database models and proposes an effective solution for the data evaluation during the retrieval, so the query request can end significantly sooner. It also points to the current problem of data management in the cloud technology as a central data repository, in which the data time perspectives can be shifted reflecting the time zone, so the results must be transformed to provide relevant data output in a client site.**

Keywords—medical data; temporal aspect; complex indexing; parallelism;

I. INTRODUCTION

Nowadays, great emphasis is done on complex patient treatment, to ensure correct health care across the world. A person is generally managed by the general practitioner, who gets the main overview and shits the patient to the specialist based on the assumptions and previous examination [10] [12]. Currently, data and results are commonly shared by the centralized hospital, regional, or even national system. As evident, data efficiency and proper management are crucial to getting the relevant data almost immediately. Performance and processing demand, mostly reflected by the time consumption cover the inevitable elements to be treated. Data are stored in a local on-premise server, but now, the strong demand to centralize management in a complex cloud system can be felt [6] [19]. The main reason is based on the data amount, which is still rising. The patient is treated more and more complexly, the medical care is still improved producing more reliable, more precise data outputs, which must be stored, analyzed, and retrieved on the demand. It is therefore clear, that the problem of the efficiency of the whole process is relevant and must be managed.

Moreover, nowadays, we can see the Covid-19 pandemics across the world influencing any sphere. Business and production are closed or strictly limited, transport is markedly monitored to limit the spread of the virus, which strongly influences the situation, fills the hospitals, and consequencing in getting the massive number of victims.

With the gradual relaxation of measures against the spread of coronavirus, there is again a strong need to monitor outbreaks and thus reduce the impact and overall problems of those infections.

Similarly, as already stated, all evaluations and systems produce various data structures but have two common aspects. Firstly, the produced data amount is significant and secondly, there is strict pressure to get the data in a massive, parallel, and mostly in a performance effective, robust, and reliable manner. Such data have a common definition of the time spectrum, as well as region assignment.

This paper deals with the temporal database architecture to provide support for complex medical data. In section 2, current technologies and temporal architectures are summarized, pointing to the specifications and limitations. The proposed extension is done by using an index relocation unit to shift the internal granularity to fit the request specified inside the query. Section 3 deals with the proposed architecture for medical data management and supervision. In section 4, the data retrieval process management handler, covered by the two-level parallel indexing strategy is defined. All data are time positioned. By shifting the environment to the cloud technology, the problem of the different time zones can be present, whereas most of the technology reference server time, not the transformation to the client [1] [4]. As a result, wrong values are provided. In section 5, we deal with the transformation covering the analysis of the additional demands to the processing. Internally, it is done by the query time replacement using the translation package. Section 6 deals with the performance evaluation.

II. TEMPORAL ARCHITECTURES

Significance of the temporal data management in database systems has been identified with the advent of the relational paradigm supervised by the transactions. Soon, the first temporal architecture delimited by the object-level temporal model was proposed. The identifier of the object itself was extended by the time attributes characterizing validity, transaction validity, or other temporal attributes. As a consequence, the object state cannot be uniquely identified by the object representation itself, the definition of the time-image should be handled, as well. The main disadvantage of such an approach is just the granularity itself. Each new state has a full definition, thus, each attribute value should be

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

present. There is no need to get an image of several states, there is no need to compose a state from several fragments. Vice versa, such an approach can produce many duplicate values, if the update operation is not synchronized across the whole object state definition. Moreover, there is no solution for dealing with just the relevance defining the region or position of interest. Finally, such a solution can be significantly ineffective in terms of performance, but mostly in the aspect of the data storage demands. Original state values can be copied to new images if no change occurs. Synchronization across the whole state should be present to obtain the optimal solution. As evident, object-level temporal architecture is not commonly suited for the medical data with the dynamic evolution reflecting the region of the interest monitoring.

Attribute-oriented architecture was complexly defined in [3] [4] [15]. Each state is delimited by the composition of individual attributes along with the time definition. Each change is divided into the attributes themselves, which are maintained separately. Namely, each attribute change forces the definition of the new state physically stored by the temporal layer. Such architecture is suitable in terms of the physical layer perspective. It is, however, clear, that the data retrieval process can be demanding forcing the system to calculate the state on demand. In ad-hoc networks forcing the systém to obtain data dynamically with a strong impact on the data transfer, such architecture is not suitable, whereas the pre-processing and state composition can last too much time. In medical data management, the situation is similar with the particular emphasis on the processing efficiency of the state retrieval as the data are commonly centralized. It creates significant pressure on the optimization of data processing, retrieval, and transmission. The architecture of the attribute-oriented temporal model is in fig. 1. It consists of four-level architecture. The first level deals with the current state of the object. Physically, data are stored in the attribute granularity, however, the retrieval representation is done using the object perspective composition. The second level is formed by the core of the solution – the temporal evidence management layer. Any change is registered there, with the pointing reflection to the particular layer in an attribute perspective. Such a module is also responsible for the object granularity state composition during the data retrieval. Level 3 deals with the historical data, level 4 manages plans (level 4 is an optional level, if no plans or states valid in the future need to be monitored, such representation does not need to be used). The temporal level is interconnected to the index relocation unit, where the requested format is created, mostly modeled by the object granularity. The index relocation unit is then interconnected to the public interface providing result sets to the client site. In comparison with the original attribute-oriented granularity presented in [13], our proposed solution uses an external source to provide any data granularity perspective using an index relocation unit.

Extension to fuzzy management is discussed in [2] [5].

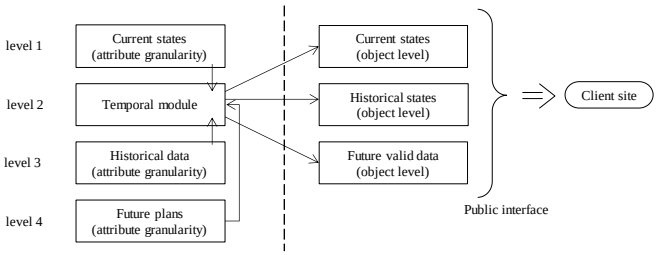


Figure 1.Attribute oriented granularity

The last solution is a group granularity composition. Instead of storing each attribute change separately, the dynamic temporal group can be identified (either manually or in an automatic manner using a data analytics perspective). Thanks to that, if several attributes are update synchronized, only one new row is added to the temporal module limiting the amount of the data to be stored, with no loss of the information value. Physically, it is covered by the analytical module extending the temporal module. In principle, each attribute represents the one element group, so the reference is always done on the group level, which can be temporarily grouped into the various segments. Fig. 2 shows the group state hierarchy. The input stream is routed to the data change barrier, where the attribute or synchronization groups are identified by the Grouper background process. The analytical module is responsible for the synchronization detection, so the storage demands are always up to date and optimized based on the data flow. Note, that in fig. 2, individual levels are visualized using the storage type, however, internally, current, future, and historical data are treated and stored separately.

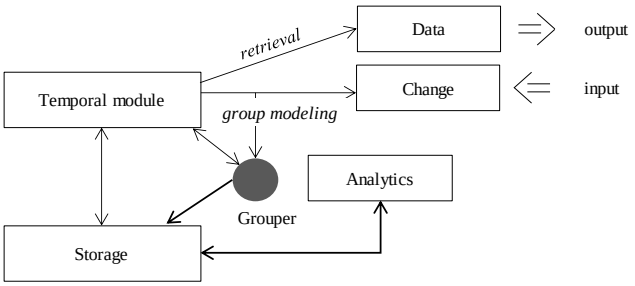


Figure 2.Group data management

Original group level architecture is defined in [13]. The stated solution introduces the Grouper background process, which brings additional benefits, whereas the group definition can be automated and treated dynamically in parallel using the Analytics module. The original solution required to specify groups either by the user or the optimizer could do that. And that is just the point. The optimizer was congested causing delays.

III. TEMPORAL GROUP REGION ARCHITECTURE

In the previous sections, we have summarized current temporal approaches by pointing to the limitations. Besides, we have proposed several extensions covered by this paper, mostly reflected by the synchronization group extension and index relocation unit. Section 2 deals just with temporality, in

this part, we propose the Spatio-temporal solution covering the positional data in the region assignment principle. Thanks to that, medical data can be delimited by the patient himself, temporal elements, but the positions of the interest can be managed, as well.

The proposed architecture overview is shown in fig. 3, reflected by the data model located in the temporal layer. A core part of the model is formed by the temporal table referencing any data update by assigning its change_id chaining individual changes for the particular object. Statement_type delimits the data modeling operation (Insert, Update, Delete). Pointer to the data object is done by the table identifier (id_tab), attribute or temporal group (data_id) definition, and row (row_id). Non-current values are grouped into individual data type categories referenced by the data_type_cat. Note, that the data type categories are maintained automatically by the internal transformation opportunity (implicit conversions). As stated, the model is spatio-temporal meaning, that the temporal table consists of the validity time frame, reliability time dimension expressed by the transactions and spatial assignment managed either locally by spatial_positions or by dynamic region covering – region_assignment.

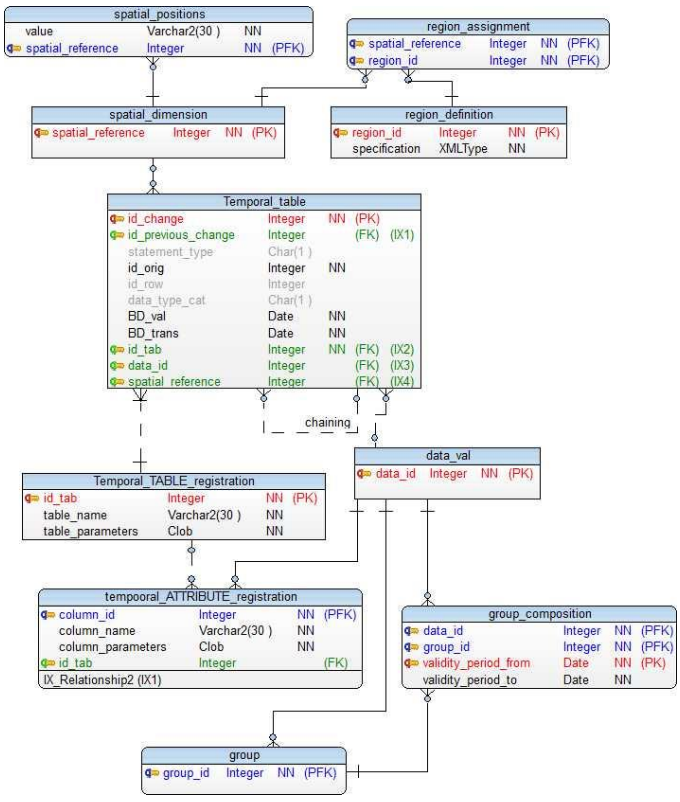


Figure 3.Group spatio-temporal model

IV. DATA RETRIEVAL

Data retrieval perspective is a core element influencing the whole performance. Optimization and evaluation is a staged process to get reliable results. Namely, the optimizer has to check the prerequisites for the processing – semantic and syntactic check reflected by the access rules, object identification by the owner, etc., followed by the execution plan composition. Generally, several plans are identified, evaluated by the estimated costs originating from the data table and database statistics. The aim is to select the best suitable plan for the consecutive evaluation. The execution process itself can be done either by using sequential scanning of the data blocks associated with the table or by using the index layer [20]. Sequential scanning is the most demanding operation, whereas the data amount to be produced cannot be evaluated directly. Moreover, individual blocks can be fragmented, located in multiple physical discs, as well, bringing additional processing demands. Moreover, some blocks can be even empty, as a result of extent initialization, instead of the blocks themselves. Each extent is composed of the defined amount of data blocks, which allows you to associate blocks dynamically by limiting the necessity of allocation any time. The second current solution is based on using an index strategy to locate relevant data blocks. Namely, instead of block scanning, a defined index is used to locate data. Pointers to the direct data are located in the leaf layer, specifying ROWID value – address of the row inside the block and data file forming the database. The index is far smaller in comparison with the whole table, can be optimized [6] [7] [9], but mostly, it is directly accessible in the database instance memory. The index itself is defined by the named set of the individual table attributes or function calls, where the element order is significant forming the structure [8]. In database systems, the B+tree index structure is used most often as a default strategy. B+tree database index is formed by the root node, internal nodes, and leaf layer pointing to the data. The limitation of the indexing is just the reliability issue – undefined table data cannot be indexed, whereas NULL values cannot be mathematically sorted and evaluated. As a consequence, commonly, if the query can produce an undefined or untrusted value, the processing is shifted to sequential data scanning. In [16] [17], it is solved by introducing NULL modules, which hold undefined data directly inside the index by using mapping function [11] [14] [18]. In comparison with function-based index transforming undefined values, mapping is done internally with no additional storage demands. NULL values can be stored either internally or in an external module interconnected to the root element.

Flower Index Approach (FIA) is formed by a specific robust index method usable in case of unavailability of the (sub)optimal access path. In that case, sequential scanning would be necessary to perform. As stated in the research paper [14] [15], if the data fragmentation caused by a huge update stream or volatility aspect is present, total performance would be poor, several data blocks would be necessary to be memory loaded, with no relevant data there. FIA approach is used as a data block locator, where at least one existing data row is present. In the leaf layer of the index, BLOCKIDs are present pointing to the whole block, instead of the data row position. By loading the data, the whole block is evaluated to focus on the data.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The performance of the index is limited by the structure, which can, in principle, degrade over time, if several updates and delete operations are present. The property of the B+tree is based on the balancing, so the traverse path from the root to any leaf node is always the same in terms of depth. If the node inside the index does not hold any data after the Update or Delete operation, a particular node remains in the index structure still, because it is assumed that the node will be used again in the early future. Such definition can have, however, significant performance impact, if the type and values are evolving responding to the accuracy and precision. In [14], database index balancing strategies are defined. Balancing operation is extracted from the main transaction and is operated separately consequencing in the ability to approve the original change operation and transaction sooner. Changed values are stored either in a separate structure by applying them by the introduced Balancer background process [14] or are placed in the leaf index layer with no reflection to the balancing itself [15].

Thanks to that, it is ensured, that the index is still up-to-date, even in terms of the structure and tree depth. Based on the computational study, it has minimal additional data retrieval demands – less than 1 percent.

Proposed spatio-temporal solution

The proposed solution covers spatial and temporal perspectives together. Architecture can be split into two internal regions supervised by the background processes and available through the public interface. The internal layer consists of two index types. B+tree index set is formed by the primary index set, where objects irrespective of the spatial and temporal dimension are registered and indexed. The secondary index set is managed by the user specification. It commonly stores B+tree definition, as well, however, the bitmap, compressed, hash, or reverse indexes can be covered by such module, as well. Spatial and temporal dimensions are secured by the separate index set. These data portions can be processed and evaluated in parallel followed by the merging operation, which is done by the bitmap index pointing to the ROWIDs. Data blocks are then loaded, the result set is created and provided to the client via a public interface. The internal B+tree index set does not point to the data blocks (like in a common conventional system, but the leaf layer references the bitmapper located in the second interval layer). Fig. 4 shows the architecture.

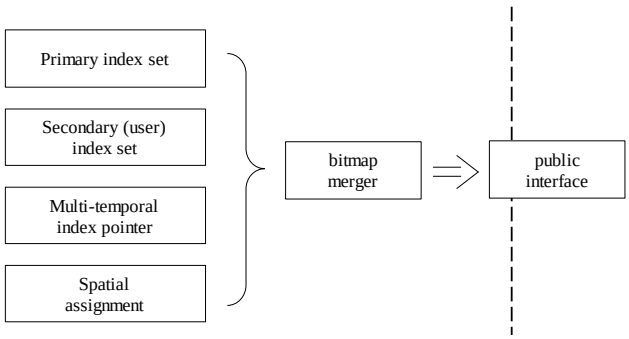


Figure 4.Group spatio-temporal model

V. TIME MANAGEMENT

Common characteristics of current information systems (not only regarding the medical data perspective) are based on time spectrum limitation, mostly reflected in the server. Most of the systems represent just server time definition, instead of the local client. It was based on the time zones - it was assumed that the server settings are always correct and provide the correct values regardless of the time zone parameter setting on the client-side. Thanks to that, it was ensured that the provided data were relevant - the server was stored directly in the medical center, in the local server room. With the spread of cloud technologies, security standards, and complexity of the requests, there has been a legitimate demand to move data to cloud storage, to manage them autonomously [21]. As we mentioned in the description of the architecture, the amount of data to be processed is constantly growing, and therefore the dedicated cloud storage provides a suitable space for growth and ensuring robustness, performance and security. The problem is just the time perspective. Based on the study provided by Oracle, most of the current systems use server time reflection (sysdate), instead of the client, which can produce incorrect data (curret_date) for getting the current date and time. The difference is just the time zone, however, if the cloud storage is in another region or is even placed in a different continent, it would be necessary to rewrite the whole code to reflect the client perspective and time zone. It would to very demanding and prone to errors. Moreover, the time evaluation perspective in the medical systém is crucial. Therefore, it is necessary to find another solution to represent local client time. In this paper, we propose and evaluate the costs and benefits of the translation profiles of the SQL.

Solution – SQL translation profile

Our proposed solution covering local client time is based on the SQL translation profile, which dynamically replaces the original definition of the server time by the client time zone. SQL translation profile is commonly set by the database administrator (DBA) by invoking the CREATE_PROFILE procedure of the DBMS_SQL_TRANSLATOR package. Each profile is delimited by its unique name, defined attributes with specified values, and a pointer to the package method, which is invoked any time the SQL query is to be executed.

The invoked function is associated with the ATTR_TRANSLATOR parameter. The translator function must be packaged with the following two procedures:

- Procedure TRANSLATE_SQL has two parameters by passing the original SQL query resulting in providing the second parameter as an output in form of character LOB value.
- Procedure TRANSLATE_ERROR is called if the translation ends with a raising exception. It has three parameters – binary integer value ERROR_CODE, TRANSLATED_CODE, and TRANSLATED_SQLSTATE.

To cover the reliability of the medical data processing, the translation function is associated with the regular expression calling REGEXP_REPLACE method. The syntax of the REGEXP_REPLACE procedure is in fig.5 [22]:

- SOURCE_CHAR is a string used as a search value.

- PATTERN is a regular expression with the same data type as SOURCE_CHAR, respectively transformable implicitly.
- REPLACE_STRING
- POSITION – positive integer defining the starting position for the evaluation.
- OCCURRENCE – non-negative value (n) delimiting the replacement of the n-th occurrence. Value 0 expresses all occurrences to be replaced.
- MATCH_PARAMETERS characterizes the sting format to be handled. Value "i" in our case defines case insensitivity.

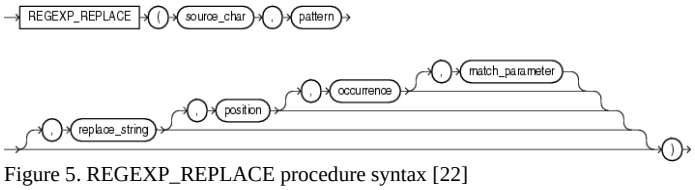


Figure 5. REGEXP_REPLACE procedure syntax [22]

Implementation

Medical data transformation is done by using the translator defined in the following code snippet. In case of ensuring server-client time transformation, SQL translation profile is defined by the following steps. First of all, the translation profile has to be created and mapped to the package definition, which consists of two methods. Then, the session is mapped to the profile.

Package definition

```
create or replace package date_translator
 is
  procedure translate_sql(sql_text in clob,
                translated_text out clob);
end;
/

create or replace package body date_translator
 is
  procedure translate_sql(sql_text in clob,
                translated_text out clob)
    is
     begin
      translated_text:=regexp_replace(sql_text, 'SYSDATE',
                                      'CURRENT_DATE,1,0,'i');
     end;
end;
/
```

Note, that the procedure TRANSLATE_ERROR is optional. If omitted, if any error occurs, translated code is ignored and the original definition is used instead.

Profile definition and association

```
begin
 dbms_sql_translator.create_profile
  (
   profile_name => 'MEDICAL_PROFILE'
  );
 dbms_sql_translator.set_attribute
  (
   profile_name => 'MEDICAL_PROFILE',
   attribute_name =>
    dbms_sql_translator.ATTR_FOREIGN_SQL_SYNTAX,
   attribute_value =>
    dbms_sql_translator.ATTR_VALUE_FALSE
  );
 dbms_sql_translator.set_attribute
  (
   profile_name => "MEDICAL PROFILE',
   attribute_name => dbms_sql_translator.attr_translator,
   attribute_value => 'DATE_TRANSLATOR'
  );
end;
/
```

Profile definition and association are covered by three-step blocks inside the anonymous block execution. The first calls the procedure create profile delimited by the name. The second block is based on the changing default parameter *ATTR_FOREIGN_SQL_SYNTAX* from the value *ATTR_VALUE_TRUE* to *ATTR_VALUE_FALSE. Such a parameter specifies the type of SQL syntax.* The third call sets the ATTR_TRANSLATOR pointer to the defined package (DATE_TRANSLATOR).

Mapping

```
alter session
  set sql_translation_profile=MEDICAL_PROFILE;
```

Mapping is done by associating profile to the session.

Result

Data transformation can be obtained by using V$SQL dynamic performance view.

```
select sql_text
 from v$sql
  where sql_text
          like 'select%medical_data%examination_date%';
```

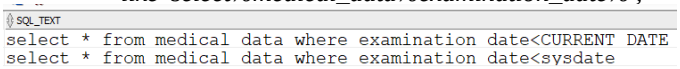


Figure 6. Transformation results

Analysis of the additional processing demands

Cloud environment can provide scalable dynamic solutions ensuring the global performance of the system. By transferring the implemented systems to the cloud environment, reference of the time zones is highlighted, as well. In this part, there is an analysis of the additional demands regarding the SQL translation profiles. Generally, there are two available domains to be implemented. The first one is simpler in terms of

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

management, whereas the translation is done automatically. The second approach is delimited by the parse code reference delimited by the base lines. It will work, however, robustly, only if the reference parse is always accessible and loaded, even after the system restart, so the administrator needs to ensure it, otherwise the processing will be aborted.

Experiments were done in magnetic resonance imaging and patient monitoring, where time precision is crucial. As evident from the following figures, the solution is scalable with minimal additional demands caused by the translation – 10ms. Note, however, that it is done only once before the hard parsing, afterwards, just the reference to the already stored library cache parse is used directly with no transformation, at all. Fig. 7 shows the original demands, fig. 8 shows the results by applying translation. Only time demands are slightly changed, access methods, as well as total processing costs expressed by the size, processing time, and resource consumption, remain the same [3]. In fig. 7, server date is used, whereas fig. 8 uses client site evaluation.

SQL | 0,113 seconds

OPERATION	OBJECT_NAME	OPTIONS	CARDINALITY	COST
SELECT STATEMENT			37	2
TABLE ACCESS	MEDICAL_DATA	FULL	37	2
Filter Predicates				
EXAMINATION_DATE<SYSDATE@!				

Figure 7. Original statement

SQL | 0,123 seconds

OPERATION	OBJECT_NAME	OPTIONS	CARDINALITY	COST
SELECT STATEMENT			37	2
TABLE ACCESS	MEDICAL_DATA	FULL	37	2
Filter Predicates				
EXAMINATION_DATE<CURRENT_DATE				

Figure 8. Translated statement

VI. COMPUTATIONAL STUDY

Performance characteristics have been obtained by using the Oracle 19c database system (Oracle Database 19c Enterprise Edition Release 19.0.0.0.0 – Production). Parameters of used computer are processor: Intel Xeon E5620; 2,4GHz (8 cores), operation memory: 16GB, HDD: 500GB.

The results of the patient monitoring using magnetic resonance imaging are used. 1000 patient data were used, each of them is delimited by the 10 results provided over time. The brain tumour is detected, located, and monitored. In the first phase, whole data about the brain are stored, later, just regions of interest with significant changes between individual results are stored to optimize consecutive evaluation and limit storage demands. The average size of the data set is 96 MB consisting of 20 markers.

In [13] [14], individual granularity types and architectures are compared technically, mostly reflecting the efficiency of the data transfer. In this paper, the evaluation study will focus on the proposed spatio-temporal architecture supervised by the parallel processing through indexes followed by the bitmap merger. In comparison with already existing approaches, the whole data should be treated, there is no robust region of interest temporal solution. Based on the provided data, just 20% of the data are monitored over time based on the significance evaluation.

For the study, we used two streams. In the first experiment, monitoring of just one patient is used. Data, which are not stored, are replaced by the baseline data results, whereas there is not a necessity to store them, they do not point to any change over time, so there is no reason to store duplicate values. The second experiment deals with the monitoring of multiple patients with the same diagnosis to compare and highlight the treatment methods and reached results.

EXPERIMENT 1

As stated, each patient is represented by 10 examination results, mostly covered by the biannual time-frequency. The first result set is always stored completely by locating potential anomalies, which are marketed as regions of interest. Besides, also the positional neighborhood is maintained to evaluate the detection progress. Besides, for each evaluation, Epsilon perspectives are handled to ensure any significant change is covered properly. Based on the evaluation, the average data amount to be stored and processed was limited to the value of 20% reflecting the temporality and spatial dimension. The significant aspect is just the reliability, such data amount reduction does not influence the information values, at all. The data retrieval process is evaluated, reflected by the storage demands and data retrieval process. Non-stored values are obtained by the base line and marked by the obtain time point.

Fig. 9 shows the results comparing temporal group dimension with the proposed spatio-temporal architecture. Fig. 9 represents the storage demands expressed in MB. In comparison with individual attribute management, storage demands are lowered to 45%, thanks to dynamic data processing and group detection. Just 27% of the storage capacity is necessary for the spatio-temporal dimension, in comparison with the reference model – attribute granularity.

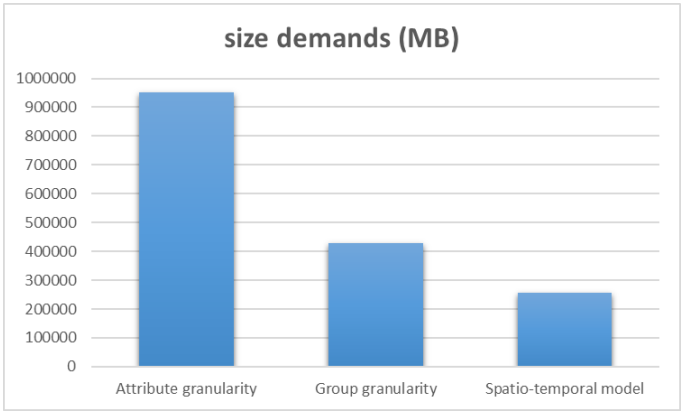


Figure 9. Size demands

Fig. 10 shows the processing time demands, which are lowered, similarly, as well. Namely, attribute granularity requires 45,74 seconds. By applying group detection, only 37,84 seconds are required, whereas the synchronization groups are managed as one element stored in the temporal

layer. Proposed group granularity of the temporality and spatial dimension requires only 25,03 seconds, 3,87 seconds are used for the non-stored values obtaining from the base line, which reflects 15,46% of the total processing for the proposed solution.

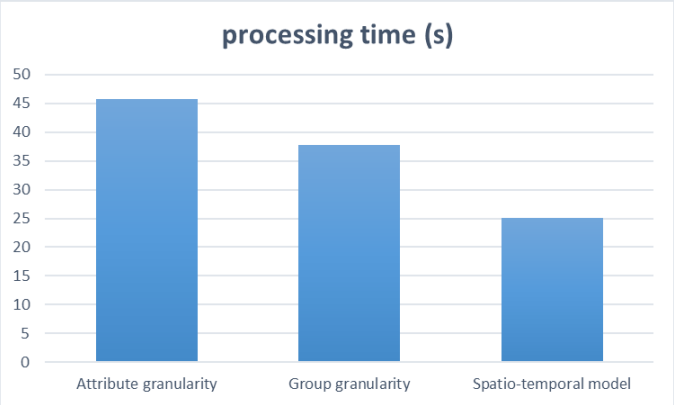


Figure 10. Processing time

EXPERIMENT 2

The second evaluated study is based on monitoring several patients over time by using the proposed spatio-temporal technique. We have identified the group with the same treatment category containing 37 patients. Fig. 11 shows the results reflecting the size demands, which are reduced from the value 35,5 GB for the attribute level architecture to the value 16,2 GB for temporal group dimension. By applying spatio-temporality, 12,72 GB. The main size demands drop possibility is based on the categorization, it is clear, that all the patients have the same disease, so the location can point to the specific region directly.

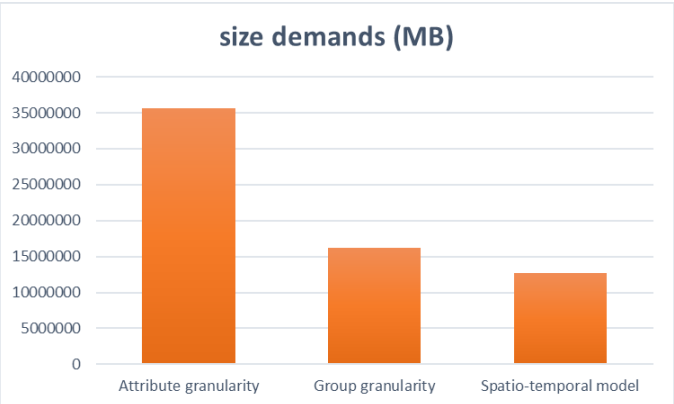


Figure 11. Size demands

Processing of such data amount is demanding and should be reduced as much as possible. Spatio-temporal model can directly locate data of interest, whereas it is just the same for each patient (with regards to the data neighborhood). Thanks to that, processing demands are lowered from the value 1712 seconds for the referential attribute granularity to the 1378 seconds for the group detection model (19,5% improvement by reducing processing time). The proposed spatial model focuses on just the spatial positions, not only temporality, thus the time demands are 946 seconds (44,7% referencing attribute model and 31,3%).

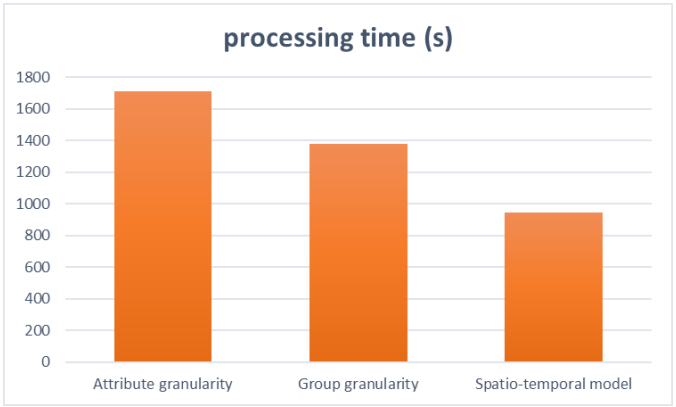


Figure 12. Processing time

VII. CONCLUSIONS

Data efficiency is the main point expressing the complexity, robustness, and reliability of the whole information system. If the data are not provided in a defined time spectrum, consecutive analysis, management, evaluation, and decision making cannot be done. In medical systems, it is necessary to manage data of the patient over time, to be provided immediately, and cover real-time activities. Therefore, the strict impact should be taken into the whole optimization of the database processing as an internal layer providing and managing data. In this paper, we summarize existing temporal systems, which can cover the whole data tuples and their changes in the whole time spectrum. The analysis emphasizes the available granularity levels by forming spatio-temporal solutions covered by four-level architecture splitting data based on the time spectrum.

The main contribution of this paper is defined in section 3 proposing spatio-temporal group solutions covering all existing approaches to reach a robust effective model dealing with Flower Index Approach, undefined states, etc. The aim is to ensure the effectiveness of the processing, mostly pointing to the data retrieval process. The proposed solution extends the existing index set management by using parallel processing merged by the bitmap system.

The limitation of the current systems is just the temporality management, mostly represented by the server time, which brings problems if the data are migrated to the cloud environment to propose a robust, reliable, and secure solution. In this paper, we also analyze the impact of the SQL translation profiles definition on performance. As evident from the reached results, the transformation is done only once, followed by the storing execution plan in the memory cache.

In the future, our emphasis will be taken to the cloud environment block storage and baseline definitions, by mapping the SQL statement to the pre-stored parsing process. Thanks to that, the translation can be hosted internally, which can lower the additional processing time demands completely.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

ACKNOWLEDGMENT

This publication was realized with support of Operational Program Integrated Infrastructure 2014 - 2020 of the project: Intelligent operating and processing systems for UAVs, code ITMS 313011V422, co-financed by the European Regional Development Fund.

REFERENCES

[1] Abdalla, H. I.: A synchronized design technique for efficient data distribution, Computers in Human Behavior, Volume 30, 2014, pp. 427-435

[2] Behounek, L., Novák, V.: Towards Fuzzy Patrial Logic. In 2015 IEEE Internal Symposium on Multiple-Valued Logic, 2015.

[3] Bryla, B.: Oracle Database 12c The Complete Reference, Oracle Press, 2013, ISBN – 978-0071801751

[4] Burleson, D. K.: Oracle High-Performance SQL Tuning, Oracle Press, 2001, ISBN - 9780072190588

[5] Delplanque, J., Etien, A., Anquetil, N., Auverlot, O.: Relational database schema evolution: An industrial case study, IEEE International Conference on Software Maintenance and Evolution, ICSME 2018, Spain, 2018, pp. 635-644

[6] Dostál, J., Wang, X., Steingartner, W., Nuangchalerm, P. , - Digital Intelligence - New concept in context of future school education, In: 10th Annual International Conference of Education, Research and Innovation (ICERI2017), Seville, SPAIN, NOV 16-18, 2017, pp. 3706-3712, 2017

[7] Eisa, I., Salem, R., Abdelkader, H.: A fragmentation algorithm for storage management in cloud database environment, Proceedings of ICCES 2017 12th International Conference on Computer Engineering and Systems, Egypt, 2018

[8] Feng, J., Li, G., Wang, J.: Finding Top-k Answers in Keyword Search over Relational Databases Using Tuple Units, IEEE Transactions on Knowledge and Data Engineering (Volume: 23, Issue: 12, Dec. 2011) , 2011.

[9] Honishi, T., Satoh, T., Inoue, U.: An index structure for parallel database processing, Second International Workshop on Research Issues on Data Engineering: Transaction and Query Processing, 1992.

[10] Janáček, J., Kvet, M. (2016). Sequential approximate approach to the p-median problem. In Computers & Industrial Engineering 94 (2016), Elsevier, ISSN 0360-8352, pp. 83-92.

[11] Kriegel, H., Kunath, P., Pfeifle, M., Renz, M.: Acceleration of relational index structures based on statistics, 15th International Conference on Scientific and Statistical Database Management, 2003

[12] Kvet, M. (2019). Complexity and Scenario Robust Service System Design. In Information and Digital Technologies 2019: conference proceedings, Žilina, 2019, ISBN 978-1-7281-1400-2, pp. 271-274.

[13] Kvet, M.: Managing, locating and evaluating undefined values in relational databases. 2020

[14] Kvet, M.: Database Index Balancing Strategy, in print (2021)

[15] Kvet, M., Kršák, E., Matiaško, K.: Study on effective temporal data retrieval leveraging complex indexed architecture, Applied Sciences 10 (2020)

[16] Lien, Y.: Multivalued Dependencies With Null Values In Relational Data Bases. In Fifth International Conference on Very Large Data Base, 1979.

[17] Mirza, G.: Null Value Conflict: Formal Definition and Resolution, 13th International Conference on Frontiers of Information Technology (FIT), 2015.

[18] Moreira, J., Duarte, J., Dias, P.: Modeling and representing real-world spatio-temporal data in databases, Leibniz International Proceedings in Informatics, LIPIcs, Volume 142, 2019

[19] Schreiner, W., Steingartner, W. and Novitzká, V.: A Novel Categorical Approach to Semantics of Relational First-Order Logic, Symmetry-Basel, Vol. 12, No. 10, MDPI, OCT 2020, doi: 10.3390/sym12101584

[20] Vinayakumar, R. Soman, K., Menon, P.: DB-Learn: Studying Relational Algebra Concepts by Snapping Blocks, International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, India, 2018

[21] https://www.oracle.com/database/what-is-autonomous-database.html

[22] https://docs.oracle.com/cd/B19306_01/server.102/b14200/functions130.htm

Mapping rules for schema transformation

SQL to NoSQL and back

Roman Čerešňák
Faculty of Management and Informatics
University of Žilina
Žilina, Slovakia
roman.ceresnak@fri.uniza.sk

Adam Dudáš
Faculty of Natural Sciences
Matej Bel University
Banská Bystrica, Slovakia
adam.dudas@umb.sk

Karol Matiaško
Faculty of Management and Informatics
University of Žilina
Žilina, Slovakia
karol.matiasko@fri.uniza.sk

Michal Kvet
Faculty of Management and Informatics
University of Žilina
Žilina, Slovakia
Michal.kvet@fri.uniza.sk

Abstract— **Efficient way of storing data has always been a key requirement for a properly designed database system. With the growing demand for this property, the first concept of an efficient data storage called relational databases was developed in the 1960s - this type of databases is still used as the primary data storage to this day. In recent years, however, relational databases have failed to deal with two aspects of modern data: large volumes of data and unstructured data. In order to solve the mentioned problems looser databases with a flexible structure and a more efficient way of working with large volumes of data have been created. In many cases, non-relational databases also called NoSQL databases, have become a replacement for relational databases. Several applications require migration from relational to non-relational databases based on suitable properties while there is number of problems associated with this migration. Moving records from a relational database to a non-relational database requires having a structured methodology for transforming existing data. This data transformation from a relational database to a non-relational database (such as MongoDB), is more difficult due to the non-existent transmission standards. The main objective of this paper is to present a proposal for mapping rules of the relation database schema to the NoSQL database schema, specifically NoSQL database of the key-value type, such as MongoDB. The mapping is performed based on the type of relationship that occurs in the relational database, and this process can also be applied in the opposite direction, from the non-relational MongoDB database to the relational database.**

Keywords—Oracle; MongoDB; Mapping rules; ETL; undefined data

I. INTRODUCTION

The relational databases are being developed since the 1960s, which resulted in a stronger theoretical model, a larger range of features and thus a greater use of these databases. The main property of the relational database is the storage of data in highly structured tables while maintaining a normalized form. These two objectives have become a limitation, which (with the growing number of data) has become an obstacle to their use.

The problem with modern data can be identified as the diversity of data and their non-normalization, which means that objects as such are not structured with the use of same formula, number of properties or the same data types. While working with objects, the relational databases cannot be compared with the non-relational ones.

Potential of working with objects and large volumes of data was understood by notable organizations such as Google, Amazon or Microsoft, who chose NoSQL databases as their primary data storage [1]. With the increasing use of non-relational databases, it became important to find a concept of proper mapping of a schema in relational databases to schemas in various non-relational databases. Proper transformation of schemas between relational and non-relational databases enables the integration of data, which is now common practice. The problem of such mapping is currently a large number of types of NoSQL databases. This is reason for several researchers trying to define different types of rules and different mapping methods based on types of NoSQL databases.

Since we dealt with the non-relational database MongoDB in a large part of our research, we also focus on this mentioned database in the presented paper. Data in the non-relational database MongoDB are based on documents, each of which is identified by a specific key [2]. These documents are grouped into collections, which are stored sequentially, and new documents can be added to any collection at any time [3]. By inserting an object into an object, the objects are gradually nested and thus certain layers in data structure are created. There are two ways to model relationships in document-based NoSQL databases - the relationships based on references and the relationships based on insertion. Referential relationships are similar to relational databases, where user's document ID becomes a foreign key in another document. While using insertion relationships, documents are truly nested in other documents so both can be accessed together.

After applying the basic rules, we use a method for data transformation, which is well known by the abbreviation ETL (Extract, Transform, Load). We create the rule set in such a way that the program is able to apply the rules based on the input data and thus autonomously and without difficulty move the data from one type of database to another.

Since it is not possible to design and implement a general module working for all relational databases such as MySQL, MsSQL or PostgreSQL and then apply mapping rules to various types of non-relational databases, such as column-oriented database, graph model and key-value database, we present rules for mapping of relational database Oracle to the non-relational database MongoDB.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The rest of presented paper is structured as follows:

- The works, which are focused on schema transformation between relational databases and NoSQL databases are presented in the section II.
- Section III contains proposed mapping rules for schema transformation from SQL to NoSQL and back.
- In the fourth section, we present experimental part of the paper – we experimentally tested proposed mapping rules with the use of NoSQL queries and ETL.

II. Related Work

The importance of a structured schema transformation between various types of databases has led many researchers to exploration of solutions for transformation of relational database schema, which is the most commonly used database today. In the last two decades, we have been able to follow a large number of transformation work focused on relational databases, e.g. [4, 5], in order to meet the growing need for semi-structured and unstructured data. In many works on schema transformation, not only the uniqueness of the newly created data structures is taken into account, but also the semantics that can be included in the relational databases are preserved.

Within the issue of record mapping between relational and non-relational databases, it is possible to find several works proposing techniques for transformation of relational databases into column-oriented NoSQL databases. In [6], the authors mapped entities and association relationships in an improved entity relationship diagram to the HBase database using following three rules. In the first rule, a column family is created for each table, and the primary key of each table becomes the row key in the column family. In the second rule, a new column family is added to another column family to become a supercolumn family. For a M:N association relationship, new column families are created in the relational database and inserted into HBase on both sides, which means that the join table in the relational database is deleted. The purpose of this rule is to maintain the referential integrity of the foreign key mechanisms in the relational databases. The third rule reduces foreign keys by merging them into a super column.

There are several works, in which authors proposed a method for schema transformation from relational database to a NoSQL database based on documents. In [7], the authors proposed a framework for implementing an algorithm that used metadata stored in the relational databases to automatically transform entities and association relationships. In [8], the authors used a separate application called MigDB, which parses the tables in the relational databases, creates a JSON file based on the tables, and then passes the JSON file to the neural network. In addition, the network decides the most appropriate structure for mapping of the JSON file – nested structure of referential structure. This work was done only to map association relationships.

In [9] the authors mapped the relationship of 1:M from relational databases to graph-based NoSQL, specifically the database GraphQL. The starting node in the graphs consists of multiple pages, and the primary key of one page is inserted into multiple pages by preserving the primary key as an edge property. The join table in the relational database is not used to store information as a relationship property. While mapping ternary relationship, the join table and foreign keys of the other tables were removed, but the relationship attributes were preserved as a property of the relationship between the nodes in the graph.

In [10] the authors presented the transformation of relational database into several types of NoSQL, specifically into a few key – value databases, column-oriented database, document database and graph-based database. The authors identified the concepts of each database using defined n-tuples. Subsequently, the authors presented algorithms for performing the transformation and a case study as proof of the concept. This work is complete in the sense that it includes all types of non-relational databases. However, it is not clear that all types of relationships are included in the relational database.

In [11], the authors introduced a data adapter used for querying and mapping between SQL and NoSQL databases. The adapter allows queries from the application and deals with the transformation of the database to a server with a relatively low time difference. Although this work implements a data adapter, it does not provide clear rules for the transformation between two types of databases - a mapping between various non-relational databases or a relational and non-relational database.

In addition to the data structure and relationships, several works have recently been published in the area of transformation implementation. In [12], the authors presented a framework, which supports convenient migration from relational to NoSQL database management system. The framework consists of two modules, namely the migration and data mapping modules. Since the work focuses more on implementation, it does not present a clear transformation existing within the data mapping module. Instead, the article presents the results of experiments with various database operations of the mapping results.

Since the lack of the structure can cause not only ambiguity but also the non-definition of certain values, we also had to deal with this issue. In the research, we applied the method presented in the work [13], where researchers dealt with temporal database architectures, which manage undefined values and propose a comprehensive classification system based on transactions, data reading and indexes. The article deals with techniques for modeling undefined values and covers synchronization processes using data groups. Authors also propose solutions for efficient data acquisition with emphasis on undefined values and states.

An interesting study regarding transformation was published in an article [14] where the authors proposed an approach to model transformation and data migration from a relational database to MongoDB. Their work is divided into four sections where in the first part of the work authors take into account the characteristics of the query and the data characteristics of the relational database. Subsequently, in the second part, they propose an algorithm for transforming the model based on description tags and action tags. The third part is focused on the automatic migration of data to MongoDB based on the result of the transformation of the model, and in the last – fourth – part of the paper a transformation tool is designed and implemented.

III. PROPOSED MAPPING RULES

Although the works mentioned in the previous section brought relevant and important ideas to the problem of schema transformation from relational to non-relational databases, a large number of studies contained ideas dealing only with the associative relationship between individual types of databases. Many of these studies do not deal with the loss of referencing in the schema or loss of values, these methods do not apply backwards compatibility and also do not address the change of values after use of proposed application based on proposed rules. In order to focus on solving the problem, we suggest ways to implement mapping rules when checking undefined values, and then we carry out the process of changing values from the relational database Oracle to the non-relational database MongoDB.

In this part of the paper, we present the rules for transforming schemas from relational databases to non-relational databases. The main objective is to present three basic types of relationships between entities - these are relationships of 1:1, 1:M and M:N types.

A. *Transformation of association relationships of One-to-One (1:1) type from SQL to NoSQL*

Since One-to-One relationship is one of basic relationships, we decided to define two relations for the relational database Oracle. These are relations *student* and *college*. The *student* relation consists of three attributes, these are the primary key *Student_ID*, the name of student *Student_Name* and the address of student *Student_Address*. The *college* relation consists of two attributes and they are *College_ID* which also represents the primary key of the college and the attribute of the name of collage *College_Name*.

To connect these relations, as shown in Figure 1, a relationship called *StudyIn* is created. Based on the E-R diagram from Figure 1, we created a rule presented in the lower part of this figure. Since this is a 1:1 relationship, the ratio of student-collage relationship is in represented in the same way - one object is nested into another object.

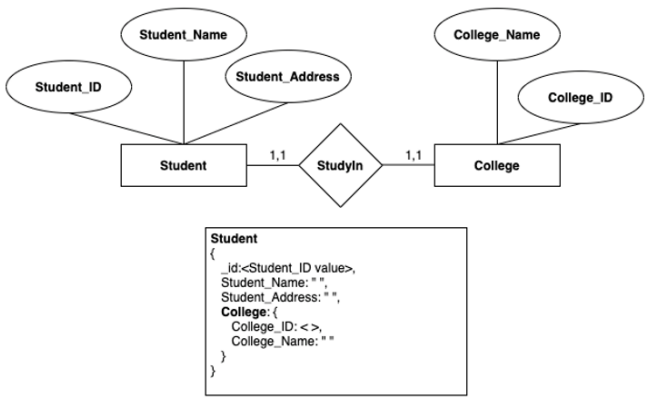


Fig. 1. 1:1 mapping rule for SQL to NoSQL

B. *Transformation of association relationships of One-to-Many (1:M) type from SQL to NoSQL*

The relationships of the type 1:M in relational databases do not complicate diametrically the relationship we presented in the previous step. Since the 1:M relationship differs only in the number of values occurring, the only difference is to increase the number of individual records in json format and apply changes. We present this relationship in the Figure 2 - it is clear, that only one change occurred (in the relationship *StudyIn* where 1 has changed to M). The presented mapping rule is proposed as follows:

- First table is created (in our case it is the table *student*)
- Subsequently, algorithm determines type of relationship based on the number of entities:
  - If number of entities is equal to one, algorithm applies relationships based on the subsection A of this section.
  - In other cases, algorithm adds new objects and creates lists in the newly created collection.

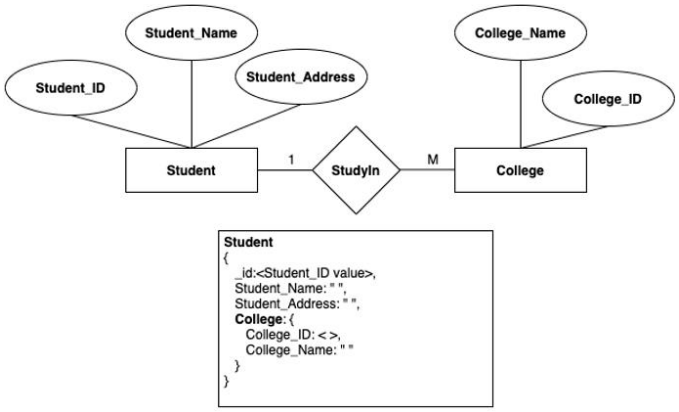


Fig. 2. 1:M mapping rule for SQL to NoSQL

C. *Transformation of association relationships of Many-to-Many (M:N) type from SQL to NoSQL*

The principle of creating a mapping rule for the relationship M:N is straightforward. First, one collection is created, in our case the collection *student*. After creating this collection, the *college* collection is created – this collection already contains values of the primary key of student (specifically the *Student_ID*). Subsequently, a new collection presenting the M:N relationship is created in the *college* collection. Since the *StudyIn* object is a nested object in the *college* collection, it can always retrieve values for references to the *student* and *college* collections. For this reason, we have defined additional values in the *StudyIn* collection – namely the value *City*.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

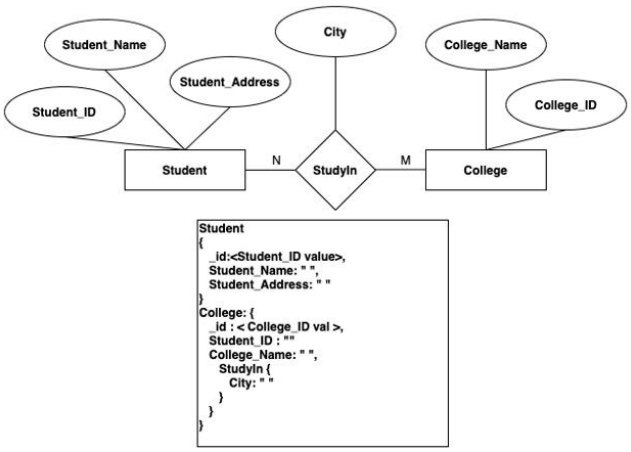


Fig. 3. M:N mapping rule for SQL to NoSQL

Since we only defined the principle of mapping from tables to collections, which means the transformation of the schema from a relational database to a non-relational database, in the next step we create mapping rules in the opposite direction. In this process, the freedom of the structure proves to be an unfavorable property in a backwards processing of schema.

D. Transformation of association relationships of One-to-One (1:1) type from NoSQL to SQL

It is more difficult to design and implement mapping rule when creating a schema from an undefined structure than in the case of creating mapping rules from a relational database to a non-relational one. Since the lack of strict structure of the schema is negative and the data types are key to the database, we need to create a universal model for the change of data types.

With the One-to-One object mapping rule, the number of nested objects in the collection student is verified. In the case the number is equal to 1, then the table student and the table college are created and a 1:1 association relationship is created between them.

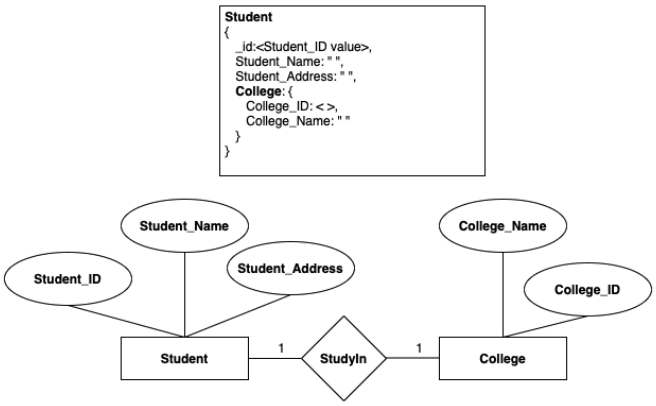


Fig. 4. 1:1 Mapping rule for NoSQL to SQL

E. Transformation of association relationships of One-to-Many (1:M) type from NoSQL to SQL

In the case, that algorithm finds an object in which the nesting of objects takes place, the object in question is verified. If the algorithm detects the number of nested objects greater than one, this suggests that the One-to-One mapping rule is not fit for use with the object - One-to-Many mapping rule is used. In such case, objects from the nested collection are read until the record corresponding to last object of original dataset is created.

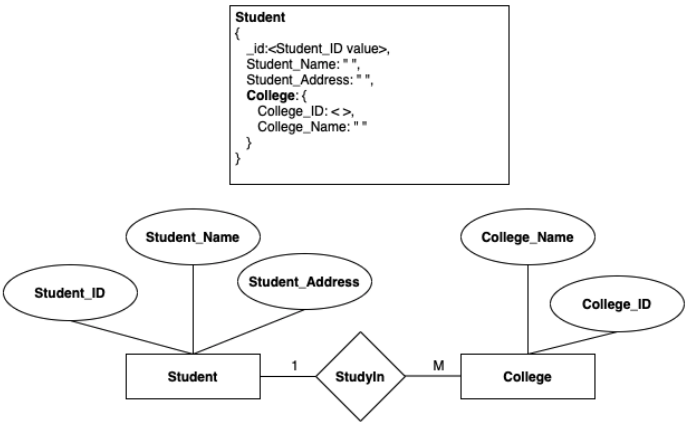


Fig. 5. 1:M Mapping rule for NoSQL to SQL

F. Transformation of association relationships of Many-to-Many (M:N) type from NoSQL to SQL

The transformation of M:N relationships is the most complex when creating a mapping rule due to multiple nesting of objects. In the case of single-layer nesting, as presented in the Figure 6, it is necessary to solve the problem with only one reference to the parent object - finding the collection in the collection and looking at the *_id* value of the parent collections. In the case of multiple nestings, there is a relationship of M:N type, which is associated with another relationship of the same type. There must be multiple use of the relationship F or other relationships applied in sections E and F according to Figures 4 and 5.

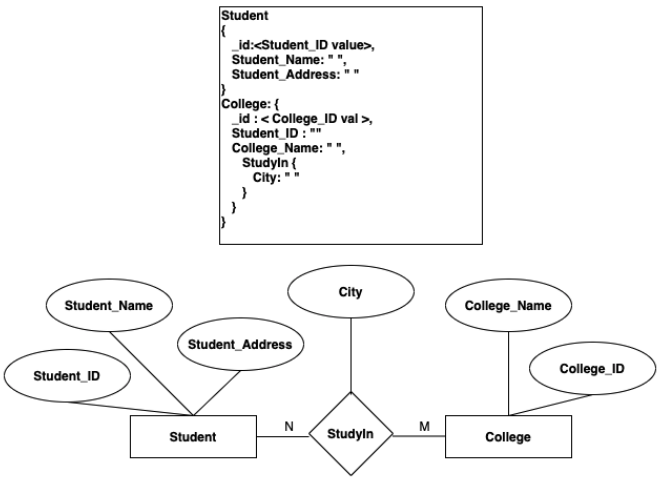


Fig. 6. *M:N Mapping rule for NoSQL to SQL*

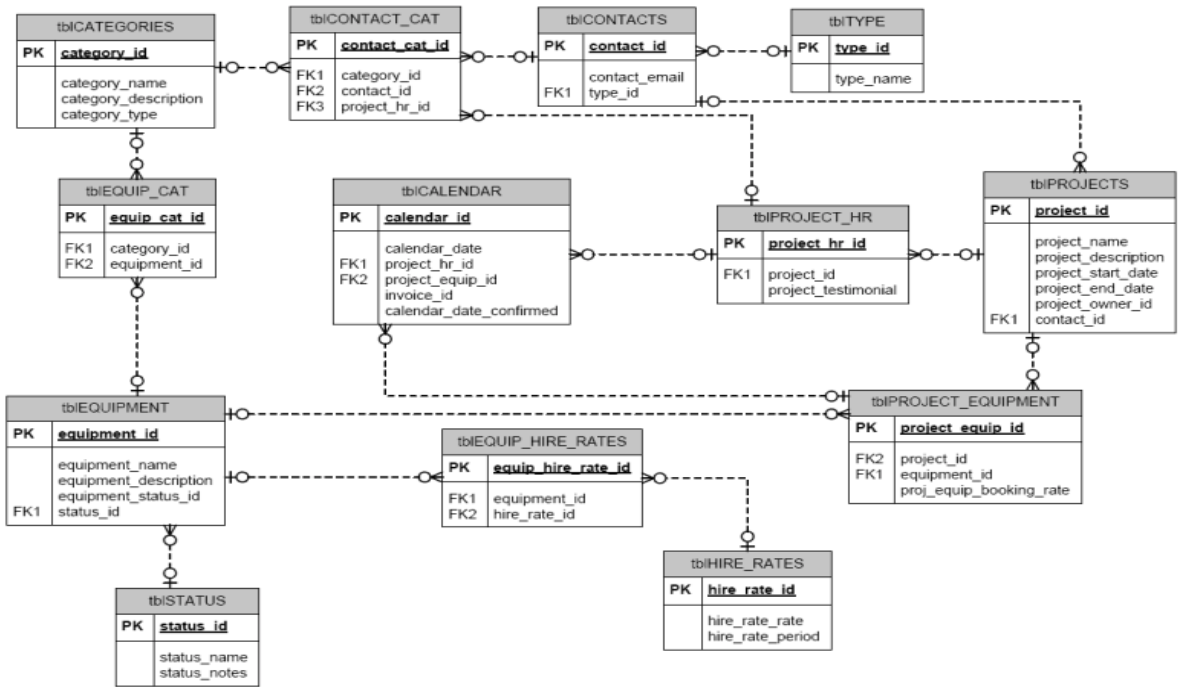


Fig. 7. Complex data model used in the testing of proposed mapping rules

G. Solution to the problem of number of attributes

The lack of strict structure of a data schema is an excellent feature in many respects, but not when creating a schema in a relational database. Since the non-relational database MongoDB does not create a schema, or said more precisely it creates it in a way that is diametrically different that relational, we needed to create a mechanism to manage the number of attributes. Since the number of attributes differs in the non-relational databases, we based our mechanism on the records containing highest number of attributes. That is, the algorithm traverses all the objects in the collection. The algorithm maintains a reference to the object and the number of its properties. If the algorithm finds an object with more attributes, it stores it in a local variable and then continues the search. At the end of the run of algorithm, it contains the object and the number of attributes. The proposed method creates the initial number of objects based on this object.

In the second cycle, the algorithm detects additional attributes and compares them with the attributes of the object in the local variable. If there is an attribute that is not contained in the local variable, the algorithm completes this attribute set and continues until all objects are verified and the remaining attributes are added to the set of attributes. This means, that an object stored in the local variable contains the attributes of all objects of the collection.

In the third cycle, objects and number of values represented in the objects are verified. Since the objects do not contain the same number of attributes (both as each other and as the model object stored in the local variable), algorithm needs to add number of attributes to all objects along with their data type and the name. This represent an operation, which tracks number of uses of various data types in given attribute and based on the number of uses decided which data type is used in the created schema. For example, if the attribute color contains values for 10 objects while six times the values is integer and four times a string, the algorithm assigns data type of the attribute as int. Remaining four values must be type consolidated by ETL.

IV. Experiments For Verification of Proposed Mapping Rules

For testing purposes, we used the model presented in the Fig. 7. The values of attributes are not significant for us at present, since in this paper we do not perform the overall transformation of data but only define the rules, it is not vital to describe them. The purpose of this model is to capture the 1:1, 1:M and M:N relationships. During the initial verification and application of the relationships when mapping the schema from chosen relational database to the non-relational database, the relationships presented in the section III were applied, specifically from subsections A, B and C. Based on these relationships, the schema transformation was performed without further additional modifications.

Since we wanted to verify the backwards compatibility and determine whether a change in structures or a change in data types can affect the whole transformation process, we decided to perform two types of experiments.

The first type of experiment consisted of us moving the schema from the relational database to the non-relational database using the rules defined in the section III. Mapping rules 1, 2 and 3 were enough for us to completely transfer the data scheme. Subsequently, we performed steps based on the proposed rules 4, 5 and 6 and transformed the schema backwards (back to relational database). While comparing pre-transformation and post-transformation schemas, we focused on consistence of schema itself and on consistence of data types.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

There were no differences - the whole structure was the same when checking the schema itself and data types.

However, the problem occurred when transforming the tables presented in the Table 1. These tables were randomly selected, and their values were randomly changed. Other than these changes, new attributes were added. In these cases, the mapping rules were able to cover all the required values when changing the schema from a relational database to a non-relational database (YES values indicate the success of the mapping (SQL to NoSQL column)).

During the experimental work, we changed and modified data types, changed values or added new attributes to the database. When changing these values, we wanted to move the schema to a relational database, and as can be seen from the results in the Table 1 (specifically NoSQL to SQL column), there were problems with this transformation. Even if the rules created by us revealed the change and worked properly, it was necessary to make additional adjustments - modifying the data type for a schema in a relational database to be compatible with original schemas, and it is also necessary to add new attributes to the database.

Our mapping rules demonstrate the detection of incompatibilities and also know how to identify collisions. Based on our research, we can already send simple SQL views to modify the data schema during data transformation, which will be related to the full compatibility of the process.

Table 1. Properties of Mapping Rules

Mappings	**Results of mapping rules**		
	Table	***SQL to NoSQL***	***NoSQL to SQL***
1	EQUIPMENT	YES	YES
2	PROJECT_EQUIPMENT	YES	NO
3	PROJECTS	YES	NO

CONCLUSION

In the presented paper, we proposed a set of rules for transforming a schema from relational database into non-relational database NoSQL, specifically the MongoDB. The rules cover the different types of relationships that can appear in the data stored in relational databases, namely associations, inheritance, and aggregation. Along with the types of relationships, cardinalities are also considered.

After defining the rules, we applied the mapping rules to the case study in the relational database, where we were able to create 16 documents for the non-relational MongoDB database from 14 tables with the correct mapping.

Proposed methodology works based on the following principle. The application exports the data schema and passes all tables on the basis of individual records. Then, it creates mapping rules for individual tables, where the 1:1, 1:N or M:N relationships are present. After all the rules have been created, the record mapping process is computed. The mapping works on the principle of ETL and applies the designed rules to the records which enter the system. After successfully mapping the data, the process ends and the rules are stored in additional storage space. When applying the process in reverse i.e., the process from the non-relational database to the relational one, we had to apply a control rule - the rule for monitoring undefined values, which often occurred in tables while working with the non-relational database.

In order to verify the proposed method, we created a data model which contains 1 000 records for each table. Experimental activities are divided into two parts - the first part is the tracking of records and mapping rules from the Oracle relational database to the non-relational MongoDB database, in which the rule of undefined values did not have to be applied. Otherwise, when we created mapping rules from the non-relational MongoDB database to the Oracle relational database, we had to apply the rule of tracking undefined values and we also monitored the poor compatibility between the changed values compared to the original relational database.

In our future research, we set out to improve the method of verification of undefined values and also to create improved mapper. The new mapper should address compatibility between individual data types, which in our current solution is not perfect and sometimes requires human input into the mapping process. As one of the possible variants, we propose to design and implement a process of mapping to all types of relational databases and then focus on all types of non-relational databases, as the popularity of non-relational databases is constantly growing.

ACKNOWLEDGMENT

The research was partially supported by the grant of The Ministry of Education, Science, Research and Sport of Slovak Republic - Implementation of new trends in computer science to teaching of algorithmic thinking and programming in Informatics for secondary education, project number KEGA 018UMB-4/2020.

REFERENCES

[1] L. Rocha, F. Vale, E. Cirilo, D. Barbosa, and F. Mourão, “A framework for migrating relational datasets to NoSQL,” in Procedia Computer Science, 2015.

[2] V. C. Storey and I.-Y. Song, “Big data technologies and Management: What conceptual modeling can do,” Data Knowl. Eng., vol. 108, pp. 50–67, 2017.

[3] P. Atzeni, F. Bugiotti, and L. Rossi, “Uniform access to NoSQL systems,” Inf. Syst., vol. 43, pp. 117–133, 2014.

[4] E. Pardede, W. Rahayu, and D. Taniar, Mapping Methods and Query for Aggregation and Association in Object-Relational Database using Collection. 2004.

[5] E. Pardede, J. W. Rahayu, and D. Taniar, “Object-relational complex structures for XML storage,” Inf. Softw. Technol., vol. 48, no. 6, pp. 370–384, 2006.

[6] C. Li, “Transforming relational database into HBase: A case study,” in 2010 IEEE International Conference on Software Engineering and Service Sciences, 2010, pp. 683–687.

[7] L. Stanescu, M. Brezovan, and D. D. Burdescu, “Automatic mapping of MySQL databases to NoSQL MongoDB,” in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, 2016.

[8] G. Liyanaarachchi, L. Kasun, M. Nimesha, K. Lahiru, and A. Karunasena, “MigDB - relational to NoSQL mapper,” in 2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), 2016, pp. 1–6.

[9] D. W. Wardani and J. Kiing, “Semantic mapping relational to graph model,” in 2014 International Conference on Computer, Control, Informatics and Its Applications (IC3INA), 2014, pp. 160–165.

[10] M. Freitas, D. Souza, and A. C. Salgado, Conceptual Mappings to Convert Relational into NoSQL Databases. 2016.

[11] Y.-T. Liao et al., “Data adapter for querying and transformation between SQL and NoSQL database,” Futur. Gener. Comput. Syst., vol. 65, pp. 111–121, 2016.

[12] L. Rocha, F. Vale, E. Cirilo, D. Barbosa, and F. Mourão, “A Framework for Migrating Relational Datasets to NoSQL1,” Procedia Comput. Sci., vol. 51, pp. 2593–2602, 2015.

[13] M. Kvet, Š. Toth, and E. Krsak, “Concept of temporal data retrieval: Undefined value management,” Concurr. Comput. Pract. Exp., vol. 32, p. e5399, Jun. 2019.

[14] T. Jia, X. Zhao, Z. Wang, D. Gong, and G. Ding, “Model Transformation and Data Migration from Relational Database to MongoDB,” in 2016 IEEE International Congress on Big Data (BigData Congress), 2016, pp. 60–67.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Improved method of selecting data in a nonrelational database

Roman Ceresnak
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
roman.ceresnak@fri.uniza.sk

Michal Kvet
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
michal.kvet@fri.uniza.sk

Karol Matiasko
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
karol.matiasko@fri.uniza.sk

***Abstract*— People surround themselves with data in many ways, which evokes the need for correct ways of storing the data. Nowadays, the trend tends to lean in favor of data storing in nonrelational (or NoSQL) databases. These databases are used in various user applications, which need a huge volume of the data highly accessible and do not require big data consistency. The problem of the data growth and its storing in the nonrelational databases results in the decreasing efficiency of searching in data. In the paper, we present use of very popular in-memory database in order to help us with this lack of efficiency of the data searching. This paper examines the data searching in applications hosted by Amazon cloud service while using nonrelational database DynamoDB. We develop new procedures to provide faster response to user and to obtain the data using nonrelational database DynamoDB. These procedures provide the queried data and subsequently, transfer them into the memory. The given procedures are based on two methods. The first method is a recognition of values, the user refers to and the provision of this data to the in-memory database. The second method is related to the automatic storing of the data transferred to the in-memory database. We perform number of experiments, which are describing a limitation of efficiency/inefficiency from a perspective computational time.**

Keywords—Data Searching, NoSQL databases, in-memory databases, DynamoDB

I. INTRODUCTION

The growth of population and amount of produced data caused, that many conventionally used procedures and systems, like traditional databases, started gradually losing their efficiency. In contrast with relational databases [2], NoSQL databases process and manage the big data, characterized by 3V (volume, variety, velocity) [3]. NoSQL databases are critical in support of various applications, which need various levels of performance consistency, availability, and scalability [4]. Social media such as Twitter and Facebook [5] generate exabytes of the data daily, which exceed processing capabilities of relational databases. These applications demand high performance, but they do not have to demand strong consistency.

It is impossible to achieve the same efficiency of searching in a large amount of the data while working with the nonrelational and relational databases. Regarding the searching in the nonrelational databases, the data, which do not have to meet the strict structural demands of system (RDMBS), are stored, because the data for the searching can be texted, semi-structured or unstructured. A search engine database is created to help the users in fast finding of the information they need in a highly qualified and cost-effective way. These databases are optimized for the use of keywords and usually offer specialized methods such as full-text searching or searching with the use of complicated expressions of various types.

Search engine database consists of two main parts. The first part is adding a search engine database index to the data. When the user queries for data, relevant results are quickly returned with the help of the search engine database index. This fast and responsive way of data searching is possible since instead of direct searching for queried text, these databases search for relevant index in the database. This can be thought of as an equivalent to looking for page number related to the term in book index, in contrast to searching for individual words on every page in the book. This type of index is called an inverted index because it transfers the data structure-oriented on a page to the data structure oriented at the keywords.

Second part of the search engine database is the data searching can be made more efficient with the use of in-memory databases. The in-memory database (IMDB) is a computer system, which stores and searches the data records situated in the main memory of computer e.g., in RAM memory. IDMB is an advantageous approach to the data storing since traditional databases work with a data access delay as a result of storing data on medias with higher time of access such as hard discs, SSDs and so on. This means, that IDMB is useful when fast reading and recording of data is crucial.

Most of IMDB implementation preserve data in the RAM. Some implementations use IDMB with the combination of disk part of the system, but RAM is still primary storing medium. Some IDMBs also store the data on disk as a preventive measure to minimalize the risk of the data loss, since RAM is volatile e.g., the data is lost when a computer loses electric energy.

The majority of IDMB also prevents the data loss in chosen data center (property known as “high availability”) preserving the copies (technically called replicas) of all the data records on several computers in a cluster. This data redundancy secures, that when any kind of error makes whichever computer in the network not available, no data record could be lost. Among the most popular in-memory databases, which use query languages for data searching, belong databases such as Redis, Memcached and similar products. Artificial intelligence can be used in order to help us with the purpose of transfer of the data situated in the nonrelational database DynamoDB.

The problem related to the data transfer from the conventional disk-based database to the in-memory database is known as velocity of the data searching and the transfer efficiency. Artificial intelligence was used for these purposes, which secures this transfer and so it also makes the data

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

searching more effective. The main objectives of this paper are as follow:

- We create a new procedure for processing the data in the memory,
- We reduce the time needed for the record searching in the nonrelational databases,
- We define the methods of automatic adjustment of the data growth to the size of in-memory database.

The rest of the paper is structured as follows. "State of the art" section examines the work related to objective of presented paper. Part III presents our proposed solution - designed searching model and the characteristics of this model. "The data transfer" and "The index creation" parts describe the performance of the operation in DynamoDB. "Experiments" part describes performed experimental testing and subsequently their results.

II. State of the art

A comparison between relational data models and nonrelational (NoSQL) data models was already stated in various papers. For example, comparisons between these two types of database were focused on the times needed to perform basic database operations such as data selection, data insert, data update, and deletion of data [1].

Several statistics point out the fact the most common operation, which is demanded in the relational and nonrelational database, is data selection operation [1, 2]. Number of authors presented in their papers [3], that the time needed to get the data in the nonrelational database is significantly higher as the time needed to get the same data in the relational database. Computational time of operations is crucial not only in database systems, but all problems related to the computer science [4, 5], hence need for optimization of conventional approaches. As a way to accelerate or improve the time needed to get the data from the nonrelational database, it is possible to store the data to buffer memory and by this reduce repeated searching in the nonrelational database. With the use of this method not only computational time is reduced, but also number of accesses to the database is lowered [6].

The authors performed various comparisons between in-memory databases such as Redis, Memcached, and nonrelational databases Mongo, Casandra, and H2 [7]. One of the main findings of these works is the verification of data update and deleting of the data with its increasing amount. During our research related to the topic of the paper, we noticed, that papers focus on problem-solving in the context of increasing amount of the data.

In the authors created a module by using the library Lontar, which sends the data to the relational database with the use of Hibernate as a framework and a relational mapper, in the case of user's demand. Subsequently, Hibernate accesses the MySQL database and maps the relational data to object-oriented [8], and then it sends the data to the nonrelational database. The searching then works with the help of the mapper, so Lontar is able to read the data in a relationship. According to the authors, searching for data in chosen data files resulted in better times for nonrelational database MongoDB than for relational database MySQL. However, in certain situations, the relational database got better results than the nonrelational database.

The authors of [9] introduced a framework capable of data manipulation in order to overcome the problems related to the decreasing efficiency of the searching in the nonrelational databases. Before performing the basic operations of data selection, data insertion, data update, and data delete this framework uses mapper function. The main role of the mapper is to change the data on the base of rules, to such form, which complies with principles of nonrelational database MongoDB more. With the use of this module, the speed of data searching is in majority of cases higher in nonrelational database MongoDB compared to the relational database MySQL. Another concept used in this framework is the Cataloging module, which uses JSP (JAVA) as a web programming technology and MySQL as DBMS. There are two frameworks supporting this concept, Structs and Hibernate. Structs are used to set users' interface and the Hibernate regime is used to map the relational data to the object-oriented data, which will be used by JSP.

In our work, we also design and implement framework, which uses two database types. The first database is nonrelational database DynamoDB serving as a primary data storage. In the case, when user demands data, the values will not be directly given to them from the nonrelational database, but the values will be transferred to the in-memory database. The main challenge in this approach is to transfer the data from the nonrelational model to the in-memory model.

III. Our Contribution

In this part of presented paper, we introduce two modules which are used in order to make the searching in the nonrelational databases more effective. Data Cached Module (DCM) and Data Elastic Module (DEM). DCM serves as a data storehouse, whose role is the data transfer to the buffer memory. DEM serves as a tool for automatic adjusting to the data size.

A. Data Cached Module

The created module serves as a way of the data processing in the memory. We connected the data we store in the nonrelational database DynamoDB to highly available buffer memory Amazon DynamoDB Accelerator, well known in short as DAX, with the help of API interface. This method helps us with Side-cache access, Read-through cache access and Write-through cache access as follows:

a) Side-cache – This principle helps us with high overload during the reading of information from the memory. The principle works as follow (see Fig. 1):
   1. An application first tries to load the data for a given key-value couple from the buffer memory. If the buffer contains queried data, the value of this key-value pair is returned as and output of the operation. Otherwise, step 2 follows.
   2. Since the demanded key-value pair was not found in the buffer, the application loads the data from the basic data storage.
   3. A key-value pair from step 2 is written to the buffer memory to make sure the data are present when the application needs to load the data again in the future uses of similar query.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

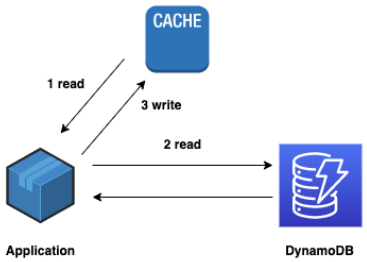


Figure 1. Side-cache algorithm

b) Read-through cache – DAX is a buffer memory for the reading - it is compatible with API for the reading of DynamoDB and stores the results *GetItem, BatchGetlem, Scan*, and *Query* to the buffer memory, if they are not currently in DAX. The buffer memory for the reading is effective in high working loads. This principle works as follow (see Fig. 2):
   1. Regarding a key-value couple from the application, algorithm first tries to load the data from DAX. In the case the buffer contains queried data, the value of this key-value pair is returned as and output of the operation. Otherwise, step 2 follows.
   2. Transparently for the application, if a semi-memory happens, DAX will load a couple of key-value from DynamoDB.
   3. To make the data available for every reading that follows, a key-value pair is stored in the semi-memory of DAX.
   4. A key-value couple is then returned to the application.

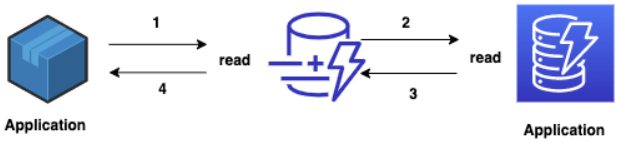


Figure 2. Read-through cache algorithm

c) Write-through cache – Similarly to the semi-memory for the reading, a semi-memory for the data writing also operates in a line with the database and updates the semi-memory, when the data is written to the basic data storage. DAX has also buffered memory for writing since it stores to the buffer memory (or updates) the items with *PutItem, UpdateItem, DeleteItem* and *BatchWriteItem* API, because the data is written or updated in DynamoDB. At first, DAX is updated (everything is transparent for the application). The following steps indicate a procedure for buffer memory of write-through type (see Fig. 3):
   1. The application will write itself to endpoint DAX for a given key-value couple.
   2. DAX will catch the writing and then will write a key-value pair into the DynamoDB.
   3. After the successful writing, DAX hydrates buffer memory with a new value so whichever following the reading of the same couple key-value results in a finding of the buffer memory. If the writing is unsuccessful an exception will return to the application.
   4. Confirmation of successful writing will then return to the application.

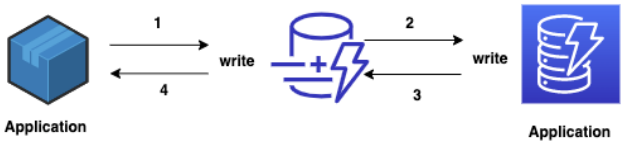


Figure 3. Write-through cache

B. Data Elastic Module

The created module works with the data which increases demands for the storage space. Nonrelational database DynamoDB fulfills the task of a wide data storage and in the case of the data transfer to the in-memory database, size of the data can grow from several megabytes to several gigabytes very easily. This problem is solved with the use of the created module.

We configured monitoring of metrics in cloud service Amazon with the help of Amazon CloudWatch service. The mentioned service makes it possible to edit (to add and to remove) new computation units in the case of enabling of horizontal or vertical scaling. An advantageous characteristic of this method is the horizontal scaling which, in the case of a large number of the data uploads, invokes warning of system overload and a script for the reading of information from other replicas in service CloudWatch. The horizontal replica is the part of the script performed automatically during the configuration of the in-memory database DAX with the following script:

```
aws dax decrease-replication-factor \
  --cluster-name MyNewCluster  \
  --new-replication-factor 3
```

The monitoring of the metrics in the same way with our method also makes the vertical scaling possible - the scaling by addition or removal of the computation units (Fig. 4).

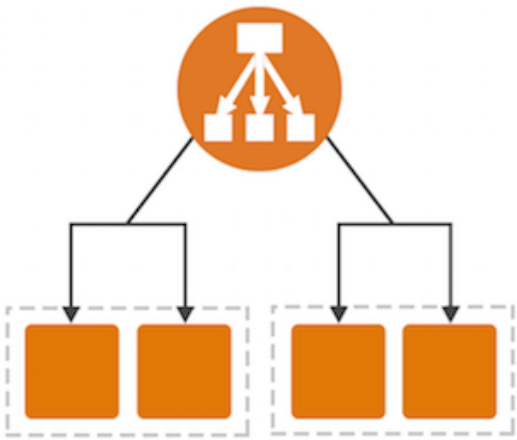


Figure 4. Application Load Balancer for in-memory database

In a situation, when the values retrieved from the nonrelational database do not fit the in-memory database, the event is invoked again with the help of service CloudWatch, which will cause the addition of a new computation unit. When the computation unit is not needed anymore, an instance is automatically released, which results in saving of the buffer memory and optimization of a price.

IV. EXPERIMETNS

In the very first step, we created a simple database model presented in the Fig 5. This database model consists of two tables - *user* and *comment*. These two tables are interconnected by identification relationship of type *1:n*, which means one user can create number of various comments and various comments in the table belong to one user.

Subsequently, we compared various commands, whose aim is to retrieve information from implemented database model. The objective is to measure time of computation of simple queries in conventional systems.

In the sections A and B, we present measurements for four sizes of queries in the relational database Oracle (section A) and nonrelational database DynamoDB (section B).

Since an important aspect of this paper is use of the in-memory database, we also compare the times of various operations during data selection in section C.

In the section D, we compare these conventional approaches to the problem with proposed solution described in the secion III.

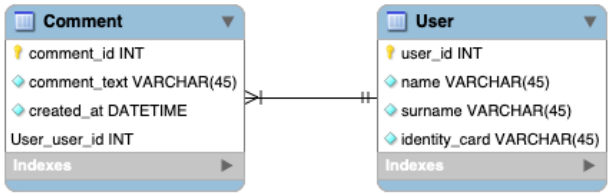


Figure 5. Database model

A. Experiments on relational database Oracle

We inserted 1000 records into table *user* and 1000 records into table *comments* which are based on the defined structure. In the case of this paper, we are interested only in information about the time needed for record searching. We created 3 data selection commands for these purposes:

```
(1) SELECT name, surname FROM user
    JOIN comment USING (user_id);

(2) SELECT * FROM user
    JOIN comment USING (user_id)
    WHERE comment_text LIKE "%today%";

(3) SELECT  name, surname,
    to_char(created_at, 'YYYY-MM-DD HH24:MI:SS') ca
    FROM user
    JOIN comment USING (user_id)
    WHERE ca >= TRUNC(current_date)
    and ca < TRUNC (current_date) + 1;
```

The first 1000 records served as benchmarking records to us - from these results, we continued our research. The main purpose of the nonrelational databases is to effectively store large amounts of data, and that is, why the records for other purposes will be created with the sizes of 100 000 records for table *user* and 100 000 records for the table *comment*. Subsequently, all records are deleted, and the dataset of size 10 000 000 records will be inserted to both tables. As the last size of the datasets, we chose a value of 100 000 000 records for *user* and *comment* tables.

A generator was used for record creation with the size of 1 000, 100 000, 10 000 000 and 100 000 000, which can be found on the following address:

https://www.generatedata.com/

The generator provides an option to define names and types of attributes and to generate an arbitrary number of the values. After fulfilling the tables by the generated values, we recorder the times needed to perform operations (1), (2), and (3). These measured times are presented in the Table 1. All measured values for processing of commands (1), (2), and (3) are stated in seconds.

Table 1. Measure time for operations (1), (2) and (3)

Count of records/operation	1 000	100 000	10 000 000	100 000 000
(1)	0,0020	0,004	0,028	0,44
(2)	0,0021	0,0042	0,031	0,45
(3)	0,0021	0,0044	0,030	0,45

B. Experiments on nonrelational database DynamoDB

We used commands (1), (2), and (3) to find out the velocity of data queries in the nonrelational database DynamoDB. The values inserted into the database were left the same as in experiments in the relational database. The structure is fully the same as is presented in the Fig. 3.

Table 2. Measure time for operations (1), (2) and (3) in DynamoDB

Count of records/operation	1 000	100 000	10 000 000	100 000 000
(1)	0,0035	0,0064	0,047	0,82
(2)	0,0035	0,0067	0,048	0,83
(3)	0,0037	0,0068	0,046	0,82

The values, needed to get the data from the nonrelational database DynamoDB are presented in the Table 2. All measured values for commands (1), (2), and (3) are stated in seconds.

When comparing values of computation time of the same operations between the relational and nonrelational databases, we can clearly see the significant difference. We

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

can conclude that data selection operation in nonrelational databases is less effective than in the relational database Oracle.

C. Experiments for the in-memory database Redis

The data storing in the in-memory database is diametrically different than in the relational or nonrelational databases. Except for the mentioned fact, there is also problem with the amount of data caused by computer memory limitations. The computer memory used for testing purposes was about 8 GB.

Table 3. Structure of the data in the in-memory database

ID	Name	Surname	Identity_card
1	John	Harper	12341324
2	Joe	Bush	12341234
3	George	Obama	23524675
.....	.....	.....	.....
.....	.....	....	.....
1000	Alan	Felps	45674866

As seen in the Table 3, we created records with the identical structure to the previously used table *user - ID, Name, Surname* and *Identity_card*. Specifically, we create datasets of 100, 300, 500 and 1000 records.

Three commands were created for the purposes of the testing of the in-memory databases' effectiveness. These commands were structured as follows:

(4) MGET Name

(5) MGET Name, Surname

(6) MGET Name, Surname, Identity_card

(7) MGET Name, Surname, Identity_card, Age

We applied the same principle during filling the database as in previous steps. In the specific case, we inserted the generated values to the database, tested operations (4), (5), (6), and (7), and recorded the measured values. Subsequently, we deleted all the records and inserted the next dataset of 300 records to the database. We continued in this fashion until the size of 1000 records in the database was reached. The measured values for the operations are presented in the Table 4.

Table 4. Measure time for Redis database

Count of records/operation	100	300	500	1 000
(4)	0,00020	0,00022	0,00021	0,00028
(5)	0,00021	0,00023	0,00025	0,00029
(6)	0,00021	0,00022	0,00025	0,00028
(7)	0,00023	0,00024	0,00028	0,00032

All measured results in the Table 4 are presented in seconds. The seventh operation is influenced by the fact, that value "age" does not exist. As can be seen, the measured values are not diametrically different with the increasing number of records. It is necessary to point out, that with defined growth of the records, it is the logic fact mirroring the efficiency of the searching in the memory.

D. Comparison of conventional methods and proposed method

The values we measured in the experimental activity presented in the sections A, B, and C serve for comparison of conventional methods with the proposed method. The compared values operation (1) on the datasets of 1 000 and 1 000 000 records are presented in the Table 5.

Table 5. Comparison of query performance

Count of records/operation	1 000	1 000 000
Oracle	0,0020	0,44
DynamoDB	0,0035	0,82
Our Approach	0,0033	0,42

As seen in the Table 5, the values measured while using operation (1) with a low number of records in the table, do not hint towards any big improvement of the searching in the nonrelational table. This is influenced by the data transfer to the memory. A factor of the transfer indicates a necessity to transfer the data from nonrelational database DynamoDB to buffer memory DynamoDAX, which takes a certain time which is combined with the computational time of the query processing itself. This means, that in the operation of data selection, the data are physically retrieved from in-memory database, not from the nonrelational database DynamoDB.

Based on the data transfer, it was possible to also compare the measured times of the experiments with the in-memory database and the data transferred to DynamoDAX with the size of 100 and 500 records and with the use of operation (5).

Table 6. In memory query performance

Count of records/operation	100	500
Redis	0,00020	0,00021
DynamoDAX	0,00015	0,00017

The values recorded in the Table 6 show efficiency of the data transfer. As cane be seen, the values in buffer memory Dynamo DAX are more effective from the time perspective than in-memory database Redis.

Whole achieved results related to operation data selection in nonrelational database DynamoDB were not, before the application of our method, timely the same effect than after the application of our method. With the use of machine learning and transferring the data to the database in memory, the efficiency of operation data selection in the nonrelational

database became more effective after achieving 1000 000 records than with the searching of the data in relational database Oracle. A huge advantage, that results in using of cloud storage Amazon, is related also to the possibility of automatic scaling respectively adding of performance and increasing of the storage not only in nonrelational database DynamoDB, but mostly in the database in memory alternatively, if we do not need as many calculation units, so the reduction of the size of the data storage happens, and so the decreasing of the cost related to running of our designed method happens.

Conclusion

NoSQL databases play a significant role in storing and processing large amounts of data and they are used in various wider social applications such as Twitter, Facebook, Google, and Yahoo, but they help also with the decision support or in the advanced analyses. These databases became the master of high effectiveness of storing and availability of the large datasets. With this came the loss of the effective searching methods, which can be found in the traditional databases. This paper was focused on the question of the data searching time optimization in NoSQL databases, specifically DynamoDB in the cloud environment of Amazon.

In this paper, we developed the data searching algorithm, which can make the searching in a nonrelational database DynamoDB more effective. The designed algorithm is composed of two parts, the first part is based on the principle of caching the data from the nonrelational database DynamoDB to buffer memory DynamoDAX. The second part is based on the effective data and system management - in the case of the large amount of data, system automatically increases the number of computation units of the buffer memory, and by this adjusts the size of the database to the increasing needs of the size of incoming data. This fact relieved us from the limitation of the database size towards the data.

The experiments provided us with useful information about the performance and the effectiveness of the created method. It is noted, that the system for processing artificial intelligence demanded higher overhead costs together with automatic creation of the database in memory, but this system was able to make the process of searching in the nonrelational database more effective. On the basis of the experiments, it is clearly seen, the created method is more and more effective with the increasing data amount, which is done by the data transfer to the memory.

Our future work will focus on a generalization of this model and provision of user interface for full use of the created procedure not only for Amazon cloud but also for other in-memory databases such as Redis and Memcached. We also plan to evaluate the suggested systems empirically, from a perspective of consistency and performance in other environments, which need fast response for the data demand.

Acknowledgment

This publication was realized with support of Operational Program Integrated Infrastructure 2014 - 2020 of the project: Intelligent operating and processing systems for UAVs, code ITMS 313011V422, co-financed by the European Regional Development Fund.

EUROPEAN UNION
European Regional Development Fund
OP Integrated Infrastructure 2014 – 2020

MINISTRY
OF TRANSPORT
AND CONSTRUCTION
OF THE SLOVAK REPUBLIC

References

[1] R. Čerešňák and M. Kvet, “Comparison of query performance in relational a non-relation databases,” *Transp. Res. Procedia*, 2019.

[2] Y. Li and S. Manoharan, "A performance comparison of SQL and NoSQL databases," *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, 2013, pp. 15-19, doi: 10.1109/PACRIM.2013.6625441.

[3] T. N. Khasawneh, M. H. AL-Sahlee and A. A. Safia, "SQL, NewSQL, and NOSQL Databases: A Comparative Survey," *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, pp. 013-021, doi: 10.1109/ICICS49469.2020.239513.

[4] Dudáš A., Škrinárová J., Vesel E.: Optimization design for parallel coloring of a set of graphs in the High-Performance Computing. In: Proceedings of 2019 IEEE 15th International Scientific Conference on Informatics. pp 93-99. ISBN 978-1-7281-3178-8

[5] Dudáš A., Škrinárová J.: Edge Coloring of Set of Graphs with The Use of Data Decomposition and Clustering. In: IPSI Transactions on internet research : multi-, inter-, and trans-disciplinary issues in computer science and engineering. Vol. 16, no. 2 (2020), pp. 67-74, ISSN 1820-4503

[6] P. T. Hulina and A. R. Hurson, "Reducing average access time of a parallel memory in a database environment by data permutation," *Twenty-Third Annual Hawaii International Conference on System Sciences*, 1990, pp. 65-71 vol.1, doi: 10.1109/HICSS.1990.205101.

[7] I. Pelle, J. Czentye, J. Dóka and B. Sonkoly, "Towards Latency Sensitive Cloud Native Applications: A Performance Study on AWS," *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, 2019, pp. 272-280, doi: 10.1109/CLOUD.2019.00054.

[8] H. Wang, Q. Zhu, J. Shen and S. Cao, "Web-Service-Based Design for Rural Industry by the Local e-Government," *2010 International Conference on Multimedia Information Networking and Security*, 2010, pp. 230-235, doi: 10.1109/MINES.2010.58.

[9] G. Karnitis and G. Arnicans, "Migration of Relational Database to Document-Oriented Database: Structure Denormalization and Data Transformation," *2015 7th International Conference on Computational Intelligence, Communication Systems and Networks*, 2015, pp. 113-118, doi: 10.1109/CICSyN.2015.30.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

On Graph Coloring Analysis Through Visualization

Adam Dudáš
Department of Computer Science
Faculty of Natural Sciences
Matej Bel University
Banská Bystrica, Slovakia
email: adam.dudas@umb.sk

Jarmila Škrinárová
Department of Computer Science
Faculty of Natural Sciences
Matej Bel University
Banská Bystrica, Slovakia
email: jarmila.skrinarova@umb.sk

Adam Kiss
Department of Computer Science
Faculty of Natural Sciences
Matej Bel University
Banská Bystrica, Slovakia
email: adam.kiss@student.umb.sk

***Abstract*—The focus of the presented article is put on the analysis of edge coloring of selected sets of graphs - we are specifically interested in edge 3-coloring of graphs called snarks. Previous research suggests, that while using a single coloring algorithm and using various initial graph coloring edges, coloring of such graph may take anywhere from time lower than one millisecond to the time ranging in hundreds of milliseconds. In our case, we use recursive backtracking coloring algorithm based on breadth-first search and implement the change of initial graph coloring edge via permutation of adjacency matrix of graph. In this article, we present a tool created for the needs of analysis of edge coloring of graphs which is based on visualization of edge coloring and we present several problematic subgraphs and patterns which increase the time of edge coloring of cubic graphs.**

I. Introduction

Graph theory includes number of problems which can be solved by operations on graphs. In our research we deal with edge coloring of graphs, which is relevant in several areas - scheduling, radio frequency allocation, compiler optimization or SAT solvers. [1], [2].

Edge coloring of graphs is an NP-complete problem [3], which can be simply defined as assigning colors to the edges of given graph. In the presented research, we specifically deal with proper edge *k*-coloring of selected sets of graphs - edge *k*-coloring is proper when no adjacent edges are colored with the use of same color of *k* colors used for given graph (see Fig.1).

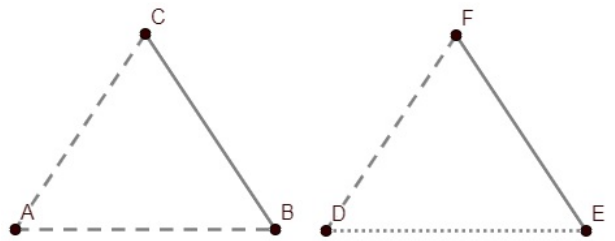


Figure 1. Example of improper (left) and proper coloring of the same graph

Fig. 1 presents graph colored with two different edge colorings - on the left is the graph described by the vertices *(A, B, C)*, which is edge colored improperly. Coloring of edges incident to vertex *A* is problematic since the two edges are colored with the use of the same color (in this case represented by a dashed line). On the same graph, now described by the vertices *(D, E, F)* the proper edge *k*-coloring is presented - where *k* is equal to three - therefore, we use three colors represented by dashed line, dotted line and solid line.

We use a specific type of graphs, which are suitable for our purposes - a group of cubic graphs (graphs in which each vertex is incident with exactly three edges) called snarks. Snarks are cubic graphs, which cannot be colored with the use of three colors [4]. However, there is no algorithm which is able to compute the "non-colorability" of a graph directly - the standard procedure for determining whether a cubic graph is a snark or not is such that the graph is edge-3-colored by possible colorings until we exhaust all possibilities. In the case, that at the end of this process the graph is edge 3-colored properly, then it is not a snark, otherwise the graph in question is a snark. Such coloring is thus an extreme case of edge coloring, in which it is necessary to recolor the graph several times.

In the research presented in this article, we use a simple recursive edge backtracking algorithm. It is known that while using this algorithm, the time of computation of edge *k*-coloring of a graph depends on the initial edge of coloring of the graph. The differences in these computational times of the edge *k*-coloring of the graph range from less than one millisecond to several hundred milliseconds.

This article presents a tool created for the needs of analysis of this edge coloring of graphs based on visualization of edge coloring and searching for subgraphs and patterns, which prolong computation of the problem. Presented paper is structured as follows:

- In the section II of the paper, we present our previous research in the area and research of other authors, which is relevant to problems described in the paper.
- Section III is focused on the principle of proper edge k-coloring of graph, cubic graphs, edge backtracking algorithm used in the proposed solution and change of initial edge of coloring of graph.
- Section IV describes design and implementation of proposed tool for edge coloring visualization purposes.
- In the section V, we compare proposed tool with other, comercialy available tools focused on graph visualization and present several subgraphs, which are problematic in the edge k-coloring of graph.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

II. Related Works

Graph coloring is NP-compelete problem, which can solved with the use of several algorithms such as:

- Edge-color algorithm presented by author of [5]. This algorithm uses polynomial space which improves over the previous, O($2^{n/2}$) algorithm of authors Beigel and Eppstein [6]. Author of [5] uses natural approach of generating inclusion-maximal matchings of the graph.
- Different approach to the graph coloring was introduced by authors of [7] who present simple but empirically efficient heuristic algorithm for the edge-coloring of graphs. The basic idea of this algorithm is the displacement of so called conflicts (adjacent edges colored with the use of same color, see. Fig 1) along paths of adjacent vertices whose incident edges are recolored by swapping alternating colors - with the use of Kempe interchange.
- Simple backtracking approach to edge coloring of graph, which uses recursive functions was presented in several research articles and publications [8], [4]. We describe this algorithm in detail in the section III of this paper.

Research presented in this article is continuation of research related to use of parallel and distributed computing in coloring of cubic graphs and visualization of graphs presented in:

- In the paper [9], we briefly introduced the proper edge coloring of graphs in the context of parallel and distributed computations while using various initial edges for coloring of the graph implemented via permutation of adjacency matrix of graph.
- The paper [10] presented use of adjacency matrix permutation as a way to minimize time of computation of proper edge coloring of large sets of graphs.
- Research presented in the [11] was focused on the visualization of graphs with specific objectives - clarity of diagram of graph, possibility of various input formats for graph visualization and possibility of simultaneous visualization of sets of graphs.

III. Edge Coloring of Cubic Graphs

Graph *G*, as we are considering it in the scope of the presented research, consists of [12]:

- vertices - elements of set *V(G)*. In the Fig. 1 vertices are labeled with capital letters A, B, C, D, E and F.
- edges – elements of set *E(G)* - edge is connection between two vertices and we can label it with the use of labels of these two vertices.

Therefore, graph *G* is pair of sets *V* and *E*, where elements of the set *E* are double element subsets of the set *V* [13]:

$$G = (V, E), E \subseteq [V]^2 \tag{1}$$

The concept of degree of vertex represents number of edges incident to the given vertex (since we work strictly with undirected graphs, by incident, we mean connected to the vertex in any way). In the case vertex V is of degree equal to three, we use notation *deg(V) = 3*. Highest (maximal) degree of vertex in graph *G* is denoted by $\Delta(G)$. In this paper we consider strictly cubic graphs. Graph is cubic when all of it's vertices are of degree equal to three.

Main objective of presented research is edge coloring of cubic graphs - operation of assignment of colors to individual edges of graph. Coloring is called proper when there is no conflict in the coloring of given graph, this means that no vertex of given graph is incident to two or more edges colored with the same color. Lowest number of colors usable in proper edge coloring of graph *G* is called edge chromatic index of graph G. For this property of graphs, we use notation $\chi'(G)$ [13].

Vizing's theorem [13], which says that minimal number of colors needed for coloring of graph is in the interval $<\Delta(G), \Delta(G)+1>$, holds true. Formal notation of Vizings' theorem focused on minimal number of colors:

$$\Delta(G) \leq \chi'(G) \leq \Delta(G) + 1 \tag{2}$$

where $\Delta(G)$ is maximal degree of vertex in the graph *G* and $\chi'(G)$ is edge chromatic index of the graph *G*. Since every vertex of cubic graph is of degree equal to three, we consider three or four colors for proper coloring of cubic graph.

There is only small group of cubic graphs which need four colors for their proper edge coloring. Graphs from this group of cubic graphs are called edge 3-uncolorable graphs or snarks [4].

Chromatic index of snarks is $\chi'(G) = 4$. In order to find out whether given graph *G* is snark, we need to edge color it with the use of three colors. Therefore coloring algorithm needs to verify every possibility of edge 3-coloring of the graph *G*. Algorithms are able to decide whether the graph is snark or not after checking all possible edge colorings with the use of three colors.

The subsection A servers as an introduction to algorithm used for proper edge 3-coloring of graphs and in the subsection B, we introduce principle and method for the change of initial edge of coloring.

A. Edge Coloring of Cubic Graphs Using Edge Backtracking Algorithm

In the presented paper, we use algorithm on the basis of breadth-first search called edge backtracking algorithm as algorithm for edge coloring of graphs.

Edge backtracking algorithm works on the basis of edge coloring of graph with predetermined succession of three colors. In the case, that algorithm finds conflict in the coloring of graph, it backtracks to the previous edge, recolors the edge and continues in coloring. If there are no possible proper colorings of problematic edge, algorithm backtracks even further back - to the edge which precedes both of the recolored edges.

Algorithm continues in this approach until either whole graph is colored properly, or until algorithm examines all possible ways of edge coloring of given graph.

Time complexity of edge backtracking algorithm is $O(2^{n-1})$, where *n* is number of vertices of given graph.

Algorithm itself is represented in these steps:

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

1) Algorithm takes three colors and colors consecutive edges of graph until:
   - either graph is colored properly,
   - or there is conflict in the graph coloring.
2) In the case of conflict in the coloring of graph, algorithm backtracks to edges which were already colored and recolors them with the use of next color in predetermined succession of colors until:
   - either conflict is solved – in this case, algorithm can continue in further edge coloring of graph as stated in the first step of the algorithm.
   - or all possibilities of edge coloring of the graph are improper. In this case the colored graph is snark (cubic graph which cannot be properly colored with the use of three colors).

B. Change of Initial Edge of Coloring

While using algorithm presented in the subsection A the time of computation of edge *k*-coloring of a graph depends on the initial edge of coloring of the graph. We showed this in the work [10] and [11]. The differences in these computational times of the edge *k*-coloring of the graph range from less than one millisecond to several hundred milliseconds. For the change of initial edge of coloring, we use graph isomorphism.

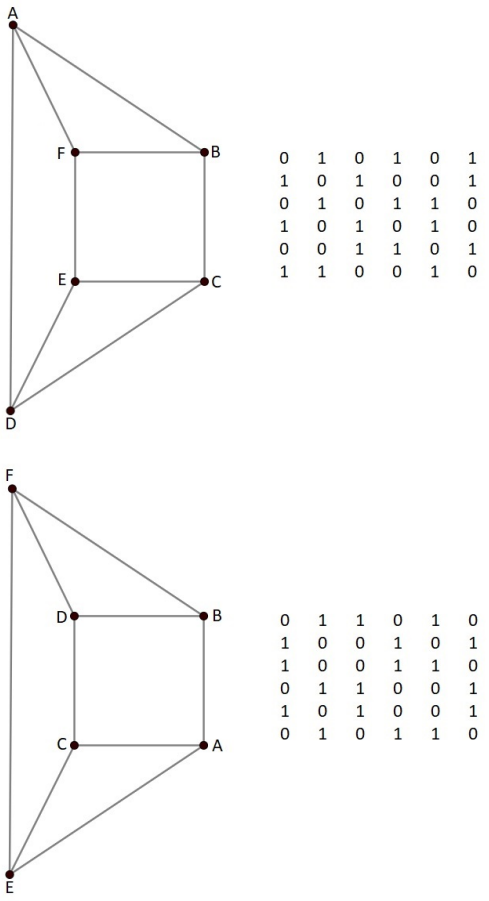


Figure 2. Example of isomorphic cubic graphs and their adjacency matrices

Let the graph *G'* and the given graph *G* be isomorphic graphs. An important feature of isomorphic graphs is that any pair of such graphs has different adjacency matrices but identical diagrams (presented in the Fig. 2 - example of isomorphic cubic graphs for simple visualization). In our case, this means that these are identical graphs with different sequences of vertices and edges. If the graph *G* is represented by the adjacency matrix A, we can create a different sequence of edges of the graph *G* (an isomorphic graph *G'*) by permuting the adjacency matrix of the graph *G*.

This permutation of adjacency matrix *A* of graph *G* can ve computed with the use of following matrix multiplication formula:

$$A' = P^{-1} * A * P \tag{3}$$

where *A'* is permuted adjacency matrix of *G*, *A* is original adjacecny matrix of graph *G*, *P* represents (randomly generated) permutation matrix (this matrix needs to be generated in proper format - containing exactly one value 1 in each row and column of the matrixm otherwise filled with the value 0) and P^{-1} is transposed permutation matrix *P*.

The permuted adjacency matrix *A'* obtained in the described way represents the changed order of the edges of the original graph *G*. Hence, while coloring graph represented by A', algorithm uses different initial edge for coloring of the graph as in the case when the matrix A is used.

IV. Tool for Visualization of Graph Coloring

In this section of the article we describe the design and implementation of the software tool which can be used to analyze edge coloring of graphs. While designing and implementing the tool, we focused on several basic criteria applicable to graph drawing tools:

- Input formats - it is necessary to be able to insert data in various formats of graph notation. The standard input formats in commercially available tools are the adjacency matrix of graph, the incidence matrix of graph, or the adjustment list of graph. However, in addition to these formats, there is number of formats that are practical and space effective for storing graphs in the memory of a computer.
- Drawing algorithm - there is large number of various graph drawing algorithms, which are the basis for graph visualization in the form of a diagram. The main required properties of graph drawing algorithms are the lowest possible computational time needed to draw the graph and clarity of the diagram itself - the clarity of the diagram is implemented as minimization of edge crossing in the diagram and maximization of symmetry of the diagram of graph.
- Specific functions - all available tools which can be used for graph visualization also contain a set of implemented graph functions. Such functions include shortest path search in the graph, shortest cycle search in the graph, Hamiltonian cycle search and so on. Rarely is any form of graph coloring among the implemented functions -

when it is included in the set of functions, only the vertex coloring is implemented.

In the rest of this section, we describe how we proceeded in design and implementation of these three criteria in the case of our proposed software tool for graph coloring.

A. Design of Tool for Visualization of Graph Coloring

Since our tool for graph visualization has a specific objective of edge coloring visualization, we focused on adapting the criteria described above to this goal.

Within the presented tool, we require three input formats for visualization of graph:

- Adjacency matrix and adjacency list - these are conventional methods of representation of graphs. The adjacency matrix is a matrix of size n x n, where n is the number of vertices of given graph. Each row and column of this matrix represents one of the vertices of the given graph - in the case there is an edge connecting two vertices in the graph, the value 1 is recorded in the adjacency matrix, the value 0 is recorded otherwise. The adjacency list is a list of vertices with assigned names of the vertices to which they are connected by an edge.
- Graph6 format - with the use of this format, we are able to store graphs in a compact manner, using vectors of printable ASCII characters. A graph in graph6 format is stored in a single line of a file, which allows us to efficiently store entire sets of graphs in a single file. Although this graph format is unreadable to humans (as opposed to the adjacency matrix and mainly adjacency list), its advantages are versatility and efficiency of use.

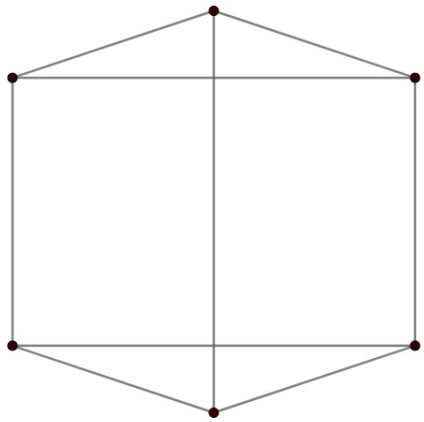


Figure 3. Example of use of circular layout drawing algorithm applied on the graph from Fig.2

After reading input graph in one of the selected formats, the application draws the graph using the selected graph drawing algorithm. For the purposes of this work, we have chosen circular layout - a popular method of graph drawing, which offers high clarity of diagram of graph (Fig. 3). Regarding number of edge crossings in the graph - this algorithm is able to draw a graph with no edge crossings only in the case of outerplanar graphs. In other cases, edge crossing occurs, but it is possible to implement optimization techniques which minimize number of crossings.

In this way we implement clarity of the diagram of graph based on minimization of edge crossing in the diagram and maximization of symmetry of the diagram of graph.

The third - and arguably the most important - criterion of the design of the graph visualization tool is the specification of a set of functions which can be used while working with the tool:

- Translation between formats of graph description - the examined graph is inputted in one of three possible formats (adjacency matrix, adjacency list or graph6 format). Due to the readability of the graph and in order to increase other possibilities of working with the graph, it is practical to translate the examined graph from the input format into the other two defined formats.
- Edge coloring of graph - the core of the presented application is edge coloring of graphs. The Edge Backtracking Algorithm presented in section III, subsection A is used for implementation of this action.
- Visualization of edge coloring of graph - an extension of the standard edge coloring of graph is the element of visualization of the computed coloring. For the edge colored graph, a graph diagram is drawn and the edges of this diagram are colored according to the computed coloring. In such case, that the graph is not proper edge k-colorable, the edges of the graph remain colored with the use of default color. The user is notified that there is no proper edge k-coloring for the given graph.
- Step-by-step edge coloring - from the point of view of analysis of edge coloring, it is necessary for the user to be able to browse through a specific computing coloring in the step-by-step manner. Each step represents coloring or recoloring of one of the edges of the graph visualized in the application. Since edge coloring can take several tens to several thousand of steps, it is advisable to implement the stepping options set to 1 step, 5 steps, 10 steps and 50 steps at a time. It is also necessary to be able to step edge colorgin in both directions (forward and backward) with the possibility of 1, 5, 10 or 50 steps at a time.

B. Implementation of Tool for Visualization of Graph Coloring

In the implementation of the required solution, we followed the design presented in the section IV, subsection A. The application was implemented using C/C++ language as follows:

- The input graph for the program is specified in a text file, which is located in the directory with executable file of the program. This graph can be stored in adjacency matrix, adjacency list or graph6 format.
- After starting the program, the input file is loaded and the format in which the graph is stored is specified. In the case of the adjacency matrix, the program proceeds to the next step of computation. In other cases, the input format is translated to adjacency matrix and the computation continues. In this step, two new files are created - each

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Table I
COMPARISON OF FUNCTIONALITIES OF OPEN SOURCE, PAID AND PROPOSED TOOLS

	Graph Online	CS Academy	Matlab	Wolfram Mathematica	Proposed Tool
Input: adjacency matrix	YES	YES	YES	YES	YES
Input: graph6 format	NO	NO	NO	YES	YES
Edge coloring of graph	NO	NO	YES	YES	YES
Step analysis of edge coloring	NO	NO	NO	NO	YES

containing the graph stored in one of the two remaining formats.

- Since the graph is represented as an adjacency matrix, it can be edge colored using the $edgeColorise$ function. The function uses Edge Backtracking Algorithm and edge 3-colors the graph. The output of this function is information on whether the graph can be colored with the use of three colors or not and a set of steps for edge coloring of the graph. These steps are later visualized.
- The graph is then visualized using the $drawVertices$ and $drawEdges$ functions in the graphical user interface run by the $drawGUI$ function. This GUI contains space for plotting graphs and buttons for stepping options which consist of eight buttons: +1, +5, +10, +50 and -1, -5, -10, -50 steps.
- Finally, the $drawSteps$ function is called, which uses step data fo edge coloring computed in the $edgeColorise$ function and, after the user interacts with the GUI buttons, draws the individual coloring steps.

The application uses two windows for its operation. The first window serves as a console in which the user is provided with information about the input graph - number of vertices, adjacency matrix, adjacency list and graph6 format of the graph, computed information about edge 3-colorability of the graph, number of steps needed in edge coloring of graph and current step of coloring visualized in the other window of application.

The second window serves as a GUI - the visualization of the graph itself takes place in it and it provides buttons for step-by-step edge coloring.

V. EXPERIMENTS ON GRAPH COLORING VISUALIZATION

We test the proposed application from the two perspectives. The first perspective is to compare the basic functionalities of the presented tool and a group of other - open source or paid - tools that are designed for a similar purpose. The properties, we focuse on include input data formats for tools, functions implemented in individual tools, the possibility of analyzing graphs in the tool or the duration of edge coloring of graphs in the application.

The second point of view of application testing is the use of the presented application in the identification of several patterns and subgraphs that cause an increase in the time of computation of edge coloring of graphs.

A. Comparison of Proposed Graph Visualization Tool with Other Available Tools

There are several tools available that are commonly used to draw graphs - whether open source tools like $GraphOnline$ or $CSacademy$, or professional paid tools like $Matlab$ or $WolframMathematica$. All of these tools share common features with small variations.

In the Tab.I we present the properties available in the individual tools and compare them with each other and at the same time with the application proposed in this paper.

From the Tab.I it is clear, that we can divide the properties essential for our research into following three groups:

- Input data formats for tools - each of the tools is able to work with an adjacency matrix as input, but only a minority of tools can work with input data in the graph6 format. This format is practical and space effective way of storing a graph in the memory of computer - it encodes an adjacency matrix of graph into a one-line character string. This format is supported by $WolframMathematica$ and proposed application.
- Functions implemented in the tool - each of the presented tools contains several implemented functions, that can be applied on graphs (searching for the shortest path in the graph, searching for the shortest cycle of graph, and so on). The function that interests us is the edge coloring of graph. This function is present only in the tools $Matlab$, $WolframMathematica$ and the presented application. Although it is not possible to edge color a graph in the mentioned open source tools, a function for the vertex coloring of the graph is present in the $GraphOnline$ tool. Thus, we would be able to project the input graph into the line graph [13] and thus obtain the edge colorability of the graph. However, such a method of coloring is not visual and the transformation of the graph itself is not available in the given tool.
- Specialized analytical functions - none of the available tools offers the possibility of visualization of the edge coloring of graph step-by-step - therefore it is not possible to analyze the coloring in any way. This feature is unique among graphing tools.

It is important to mention that the possibility of step-by-step visualization of edge coloring of the graphs is advantageous from the point of view of the analysis of the coloring itself, but it brings increased costs in the form of substantial increase of computational time needed for the problem. In the Tab.II we present a comparison of computation times for coloring of the followings: Petersen graph - simplest example of snark (10

vertices), the first Blanuša snark (18 vertices) and 34-vertex snark generated by the $snarkhunter$ tool [15]. All of these graphs are cubic, so the number of edges which are colored, can be computed as $(3/2) * n$, where n is number of vertices of the graph.

It is clear that the presented tool needs much higher time to color the graph itself. This is due to the need to compute and store all the edge coloring steps of the input graph so that the graph can be subsequently analyzed. We do not consider this shortcoming to be important at the moment, as this tool is used to analyze graphs and not as a tool which computes the edge colorability of a graph in the shortest possible time (we presented such an algorithm in [10]).

Table II
COMPARISON OF COMPUTATION TIME OF EDGE COLORING FOR CHOSEN SNARKS

	Wolfram Mathematica	Proposed Tool
Petersen graph	550 ms	670 ms
Blanuša snark	540 ms	840 ms
34-vertex snark	1080 ms	12 440 ms

B. Patterns and Subgraphs of Interest

In this subsection, we present point of view of application testing focused on the use of the presented application in the identification of patterns and subgraphs that cause an increase in the time of computation of edge coloring of graphs.

In the Tab. III we present the number of necessary recolorings of the edges of individual graphs needed for the algorithm to be able to determine whether the graph is edge 3-colorable or not. These values are measured on graphs in standard form - in the form without applying any permutation on the graph - while using Edge Backtracking Algorithm presented in the section III, subsection A.

Table III
COMPARISON OF NUMBER OF RECOLORINGS FOR CHOSEN GRAPHS IN STANDARD FORM

	Number of Recolorings
Petersen graph	91
Blanuša snark	1029
34-vertex snark	18 253

The graph with the lowest number of vertices we worked with was 10 vertex (15 edge) snark - Petersen graph. Edge 3-colorability (or more precisely uncolorability) of this graph was computed in 91 steps. On the Fig. 4 we see that the edge coloring of the graph was started on the edge marked with vertices $(Initial_S, Initial_E)$ and the graph was being properly edge colored until the coloring of edge represented by dashed line. Uncolored edges are marked by a dotted line.

The coloring of this graph was done in 91 steps - out of these 91 step 80 were dedicated to the recoloring of graph related to the problematic edge marked with the use of dashed line. This represents 87.9 percent of computation steps needed for the edge 3-coloring of the graph.

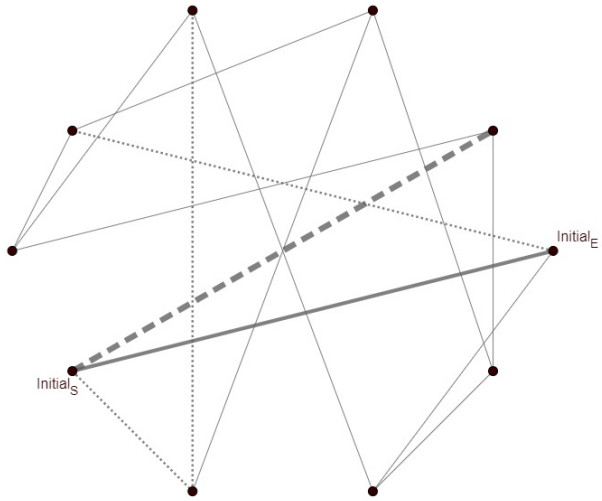


Figure 4. Analysis of edge coloring of Petersen graph

From the Fig.4, we can identify some properties and information about the problematic edge and edge coloring of this graph:

- problematic edge is incident with initial edge of coloring,
- three of four edges which were not colored are incident to the initial edge of coloring,
- fourth edge, which was not colored is incident to other uncolored edge.

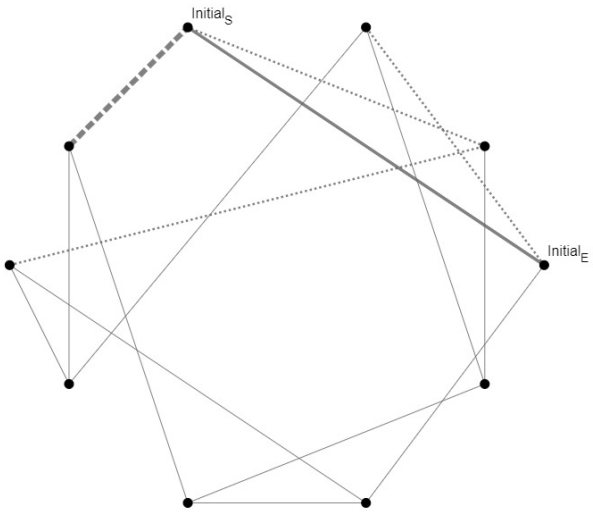


Figure 5. Analysis of edge coloring of permuted Petersen graph

To compare this edge coloring of Petersen graph, we used edge 3-coloring of the permutation of the same graph as presented in the section III, subsection B (see Fig.5). The permutation used for the coloring of the graph was randomly generated, while complying the criteria for permutation matrix presented in the section III, subsection B. The edge coloring of graph presented in the Fig.5 was computed in 73 steps. After analysis of this edge coloring we can identify same properties and information about the problematic edge and edge coloring

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

of this graph as in previous case. The only difference is number of steps needed in order to find the edge coloring of the graph (see Tab.IV).

Table IV
COMPARISON OF NUMBER OF RECOLORINGS AND TIME OF COMPUTATION OF EDGE COLORING FOR PETERSEN GRAPH IN STANDARD FORM AND PERMUTED FORM

	10-vertex, 15-edge snark
Number of Recolorings in SF	91
Number of Recolorings in PF	73
Edge Coloring Time in SF	670 ms
Edge Coloring Time in PF	660 ms

As we can see from the Tab.IV, the number of steps needed for computation was lower while using permuted graph, but the time of computation of the edge coloring of graph was almost the same. The values of computational time presented in the Tab.IV were computed as an average from five measurements.

In order to further test the application we used the first Blanuša snark in the same way as the Petersen graph above. The first Blanuša snark consists of 18 vertices and 27 edges. The edge coloring of the graph was computed in 1029 steps. Out of these 1029 steps, steps 1-64 colored edges properly with minor recolorings needed. When algorithm found the problematic edge, it took remaining 965 recolorings to try and color this edge (which corresponds to 93.8 percent).

Before recoloring, only four edges are not colored properly - in this case, problematic edge is not incident to the initial edge of coloring (as oppose to the first examined case), three of these four edges are incident to the initial edge of coloring and fourth edge, which was not colored (the problematic one) is incident to other uncolored edge. We present this coloring on the Fig.6.

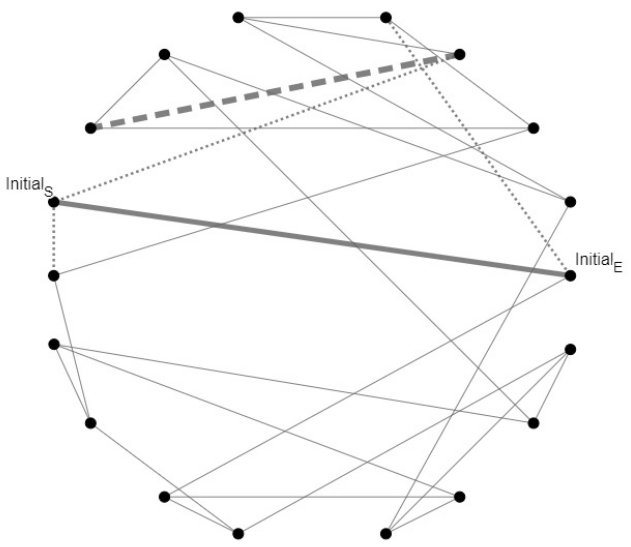


Figure 6. Analysis of edge coloring of first Blanuša snark

We also edge colored permuted first Blanuša snark. The edge coloring of the permuted graph was computed in 401 steps and contained no single problematic edge. While coloring this graph, algorithm encountered number of smaller problems in coloring which were recolored in 14 - 60 steps. Therefore, we need to study permutation used in this case and try to generalize its properties in order to further optimize the edge 3-coloring of graphs.

In the Tab.V we present comparison of computation times for edge coloring of the first Blanuša snark and comparison of number of step needed for edge coloring of given graph.

Table V
COMPARISON OF NUMBER OF RECOLORINGS AND TIME OF COMPUTATION OF EDGE COLORING FOR THE FIRST BLANUŠA SNARK IN STANDARD FORM AND PERMUTED FORM

	18-vertex, 27-edge snark
Number of Recolorings in SF	1029
Number of Recolorings in PF	401
Edge Coloring Time in SF	840 ms
Edge Coloring Time in PF	710 ms

VI. CONCLUSION

Main objective of the article was to present a tool created for the need of analysis of edge 3-coloring of graphs based on visualization of edge coloring and searching for subgraphs and patterns, which prolong computation of the chosen problem.

We designed and implemented software tool, which can be used for step-by-step analysis of edge coloring of graphs. On the basis of this step analysis, we can decide which edge of the graph is most fitting as an initial edge of coloring of given graph.

In the section V, subsection B, we presented use of step-by-step analysis on chosen types of graphs - Petersen graph and the first Blanuša snark. With the use of our tool, we were able to identify some properties and information about the problematic edges and edge coloring of these graphs.

Even though the function of step analysis of edge coloring of graph is unique and practical, it also has some drawbacks. Main shortcoming of proposed tool is time needed for computation of all steps of coloring - in some cases this visualization is computed in the time 10-times higher than standard edge coloring of graph.

Future work in this area contains:

- Use of parallel and distributed computing for optimization of computational time needed for visualization of edge coloring of graph similar to [10], [16]. The edge coloring can be computed in advance and the visualization of individual steps of the coloring can be done in parallel via either simple thread-based model or via use of distributed computing.
- Implementation of methods of artificial intelligence, which would be able to analyze graph coloring and propose fitting permutation of given graph. We call permutation fitting in such case, that after applying it on the

graph (via formula 3), the computational time for edge coloring of the graph is reduced.
- Using proposed model for analysis of large sets of edge colored graphs.
- Implementation of user interface for the general use [17] - mainly formats of data input for application are not user friendly at the moment.

ACKNOWLEDGMENT

The research was partially supported by the grant of The Ministry of Education, Science, Research and Sport of Slovak Republic - Implementation of new trends in computer science to teaching of algorithmic thinking and programming in Informatics for secondary education, project number KEGA 018UMB-4/2020.

Computing was performed in the High Performance Computing Center of the Matej Bel University in Banská Bystrica using the HPC infrastructure acquired in project ITMS 26230120002 and 26210120002 (Slovak infrastructure for high-performance computing) supported by the Research & Development Operational Programme funded by the ERDF.

REFERENCES

[1] Marx D.: Graph colouring problems and their applications in scheduling. In: Periodica Polytechnica, Electrical Engineering, Vol. 48, 2004, No. 1-2, pp. 11-16.

[2] Chaitin G. J.: Register allocation & spilling via graph colouring. Proc.1982 SIGPLAN Symposium on Compiler Construction, pp. 98-105, 1982. ISBN 0-89791074-5

[3] Holyer I.: The NP-Completeness of Edge-Colouring. In: SIAM J.COMPUT, Vol. 10, 1981, No. 4, pp. 718–720. ISSN 0097-5397.

[4] Karabáš J.—Máčajová E.—Nedela R.: 6-decomposition of snarks. In: European Journal of Combinatorics: 20th International workshop on combinatorial algorithms (IWOCA), Elsevier, Vol. 34, 2013, No. 1, pp. 111–122. ISSN 0195-6698.

[5] Kowalik L.: Improved edge-coloring with three colors. In: Theoretical computer science, Vol. 410, 2009, No. 38-40, pp. 3733-3742. ISSN 0304-3975.

[6] Beigel R., Eppstein D.: 3-coloring in time O(1.3289^n). In: J. Algorithms 54(2), 168–204, 2005

[7] Fiol M. A.–Vilaltella J.: A Simple and Fast Heuristic Algorithm for Edge-coloring of Graphs. In: AKCE International Journal of Graphs and Combinatorics, Vol. 10, 2013, No. 3, pp. 263-272

[8] Nedela R., Karabáš J., Škoviera M.: Nullstellensatz and Recognition of Snarks. In: 52th Czech-Slovak Conference Grafy, 2017

[9] Dudáš A., Škrinárová J., Voštinár P., Siláči J.: Improved process of running tasks in the high performance computing. In: ICETA 2018: Proceedings: 16th IEEE International Conference on Emerging eLearning Technologies and Applications. pp 133-140. ISBN 978-1-5386-7912-8.

[10] Dudáš A., Škrinárová J, Vesel E.: Optimization design for parallel coloring of a set of graphs in the High-Performance Computing. In: Proceedings of 2019 IEEE 15th International Scientific Conference on Informatics. pp 93-99. ISBN 978-1-7281-3178-8.

[11] Dudáš A., Janky J., Škrinárová J.: Web applicaiton for graph visualization purposes. In: ICETA 2020: Proceedings: 18th IEEE International Conference on Emerging eLearning Technologies and Applications. pp 90-96. ISBN 978-0-7381-2366-0.

[12] Palúch S.— Peško Š.: Quantitative methods in logistics ($in\ Slovak$). EDIS, Žilina, Slovakia, 2006, ISBN 80-8070-636-0.

[13] Diestel R.: Graph theory. Springer - Verlag, Heidelberg, 2016, ISBN 978-3-662-53621-6.

[14] Häglund J.: On snarks that are far from being 3-edge-colorable. In: The Elcectronic Journal of Combinatorics, Vol. 23, 2016, No. 2, Paper P2.6. ISSN: 1077-8926.

[15] Brinkmann G.—Coolsaet K.— Goedgebeur J.—Mélot H.: House of Graphs: a database of interesting graphs, Discrete Applied Mathematics, 161(1-2):311-314, 2013 (DOI). Available at http://hog.grinvin.org

[16] Melicherčík M., Siladi V., Svítek M., Huraj L.: Spreading High Performance Computing Skills with E-Learning Support, ICETA 2018 - 16th IEEE International Conference on Emerging eLearning Technologies and Applications, 2018, pp. 361–366.

[17] Sedlacek P., Kmec M., Rusnak P.: Software Visualization Application for Threads Synchronization Handling in Operating Systems. ICETA 2020 - 18th IEEE International Conference on Emerging eLearning Technologies and Applications, 2020, pp. 580–585

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Software reliability model based on syntax tree

Peter Sedlacek
Faculty of Management Science and Informatics,
University of Zilina, Slovakia
Email: Peter.Sedlacek@fri.uniza.sk

Elena Zaitseva
Faculty of Management Science and Informatics,
University of Zilina, Slovakia
Email: Elena.Zaitseva@fri.uniza.sk

Jan Rabcan
Faculty of Management Science and Informatics,
University of Zilina, Slovakia
Email: Jan.Rabcan@fri.uniza.sk

***Abstract*—In this paper new model for software reliability is suggested. There are several software reliability models, however most of them cannot be used for analysis of system components, such as importance measures calculation. Therefore new model was proposed. The creation of this model is based on the source code, that is in the next step used to generate a syntax tree. Syntax tree is a hierarchical representation of the given source code, that is independent on programming language. This tree is then used to create reliability model, in this case fault tree and transformed into structure function. The process of creating reliability model from source code is demonstrated in this paper. The created reliability model can be then used to calculate system characteristics, such as importance measures. The main advantage of suggested model is the possibility to use traditional methods for system evaluation such as logic differential calculus. However this model has also several disadvantages, for example its large dimension.**

***Index Terms*—Software reliability, Syntax tree, Source code, Reliability analysis, Fault tree, Structure function**

ACKNOWLEDGMENT

I. INTRODUCTION

Reliability is important characteristic of any system in these days and software systems are no exception. Therefore the development of methods for the quantification of reliability for software is relevant problem in reliability engineering [23], [27]. The conception of software reliability according to [12] is defined as the probability of a software operates failure-free in a specified environment for a specified exposure period. But probabilistic performance of software cannot be defined as the hardware performance due to the uncertainty in operation. According to this conception a software failure is discrepancy between intended and actual output, and software reliability engineering provides quantification of two things, expected use and desired major quality characteristics. A failure of software can occur at different stages of development and exploitation. But most often analysed stage is software development [12]. One of the possible methods for analyzing software reliability at the development stage is to consider a software testing process [32], [31]. During the testing defects and failures of software are detected, limited and removed. It allows growing of software reliability. The estimation of software reliability based on information from the software testing process, as a rule, is implemented by methods named as Software Reliability Growth Models (SRGMs) [32], [25].

The SRGMs can be considered as stochastic counting processes regarding the number of faults detected or failures experienced in software testing . Based on this data some of software reliability measures are computed as a probability that no failure occurs during a certain time interval. One of the first SRGM is known as Jelinski-Moranda model [13] which assumes that the elapsed time between failures is governed by the exponential probability distribution. There are some developments of this model with application of Markov based methods [32], [20], [31]. The main idea of these models is that software failure is random event (caused by errors in software) and this can be modelled using some well-known probability distribution. For example in Markov model, mainly exponential distribution is used. These models can be easily used to calculate software reliability and, for example predicts software failure using historical information about analysed software. The development of Jelinski-Moranda model named as Non-Homogeneous Poisson Processes (NHPPs) is often used in software reliability too [32], [25], [31]. All these methods of software reliability according to [10], [4] are time-dependent, but not time-data-dependent. In addition the structure or topology of the software are not taken into account in the analysis based on SRGM [24], [5]. There are several models suitable also to calculate these characteristics of software. One of them was proposed in paper [35] and can be used in case software is implemented with micro-service architecture. In this case dependence graph of individual services is created and this graph is used to create fault tree. Reliability of each service is calculated using historical data about its functioning and failures. The next model was proposed in paper [28] and is based on UML/DAM diagrams. In this case these diagrams are created as a representation of software, for example use case diagram, deployment diagram and activity diagrams. Using all of the mentioned diagrams fault tree is created. The software time-data-dependent analysis based on fault tree are considered in [5]. Risk analysis method for software evaluation is proposed in[11].

According to studies in[10], [11], [5] topological properties of software can be take into account mainly based on data-depend and structure-depend model of software. This approach developed in this paper. The main idea of our contribution is to use source code and syntax tree to create software reliability model. Syntax tree, or Abstract Syntax Tree (AST) is an abstract graph representation of source code. This representation has many usages in different fields of computer science, for example plagiarism detection [8], [9], code evolution [22] and

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Fig. 1. Example of software failure

summarization [7], etc. [3]. Using this representation of source code instead of using source code itself has several advantages. One of them is, for example, language independence. As AST is an abstract representation, using this form to create reliability model allow us to analyse software implemented in any programming language (as long as we are able to create AST from it). In our paper we are describing method to use source code represented in form of AST to create reliability model. Model constructed using these information will allow us to analyse system characteristics together with analysis of system components, in our case functions or blocks of code and their importance for system reliability. The advantage of this method is the possibility to use traditional methods for reliability analysis, such as logic differential calculus also in software reliability.

II. Theoretical Background

Many concepts of software reliability engineering were adapted from the techniques of hardware reliability, because these methods were efficiencies in analysis of hardware based system. The application of such methods to software has to take into account important differences in the nature of hardware and software. Therefore, there are many special methods for software reliability analysis, which developed for such system.

From reliability point of view software systems have some differences compared to other types of systems. Therefore it is necessary to specify some basic terms specific for software reliability. The first one is software failure.

Software failure is a state in which software is not performing expected functions [29]. Please note that it is not only state when software crash but also state when the output of executed action is different to the expected one. Example of such failure can be seen in Fig. 1.

Software failures are evaluated by difference indices, for example, as failure rate. *Failure rate* is number of failures per time unit [29]. The failure rate similar to other indices is computed based on results of software testing with application of SRGM. Therefore this evaluation is possible after the software development and testing. Software fault prediction is an important concept that can be applied at an early stage the software life cycle [25]. Effective prediction of faults may improve the reliability and testability of software systems. This analysis should be developed based on evaluation software structure.

A. Reliability model creation

The main idea of the proposed method is to use source code as a base for reliability model creation. The principle of this method can be seen in Fig. 2. This method consists of 2 essential steps. In the first step, source code is used to construct abstract syntax tree.

Abstract Syntax Tree (AST) is an abstract graph representation of the source code. AST consists of nodes, where each of them represents one element of the source code. Elements in the same block of code are placed on the same level of the syntax tree. In case element of the code can be divided into parts, these parts will be placed as a nodes in the next level and all of them will be connected to element this elements consists of. AST can be easily used to present a structure of the source code and to remove unnecessary information from it, such as gaps, variable names, etc. [9] that will be unnecessary for the purposes of reliability analysis.

The next step of this method is to construct a reliability model using created AST. For this purpose we decide to use fault tree.

Fault tree is an acyclic graph that consists of 2 types of nodes: events and gates. An event is an occurrence within the system, that typically describes a failure or a degradation of subsystem or a degradation of component of the system. These events are connected via logic gates, typically AND or OR gates. These events together describes situation, when system fails [14], [15]. Fault tree can be easily transformed into structure function, that allow us to represent system as a Boolean function and use methods for its analysis such as logic differential calculus to perform reliability analysis [33].

Structure function $\phi(\boldsymbol{x})$ express dependencies between system state and state of its components [16]. In case system consists of n components and each component can be in two possible states: 0 (failure) and 1 (working), structure function can be expressed in the following form [16], [34]:

$$\phi(x_1, x_2, ..., x_n) = \phi(\boldsymbol{x}) : \{0,1\}^n \rightarrow \{0,1\} \tag{1}$$

In this form $\boldsymbol{x}$ represents a components state vector, where i-th element of the given vector (element x_i) represents a state of a component i of the system.

Structure function allow us to perform topological analysis, but is not sufficient to perform also probabilistic analysis. For this purposes we need also information about probabilities individual system components will be in specific state. For the i-th system component, probabilities can be defined as following [16], [34]:

$$\begin{aligned} p_i &= \Pr\{x_i = 1\}, \\ q_i &= \Pr\{x_i = 0\} = 1 - p_i, \end{aligned} \tag{2}$$

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

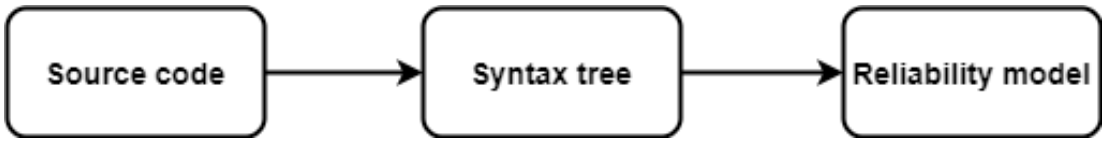


Fig. 2. Principle of proposed method

where p_i is the probability system component x_i is working and q_i is the probability system component fails.

Structure function can be represented in different ways, such as reliability block diagram, decision diagram, truth table, etc. In this paper we will use a truth table. This table contains information about every possible combination of components states and for this combination corresponding value of the structure function. In case every system component has exactly two possible states, the table consists of 2^n rows.

B. Reliability analysis

Structure function created using proposed method can be used to perform reliability analysis and to calculate different characteristics of the system and its components. One of the most important characteristics are reliability and unreliability.

Software reliability can be defined as probability software performs expected functions (according to specification) without failures for given time period or in specific time point [29]. Structure function together with information about probabilities of system components can be used to calculate reliability R and unreliability U functions as following [16], [26]:

$$\begin{aligned} R &= \Pr\{\phi(\boldsymbol{x}) = 1\}, \\ U &= \Pr\{\phi(\boldsymbol{x}) = 0\} = 1 - R \end{aligned} \tag{3}$$

Structure function can be seen as a Boolean function. Therefore it is possible to use methods for analysis of Boolean functions also in reliability analysis. One of them is logic differential calculus [18]. The most useful tool for reliability analysis is Direct Partial Boolean Derivative (DPBD). This can be defined as following [33], [30]:

$$\frac{\partial\phi(1 \to 0)}{\partial x_i(1 \to 0)} = \phi(x_1, ..., 1, ...x_n) \wedge \overline{\phi(x_1, ..., 0, ...x_n)} \tag{4}$$

DPBD allow us to find cases, where the change of function $\phi(\boldsymbol{x})$ from state 1 to state 0 is caused by change of variable x_i from state 1 to state 0. In reliability analysis this allow us to find cases, where failure of analysed system is caused by failure of system component x_i [34]. These cases are important to analyse influence of individual components to system operation state. This is performed using importance measures.

Importance measures are characteristics of system components that allow us to quantify importance of individual component on the system state. They can be divided into two types: structure and probabilistic. Structure importance measures are focused to analyse system from topological point of view and probabilistic takes into account also probability system component states. In this section we will present two of them. The first one is Structure Importance (SI). This measure is defined as a relative number of cases, in which system failure is caused by failure of given component. Using DPBD these cases can be identified and SI can be calculated using following formula [34]:

$$\mathrm{SI}_i = \mathrm{TD}\left(\frac{\partial\phi(1 \to 0)}{\partial x_i(1 \to 0)}\right) \tag{5}$$

Function TD is Truth Density and express relative number of cases in which function takes value 1 to all possible cases [19].

Structure importance can be used to analyse importance of system component from topological point of view. However this measure ignores probabilities of component states. In case we have this information, we can calculate also other importance measure, Birnbaum's Importance (BI). Using DPBD, Birnbaum's importance can be calculated as following [34]:

$$\mathrm{BI}_i = \Pr\left\{\frac{\partial\phi(1 \to 0)}{\partial x_i(1 \to 0)} = 1\right\} \tag{6}$$

This measure express the probability system failure is caused by failure of given component [16], [34].

Importance measures allow us to detect the most critical parts of our system. The higher the value of importance measure is, the more is the component important for system functioning. This help us in case of system failure to find parts that is the most likely causing this failure. In period of design and implementation of the system this can help us to put more emphasis on these critical parts of the system.

III. Example

In this section we will demonstrate whole process of reliability analysis of software system from source code, through syntax tree and reliability model creation to calculation of software characteristics using typical method of reliability analysis - differential calculus.

Let us take the source code from Fig. 3. It is a simple program written in C# language in which 2 local variables are declared. Then there is a for loop statement in which it is iterated between these two local variables. In case current value is divisible by 3, value "1" is written into standard output. Otherwise value "0" is written.

A. Syntax tree creation

This source code can be interpreted in form of syntax tree in the following way. The main block of code consists of three elements, as can be seen in Fig. 4. There are two

```
static void Main(string[] args)
{
    int x = 10;
    int y = 20;
    for (int i = x; i < y; i++)
    {
        if (i % 3 == 0)
            Console.WriteLine("1");
        else
            Console.WriteLine("0");
    }
}
```

Fig. 3. Example of source code

local declaration statements and one for loop statement. In the syntax tree, local declaration can be taken as is and nodes for them are created. The for loop statement can be analysed deeper, as can be seen in Fig. 5. For loop consists of the header and the body. The header contains declaration, iteration step and condition to stop the loop. Fig. 6 describes the body of the for loop. It contains only one element - if statement. And finally, if statement consists of 3 elements: condition, true block and false block. This can be seen in Fig. 7.

All of the mentioned elements are taken into the final syntax tree as can be seen in Fig. 8. The elements of the source code from one block are on the same level of the syntax tree. Please note, that it is possible to continue and split some nodes into parts. For example there is a node for the header element in the for loop statement. In our case, we take it only as one node - variable declaration, but actually this consists of three elements: variable declaration, iteration step and condition to stop. Similarly this split can be applied also to the other elements.

In practise there are several tools, that constructs syntax tree from the source code, for example .NET compiler platform Roslyn for C# language [2] or AST parser for Java language [1]. But a syntax tree created using these tools has too large dimension and is not suitable for the creation of the reliability model unless unnecessary nodes are removed or merged together. Example of the syntax tree generated using Roslyn can be seen in Fig. 9. This is representation of the local declaration statement. As we can see, there are 16 nodes, although many of them are not important for our purposes. Therefore this tree has to be processed before using to create a reliability model.

Either way we process our source code into syntax tree, i.e. either we split blocks of code into nodes or we merge unnecessary nodes, the main decision that has to be taken is what depth should be taken in order to create reliability model from syntax tree. There are several options. We can for example use fixed depth and process syntax tree only on the first level. In this case our resulting syntax tree will consist only of 4 nodes - block, 2 local declaration statement and for loop. Other option, we also have used in our work is not to use fixed depth, but to use specific depth for specific elements, i.e. depending on the type of the element, we decide to split or not. Actual decisions for each element can be seen in Table I. In this table, value ”no split” means, we stop on the given element and don’t continue in current branch. This is used for the true and the else clause of the if statement as they contains only one element of code. In case there will be more than one element, we would decide to continue splitting. Value ”split” means we continue in actual branch. In our case we are using it in the block element. That means we take the block as is and all containing elements will be placed in the next level. For other elements of our example, table contains information what element will it be divided into. Please note that this table does not cover all elements of source code, but only the ones used in example. There can be also other elements like other types of loops, function calls, more complex conditions etc.

TABLE I
INDIVIDUAL SOURCE CODE ELEMENTS AND HOW TO SPLIT THEM

Source code element	Split
block	split
declaration statement	no split
for loop	declaration statement & block
if	equal expression & true clause & else clause
equal expression	no split
true clause	no split
else clause	no split

B. *The creation of the reliability model from syntax tree*

Created syntax tree will be in the next step used to create a reliability model. We have decided to use fault tree. Each node of syntax tree will match with one event in fault tree. This event describes situation where the given elements fails. For example a failure of the local declaration statement means there is an unexpected value assigned to this variable or the local declaration statement is defined with unexpected type as it was expected etc. A failure of the for loop means, at least one of its child elements fails, e.g. wrong iteration step will be set or it will iterate between incorrect boundaries etc.

Another decision that has to be taken is how to connect individual events together. The main problem is that failure of these events is not independent. In our example failure of the first local declaration statement will lead to failure of the whole for loop as the for loop iterates between these two local variables. And furthermore a variable declared in the for loop statement is then used in the if statement. So the mentioned failure of the first local declaration statement can lead to the failure of the whole software. Therefore we decide to connect these events using AND gates. This will ensure that this model will not work incorrectly. But using AND gates between all events can lead to an inaccuracy of this model as not every failure lead to the failure of the whole software. Let us consider situation, where a value assigned to the local declared variable ”x” will be 1 instead of 10 and a value assigned to the variable ”y” will be 11 instead of 20. The number of iterations in the for loop statement will be the same (10 iterations) and even the output of the software will match with the expected one. In this

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

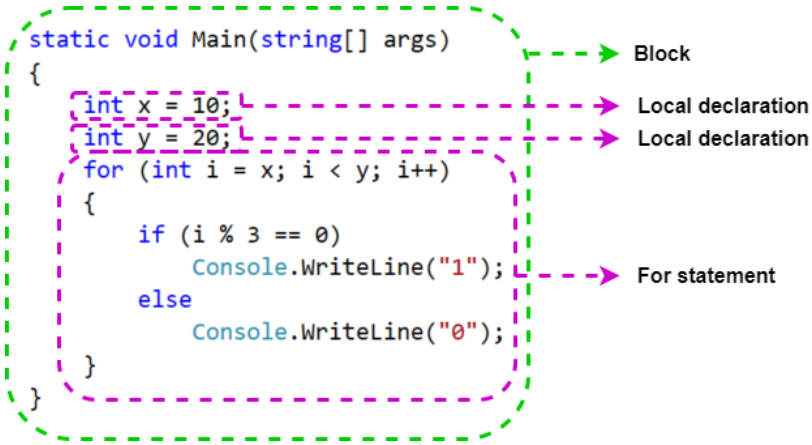


Fig. 4. Top level of source code description - main function

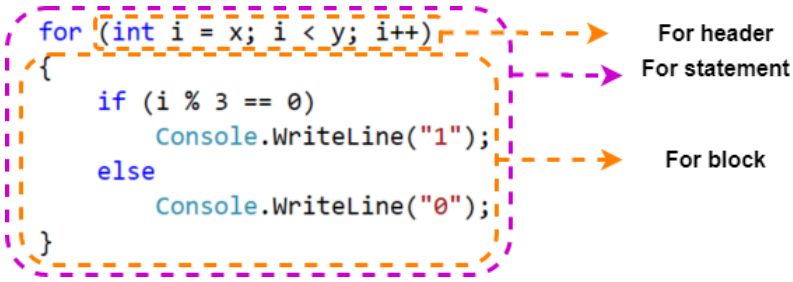


Fig. 5. Second level of source code description - for loop

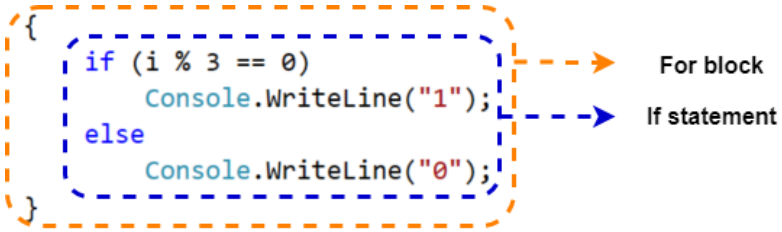


Fig. 6. Third level of source code description - body of for loop

TABLE II
PROBABILITIES OF STATES FOR EACH NODE IN FIG. 10

Node name	Component label	p_i	q_i
Local declaration statement	x_1	0.98	0.02
Local declaration statement	x_2	0.98	0.02
Variable declaration	x_3	0.90	0.10
Equal expression	x_4	0.95	0.05
True clause	x_5	0.96	0.04
Else clause	x_6	0.96	0.04

case a mutual failure of both local declaration statements lead to the correct behavior of the software. Mentioned inaccuracy is in this model neglected.

The resulting reliability model created using syntax tree from Fig. 8 can be seen in Fig. 10. Structure of this model corresponds to the syntax tree this model was constructed from and individual events are connected via AND logic gates together.

Created fault tree can be easily transformed into structure function in the form of the truth table. This allow us to use the logic differential calculus and using this tool perform topological analysis and calculate also other characteristics of the system such as system reliability. Calculation of some of these characteristics is possible only with information about probabilities of individual system component states. These probabilities can be seen in Table II. Probabilities listed in the mentioned table are only example values. Real values can be obtained using statistical information or estimations of experts. Please note, that only leaf nodes of the fault tree is required as probabilities of non-leaf nodes depend only on these leaf nodes. This table also contains information about mapping node names into variable names. This make work with these nodes simpler. Structure function in form of the truth table can be seen in Table III.

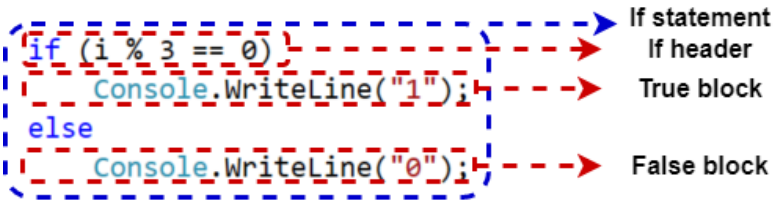


Fig. 7. The final level of source code description - if statement

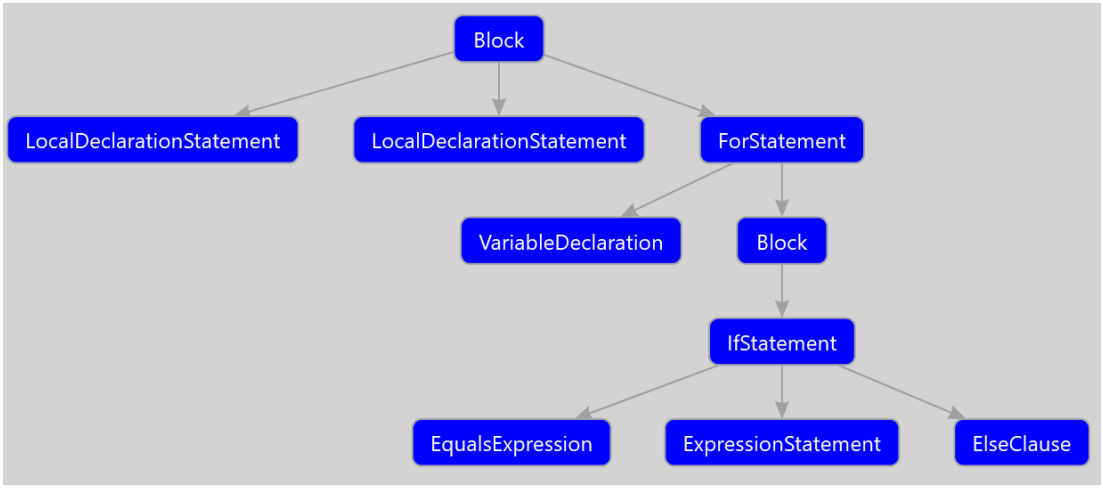


Fig. 8. Resulting syntax tree corresponding to source code in Fig. 3

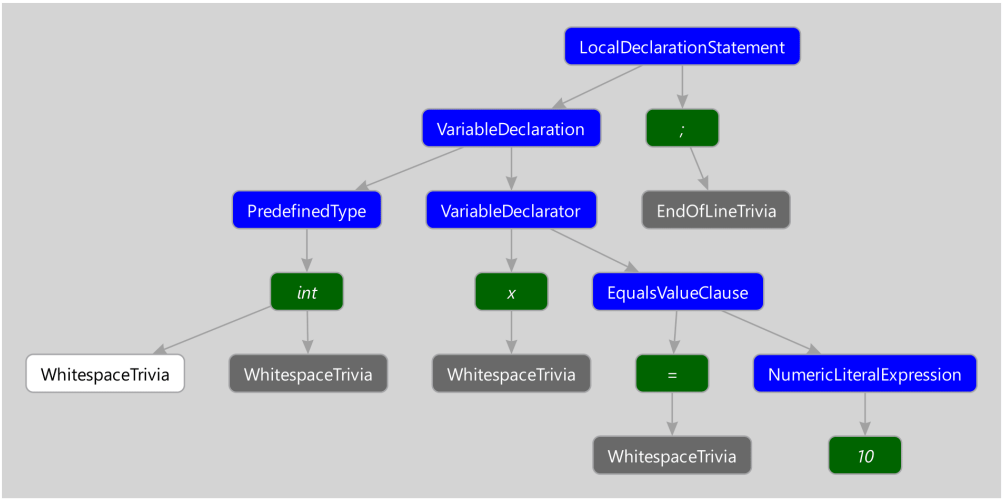


Fig. 9. Syntax tree of local declaration statement

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

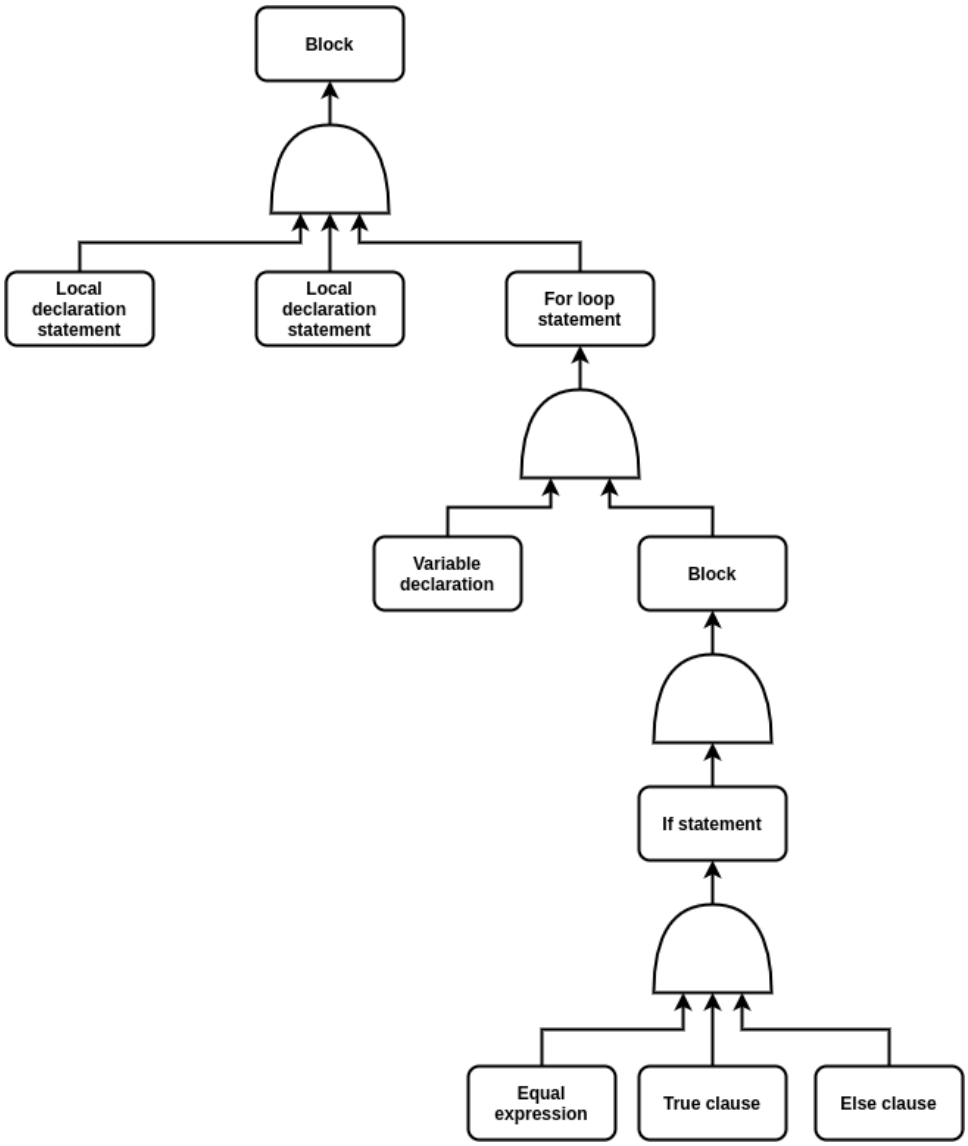


Fig. 10. Example of fault tree created using syntax tree from Fig. 8

C. Quantitative analysis

These information allow us perform quantitative analysis of the system and to calculate characteristics such as a reliability function R. As we mentioned, reliability is a probability system works without failure. From structure function represented in the form of truth table as can be seen in Table III, there is only one row, where system is working and that is the case, when all of its components works. Therefore the reliability of this system can be calculated as following:

$$\begin{aligned} R &= p_1 \times p_2 \times p_3 \times p_4 \times p_5 \times p_6 \\ R &= 0.98 \times 0.98 \times 0.90 \times 0.95 \times 0.96 \times 0.96 \\ R &= 0.7567 \end{aligned} \tag{7}$$

Similarly unreliability U is defined as a probability system will fails. Using the truth table we can see, there are 63 cases, when system will fails. To calculate the unreliability we can either sum the probabilities of these cases or to use calculated reliability:

$$\begin{aligned} U &= 1 - R \\ U &= 1 - 0.7567 \\ U &= 0.2433 \end{aligned} \tag{8}$$

Structure function can be also used to calculate importance measures for individual components. Firstly we will demonstrate the calculation of the structure importance for component x_1. Using direct partial Boolean derivative and information from Table III, we can see, that there is only one case, when the failure of this component results in the failure of the whole system and it is in the case, when all other components are working. The number of all possible

TABLE III
STRUCTURE FUNCTION IN FORM OF TRUTH TABLE

x_1	x_2	x_3	x_4	x_5	x_6	$\phi(\boldsymbol{x})$	x_1	x_2	x_3	x_4	x_5	x_6	$\phi(\boldsymbol{x})$
0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	1	0	0	0	0	1	0
0	0	0	0	1	0	0	1	0	0	0	1	0	0
0	0	0	0	1	1	0	1	0	0	0	1	1	0
0	0	0	1	0	0	0	1	0	0	1	0	0	0
0	0	0	1	0	1	0	1	0	0	1	0	1	0
0	0	0	1	1	0	0	1	0	0	1	1	0	0
0	0	0	1	1	1	0	1	0	0	1	1	1	0
0	0	1	0	0	0	0	1	0	1	0	0	0	0
0	0	1	0	0	1	0	1	0	1	0	0	1	0
0	0	1	0	1	0	0	1	0	1	0	1	0	0
0	0	1	0	1	1	0	1	0	1	0	1	1	0
0	0	1	1	0	0	0	1	0	1	1	0	0	0
0	0	1	1	0	1	0	1	0	1	1	0	1	0
0	0	1	1	1	0	0	1	0	1	1	1	0	0
0	0	1	1	1	1	0	1	0	1	1	1	1	0
0	1	0	0	0	0	0	1	1	0	0	0	0	0
0	1	0	0	0	1	0	1	1	0	0	0	1	0
0	1	0	0	1	0	0	1	1	0	0	1	0	0
0	1	0	0	1	1	0	1	1	0	0	1	1	0
0	1	0	1	0	0	0	1	1	0	1	0	0	0
0	1	0	1	0	1	0	1	1	0	1	0	1	0
0	1	0	1	1	0	0	1	1	0	1	1	0	0
0	1	0	1	1	1	0	1	1	0	1	1	1	0
0	1	1	0	0	0	0	1	1	1	0	0	0	0
0	1	1	0	0	1	0	1	1	1	0	0	1	0
0	1	1	0	1	0	0	1	1	1	0	1	0	0
0	1	1	0	1	1	0	1	1	1	0	1	1	0
0	1	1	1	0	0	0	1	1	1	1	0	0	0
0	1	1	1	0	1	0	1	1	1	1	0	1	0
0	1	1	1	1	0	0	1	1	1	1	1	0	0
0	1	1	1	1	1	0	1	1	1	1	1	1	1

DPBDs for the component x_1 are 32. The resulting SI for component x_1 can be therefore calculated as following:

$$\mathrm{SI}_1 = \mathrm{TD}\left(\frac{\partial\phi(1\to 0)}{\partial x_1(1\to 0)}\right) = \frac{1}{32} \approx 0.03125 \tag{9}$$

Structure importance can be similarly calculated also for other components. The results of this calculation can be seen in Table IV. As we can see, the resulting SI is equal for all components. That means, that from topological point of view every component, or in our case every element of the source code, has the same impact on system reliability, or in our case the same impact for the software to work correctly.

TABLE IV
RESULTS OF SI_i CALCULATION

Component	SI_i
x_1	0.03125
x_2	0.03125
x_3	0.03125
x_4	0.03125
x_5	0.03125
x_6	0.03125

The usage of the information about probabilities of the individual component states presented in Table II allow us to calculate also different importance measure, Birnbaum's importance. A calculation of BI for the component x_1 is similar to the calculation of SI. Using DPBD we find cases, where failure of the component x_1 results in the failure of the system. Then we will calculate the probability of the occurrence of this cases. There is only one case for the component x_1 where DPBD takes value 1 - when all components are working. The Birnbaum's importance for this component can be then calculated as following:

$$\begin{aligned} \mathrm{BI}_1 &= p_2 \times p_3 \times p_4 \times p_5 \times p_6 \\ \mathrm{BI}_1 &= 0.98 \times 0.90 \times 0.95 \times 0.96 \times 0.96 \\ \mathrm{BI}_1 &\approx 0.7722 \end{aligned} \tag{10}$$

The Birnbaum's importance calculus for the other components will be similar to the calculus of the first one. The results of this calculation can be seen in Table V. The results show, that component x_3 is slightly more important for our system that the other components. Component x_3 represents a variable declaration statement in the for loop statement and it is caused by the fact, there is a higher probability of its failure as it is for the other elements of the source code. The second most important component according to results is the component x_4, what represents condition in the if statement. Then there is components x_5 and x_6 and the least important according to Birnbaum's importance are components x_1 and x_2. The results show us, that we need to put a bigger effort and be more careful when writing declarations of the for loop statement and conditions compared to declaring local variables.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE V
RESULTS OF BI_i CALCULATION

Component	BI_i
x_1	0.7722
x_2	0.7722
x_3	0.8408
x_4	0.7966
x_5	0.7883
x_6	0.7883

IV. CONCLUSION

Quantification of the software reliability is a relevant problem of the reliability engineering. Most existing models for the analysis of software reliability are probabilistic models. However these models doesn't take into account topology or structure of the software. There are several models suitable to perform such analysis and in this paper new model is proposed.

The main idea of this approach is to use the source code as a base to create a reliability model. Source code is not easy to use for reliability model creation as is and has to be represented in the suitable way. For this purpose an abstract syntax tree is used. This is an abstract graph representation of the source code. Usage of the abstract syntax tree has several advantages such as programming language independence and it gives us correct description of the structure of the source code. The next advantage is the fact, that correctly constructed abstract syntax tree will automatically ignores unnecessary information present in the source code such as gaps.

A source code represented in the form of the syntax tree can be in the next step used to create reliability model. In this paper fault tree is used. In this model each node of the syntax tree represents one event in the created fault tree. These events are connected via AND gates. Fault tree can be easily transformed into the structure function. This allow us to represent software system as a Boolean function and apply tools for Boolean function analysis, such as differential calculus. This can be used to calculate various characteristics of the software system such as reliability and unreliability using traditional methods of reliability engineering. It also allow us to calculate the importance measures, e.i. characteristics of system components that express us the importance of each component on the system working and failure.

This paper also contains demonstration of the proposed method on a simple source code. This source code is processed into syntax tree, fault tree and the structure function. Structure function is then used to calculate reliability, unreliability and importance measures, specifically structure and Birnbaum's importance using direct partial Boolean derivative.

In our future work, we will focus on the 2 main problems of this model. The first one is its large dimension. As can be seen in this paper, even quite simple code results in fault tree with 9 components and structure function with 6 components. For real software systems, the reliability model will have very large dimension. There are several ways to solve this problem. The first one is by reducing the size of the syntax tree necessary to create this model merging or removing nodes until resulting syntax tree has more suitable size. This way also created reliability model will have reduced size and therefore the whole analysis will be simpler. The other solution is to use methods such as modular decomposition [6], [21], [17]. Using this method reliability model can be divided into several modules and each module can be analysed separately. This will reduce time required to perform reliability calculation of the whole system with usage of methods like parallelism.

The next problem is the usage of the AND gates what can leads to inaccuracy in the created model. In order to solve this problem, the future research is required. This will help us to determine how each event should be connected to other events in order to increase accuracy of the created model.

REFERENCES

[1] Class ast. https://www.ibm.com/support/knowledgecenter/SS5JSH_9.5.0/org.eclipse .jdt.doc.isv/reference/api/org/eclipse/jdt/core/dom/AST.html. (Accessed on 11/07/2020)

[2] The .net compiler platform sdk (roslyn apis) — microsoft docs. https://docs.microsoft.com/sk-sk/dotnet/csharp/roslyn-sdk/. (Accessed on 11/02/2020)

[3] Agrahari, V., Chimalakonda, S.: AST[AR] – towards using augmented reality and abstract syntax trees for teaching data structures to novice programmers. In: 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT). IEEE (2020). URL https://doi.org/10.1109/icalt49669.2020.00100

[4] A.Pasquini E. De Agostino, G.D.M.: An input-domain based method to estimate software reliability. IEEE Transaction on Reliability **45**, 95–105 (1996). URL https://doi.org/10.1109/24.488923

[5] B. Kaiser C. Gramlich, M.F.: State/event fault trees—a safety analysis model for software-controlled systems. Reliability Engineering & System Safety **92**, 1521–1537 (2007). URL https://doi.org/10.1016/j.ress.2006.10.010

[6] Birnbaum, Z.W., Esary, J.D.: Modules of coherent binary systems. Journal of the Society for Industrial and Applied Mathematics **13**(2), 444–462 (1965). URL https://doi.org/10.1137/0113027

[7] Chen, Q., Hu, H., Liu, Z.: Code summarization with abstract syntax tree. In: Communications in Computer and Information Science, pp. 652–660. Springer International Publishing (2019). URL https://doi.org/10.1007/978-3-030-36802-9_69

[8] Duracik, M., Krsak, E., Hrkut, P.: Issues with the detection of plagiarism in programming courses on a larger scale. In: 2018 16th International Conference on Emerging eLearning Technologies and Applications (ICETA). IEEE (2018). URL https://doi.org/10.1109/iceta.2018.8572260

[9] Ďuračík, M., Kršák, E., Hrkút, P.: Source code representations for plagiarism detection. In: Communications in Computer and Information Science, pp. 61–69. Springer International Publishing (2018). URL https://doi.org/10.1007/978-3-319-95522-3_6

[10] Finkelstein, M.: A point-process stochastic model for software reliability. Reliability Engineering & System Safety **63**, 67–71 (1999). URL https://doi.org/10.1016/S0951-8320(98)00014-3

[11] Hegde, C.A.T.M.B.U.: Incorporating software failure in risk analysis – part 1: Software functional failure mode classification. Reliability Engineering & System Safety **197**, 106803 (2020). URL https://doi.org/10.1016/j.ress.2020.106803

[12] J D Musa A. Iannino, K.O.: Software reliability-measurement, prediction, application. McGraw-Hill (1987)

[13] Jelinski, Z., Moranda, P.: SOFTWARE RELIABILITY RESEARCH. In: Statistical Computer Performance Evaluation, pp. 465–484. Elsevier (1972). URL https://doi.org/10.1016/b978-0-12-266950-7.50028-1

[14] Kabir, S.: An overview of fault tree analysis and its application in model based dependability analysis. Expert Systems with Applications **77**, 114–135 (2017). URL https://doi.org/10.1016/j.eswa.2017.01.058

[15] Kang, J., Sun, L., Soares, C.G.: Fault tree analysis of floating offshore wind turbines. Renewable Energy **133**, 1455–1467 (2019). URL https://doi.org/10.1016/j.renene.2018.08.097

[16] Kuo, W., Zhu, X.: Importance Measures in Reliability, Risk, and Optimization. John Wiley & Sons, Ltd (2012). URL https://doi.org/10.1002/9781118314593

[17] Kvassay, M., Rusnak, P., Rabcan, J.: Time-dependent analysis of series-parallel multistate systems using structure function and markov processes. In: Advances in System Reliability Engineering, pp. 131–165. Elsevier (2019). URL https://doi.org/10.1016/b978-0-12-815906-4.00005-1

[18] Kvassay, M., Rusnak, P., Sedlacek, P.: Computation of birnbaum's importance using logic differential calculus. In: 2019 42nd International Conference on Telecommunications and Signal Processing (TSP). IEEE (2019). URL https://doi.org/10.1109/tsp.2019.8768854

[19] Kvassay, M., Zaitseva, E., Levashenko, V., Kostolny, J.: Minimal cut vectors and logical differential calculus. In: 2014 IEEE 44th International Symposium on Multiple-Valued Logic. IEEE (2014). URL https://doi.org/10.1109/ismvl.2014.37

[20] Min Xie Kim-Leng Poh, Y.S.D.: Computing System Reliability. Kluwer Academic Publishers (2004). URL https://doi.org/10.1007/b100619

[21] Natvig, B.: Multistate Systems Reliability Theory with Applications. John Wiley & Sons, Ltd (2011). URL https://doi.org/10.1002/9780470977088

[22] Neamtiu, I., Foster, J.S., Hicks, M.: Understanding source code evolution using abstract syntax tree matching. In: Proceedings of the 2005 international workshop on Mining software repositories - MSR '05. ACM Press (2005). URL https://doi.org/10.1145/1083142.1083143

[23] P. Govindasamy, R.D.: Development of software reliability models using a hybrid approach and validation of the proposed models using big data. The Journal of Supercomputing **76**, 2252–2265 (2020). URL https://doi.org/10.1007/s11227-018-2457-8

[24] P.Munk, A.N.: Model-based safety assessment with sysml and component fault trees: application and lessons learned. Software and Systems Modeling **19**, 889–910 (2020). URL https://doi.org/10.1007/s10270-020-00782-w

[25] P.Roy, Mahapatra, G., P.Rani, S.K.Pandey, K.N.Dey: Robust feedforward and recurrent neural network based dynamic weighted combination models for software reliability prediction. Applied Soft Computing **22**, 629–637 (2014). URL http://dx.doi.org/10.1016/j.asoc.2014.04.012

[26] Rausand, M., Høyland, A.: System Reliability Theory, 2 edn. John Wiley & Sons, Inc., Hoboken, NJ (2004). URL http://www.theeuropeanlibrary.org/tel4/record/3000030635304

[27] Roberto Pietrantuono Peter Popov, S.R.: Reliability assessment of service-based software under operational profile uncertainty. Reliability Engineering & System Safety **204**, 1–13 (2020). URL https://doi.org/10.1016/j.ress.2020.107193

[28] Sedaghatbaf, A., Azgomi, M.A.: Reliability evaluation of UML/DAM software architectures under parameter uncertainty. IET Software **12**(3), 236–244 (2018). URL https://doi.org/10.1049/iet-sen.2017.0077

[29] Shooman, M.L.: Reliability of Computer Systems and Networks. John Wiley & Sons, Inc. (2002). URL https://doi.org/10.1002/047122460x

[30] Tapia, M., Guima, T., Katbab, A.: Calculus for a multivalued-logic algebraic system. Applied Mathematics and Computation **42**(3), 255–285 (1991). URL https://doi.org/10.1016/0096-3003(91)90004-7

[31] Xie, M.: Software Reliability Modeling. Springer (1991)

[32] Yamada, S.: Software Reliability Modeling. Springer (2014)

[33] Yanushkevich, S., Michael Miller, D., Shmerko, V., Stankovic, R.: Decision Diagram Techniques for Micro- and Nanoelectronic Design Handbook, vol. 2. CRC Press, Boca Raton, FL (2005). URL http://www.crcnetbase.com/doi/book/10.1201/9781420037586

[34] Zaitseva, E.N., Levashenko, V.G.: Importance analysis by logical differential calculus. Automation and Remote Control **74**(2), 171–182 (2013). URL http://link.springer.com/10.1134/S000511791302001X

[35] Zang, Z., Wen, Q., Xu, K.: A fault tree based microservice reliability evaluation model. IOP Conference Series: Materials Science and Engineering **569**, 032069 (2019). URL https://doi.org/10.1088/1757-899x/569/3/032069

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Gender recognition using thermal images from UAV

Kateřina Příhodová
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
katerina.prihodova@upce.cz

Jakub Jech
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
jakub.jech@upce.cz

Abstract— **Gender recognition is one of the issues that computer vision deals with. It is useful for analysing human behaviour, intelligent tracking, or human-robot interaction. The aim of this paper is to recognise the gender of people in outdoor areas, where it is very difficult or impossible to guard all access roads to the place, even in poor lighting conditions or in the dark. In this paper, a model will be designed and tested using a controlled UAV flight, during which images of people were obtained. The sensor is a thermal camera located on the UAV, which is not dependent on ambient lighting, and deep learning methods are used for subsequent image processing and classification. These are convolutional neural networks (AlexNet, GoogLeNet), which will be used to solve binary classification. Optimized networks achieve classification accuracy of 81.6 %% (GoogLeNet) and 82.3% (AlexNet). A freely available database [21] was used to learn CNNs, and a self-created database (images obtained with a thermal camera attached to a UAV) was used to test the networks.**

Keywords—gender recognition; thermal image; UAV; convolutional neural networks.

I. INTRODUCTION

In today's modern world, it is often important to reliably automatically identify people, mainly because of the increased security risks. Biometric recognition can be used, defined in [1] as: *"Automatic identification or automatic verification of a person's identity based on physiological and behavioural characteristics."* Physiological characteristics are those with which we were born and included, for example, the iris [2, 3], the retina [4], the face [5], DNA [6]. Behavioural characteristics include, for example, voice [7], writing [8], signature [9], and walking [10].

Next, we will deal with the physiological characteristics - the face. In addition to recognising a person, the human face can also be used to recognise gender, age, origin, and mood [11]. These recognitions are used in many real-world applications, such as marketing or advertising, where gender or age ad serving is very often used. Control of access to premises by gender, e.g., changing rooms, is also addressed [12, 13]. There are other uses in practice, so it is important to research gender recognition to support its role in various areas and achieve an increased level of accuracy.

Gender recognition of persons (male, female) is a classification problem – binary classification. The system that solves this classification problem consists of five modules: sensor, image processing, features extraction, classification and performance evaluation. The necessary data is collected using sensors such as cameras. Image processing includes, in particular, face detection, normalisation and suppression of irrelevant information and noise. This is followed features extraction and classification. The last step is to use test data to evaluate model performance.

One way to improve gender recognition in low light conditions is to use thermal face images. Thermal face images are not used very often. The main reason is the higher purchase price of a thermal camera [14]. However, the use of thermal images has undeniable advantages, especially in terms of lighting conditions. Changes in ambient light have less impact on face images obtained in the infrared spectrum than in the visible spectrum. If it is necessary to determine the sex of people in poor lighting conditions or the dark, it is appropriate to use thermal images.

Another advantage of using thermal images is the possibility of detecting objects under clothing or checking the body temperature of people. This approach is also a non-invasive method for determining the temperature of people. It can be used in various areas of life. Health care represents one of the areas because elevated body temperature is one of the symptoms of infectious diseases. In the last year, the whole world has been paralysed by the COVID-19 pandemic. In this disease, fever is one of the first symptoms, even in 83-99% of cases. Other primary symptoms are significant fatigue and shortness of breath. Later may add dry cough, muscle and joint pain, loss of smell [15, 16]. Just like gender recognition of people, recognising people with fever is a classification problem.

There are situations where we need to solve these recognition problems for people who are outdoors, where it is complicated, if not impossible, to scan all access roads.

There are various ways to capture people in larger areas. In recent decades, there has been a significant expansion of the use of unmanned aerial vehicles as a source of data acquisition and the development of informatics. There is also minimisation of these machines while maintaining aviation parameters. Therefore, UAV with a thermal sensor is a suitable means for obtaining thermal images of people (faces) in public spaces. Flying with a UAV is subject to legislation on the operation of unmanned aerial vehicles [17]. However, under the new legislation, it is possible to fly over uninvolved persons with a UAV weighing up to 250 grams, which can be equipped with a camera. This case concerns the use of this technology over public spaces. In the case of private spaces, different rules apply.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

For private areas, people take part in the scan, and it is possible to fly over them with machines of higher weight, but provided that all safety rules are observed. When using UAVs as a source of thermal images of people in a larger area, it is necessary to thoroughly cover the monitored area, which can be achieved by the planned flight, case study in Figure 1.

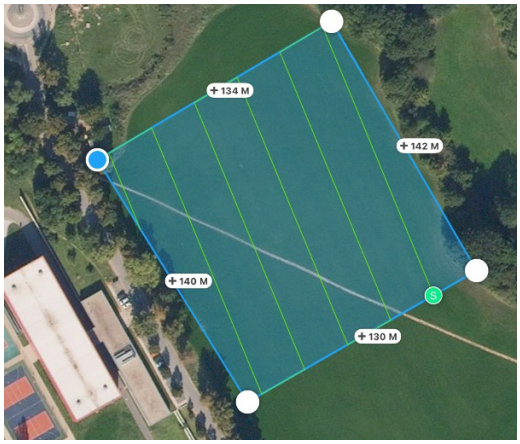


Figure 1. Planned flight

Source: authors

The planned flight can be used to obtain images at waypoints, video or time-lapse with specific settings of the sensor, especially its angle of viewing. To ensure the optimal angle between the scanned face and the sensor, the recommended value is -15 ° against horizontal sensing. And this value is related to the flight level of the UAV. This value was determined on the basis of experimental tests of setting the angle between the face and the UAV and is also related to the flight level of the UAV. When scheduling with a time-lapse acquisition, the appropriate shooting interval needs to be set. This parameter is related to the horizontal flight speed of the UAV. The output of the scheduled flight and capture set in this way is a set of thermal images or video recording in the infrared spectrum.

The aim of this paper is to recognise the gender of people in outdoor areas, where it is very difficult or impossible to guard all access roads to the place, even in poor lighting conditions or in the dark. In this paper, a neural network model will be designed and tested using images collected during a controlled UAV flight.

II. Current Gender Recognition Systems

Many different techniques for determining gender are mentioned in the literature. The most important features that can be used to determine gender are the face (chin, jaw) and the pelvis [18]. Current gender recognition systems mainly use images obtained in the visible part of the electromagnetic spectrum [19,20]. Only a small number of studies focus on gender recognition of face images obtained in the infrared spectrum. Samples of thermal images of the face are shown in Figure 2. Samples of the images are from a publicly available, fully annotated database of thermal images of the face [21].

Figure 2. Examples of images obtained in the infrared part of the electromagnetic spectrum in the first line are images of a man. In the second line, there are images of a woman.

Source: [21]

Thanks to the success of deep learning in many areas of computer vision, it has also begun to gain ground in the field of gender recognition. The advantage of convolutional neural networks (CNNs) is their ability to automatically extract features for classification, while the classical approach involves complex manual creation of features. Several systems have been based on CNNs in recent years that recognise gender and estimate age by face [22]. Not only convolutional neural networks are used for the classification problem of gender recognition, but random forest methods are used [23], and a combination of principal component analysis principles and a genetic algorithm has also been used [24].

There are a large number of contributions on the topic of face recognition. On the other hand, specific gender recognition has been addressed by a lower number of contributions. These papers gender recognition posts work with data in the visible spectrum. If the requirement for contributions for face or gender recognition in the non-visible spectrum, namely in the infrared spectrum, is added, then only a very small number of these contributions can be found [25, 26]. This also applies to contributions where there is a requirement to obtain data using UAVs [27]. As the authors of this paper have previously researched similar topics [28], there is much room for research in this area. See table 1 below.

TABLE I. Search terms in databases Scopus and WOS. Source: Authors.

Search term	***Scopus***	***WOS***
Face recognition	54253	26107
Gender recognition	945	791
Face recognition and thermal image	255	58
Face recognition and UAV	82	13
Gender recognition and thermal image	2	1
Gender recognition and UAV	1	1

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

III. Background of Convolutional Neural Networks

The classification problem can be described by the following formula [29]:

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(l_i, \mathbb{C}(X_i, \theta)) \qquad (1)$$

where $\mathbb{C} : X \rightarrow \hat{l}$ is a classifier that gives the estimated class $\hat{l}$ of features $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ for a given training data which contains n records, and θ is the parameters vector of $\mathbb{C}$. By minimising the loss function $\mathcal{L}$ with respect to θ, $\hat{l}$ approaches to the true label l and the misclassification error is reduced [29].

In recent years, convolutional neural networks have received much attention. These can be used to solve classification problems. Convolutional neural networks first feature extraction and then classify the image. All CNNs have a typical basic architecture, which is shown in Figure 3.

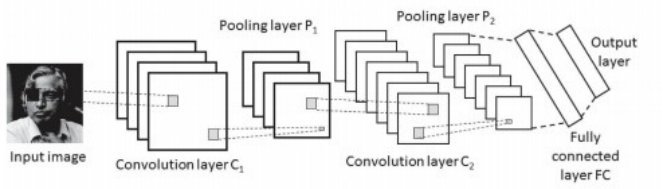


Figure 3. Typical architecture of a convolutional neural network.

Source: [30]

All convolutional neural networks consist of an input layer and hidden layers that have different features. The first hidden layer is a convolution layer with activation functions, such as Relu. This convolution layer features extraction and is used to detect colours and edges. Deeper convolution layers detect more complex tasks. The second hidden layer is the pooling layer. The pooling layer changes the image, followed by another convolutional layering with activation functions followed by a pooling layer. All of these layers provide features extraction. The extraction features are followed by a classification, which is performed using fully interconnected layers with the Softmax activation function.

IV. Methodology

A convolutional neural network-based gender recognition system was developed and validated in MATLAB, and experiments were performed on an Intel Core i7 at 1.2 GHz. The most commonly used convolutional neural networks were gradually tested (AlexNet and GoogLeNet) [31].

The aim of this paper is to recognise the gender of people in outdoor areas, where it is very difficult or impossible to guard all access roads to the place, even in poor lighting conditions or in the dark. In this paper, a model will be designed and tested using a controlled UAV flight, during which images of people were obtained.

As mentioned in the introduction, it is advantageous to use thermal images of the face in poor lighting conditions. The thermal camera can be easily placed at the entry points of buildings. To obtain thermal images of the face in public spaces using fixed thermal cameras at the entry points (many entry points) is costly. Therefore it is advisable to use a combination of thermal camera and UAV. The thermal camera will be attached to the UAV. It is a suitable non-invasive method for obtaining thermal data in areas where coverage by static thermal cameras cannot be sufficiently ensured, typically open space. In this paper, the drone DJI Phantom 3 was used as a sensor, to which was added an external thermal camera FLIR Duo with a resolution of 160x120 pixels, see in Figure 4. The thermal image sensor was set to a temperature scale with fixed points for maximum and minimum sensing values. For quality coverage of the scanned area, it is necessary to use the planned flight. The planned flight guarantees quality coverage of the scanned area and reduces data redundancy compared to an unplanned flight. Flight planning has basic rules for longitudinal and transverse overlaps between individual images. These values can be adapted to the required parameters for the scanned area. A software tool (DJI GS PRO) is used for the planned flight, which calculates and plans flight lines with individual image capture points, so-called waypoints, from the entered parameters for image overlap. The input parameters were a flight altitude of 5 meters, transverse and longitudinal overlap of 60% and manual adjustment of the sensor with a fixed thermal scale. After scheduling the flight, the software calculates the flight time and uploads the flight plan to the UAV, and a separate flight can begin. The output of the flight is individual images with GPS coordinates taken on waypoints, according to the planned flight. The output data set of thermal images is input for other data processing methods are drawn.

Figure 4. Duo DJI Phantom 3 drone, with external FLIR Duo thermal camera.

Source: authors

It was necessary to detect the faces of individuals in the images obtained during the planned flight. The face was chosen for gender recognition for practical reasons. This is a part of the body that does not require the cooperation of the scanned persons. For the last two decades, the face has been used to identify people, recognise gender, ethnicity, estimate age [19, 20].

After face detection using an algorithm from Ren et al. [32], it was necessary to classify the detected face. The classification was performed by convolutional neural networks. The most commonly used convolutional neural networks (AlexNet, GoogLeNet) [31] were gradually tested in the models.

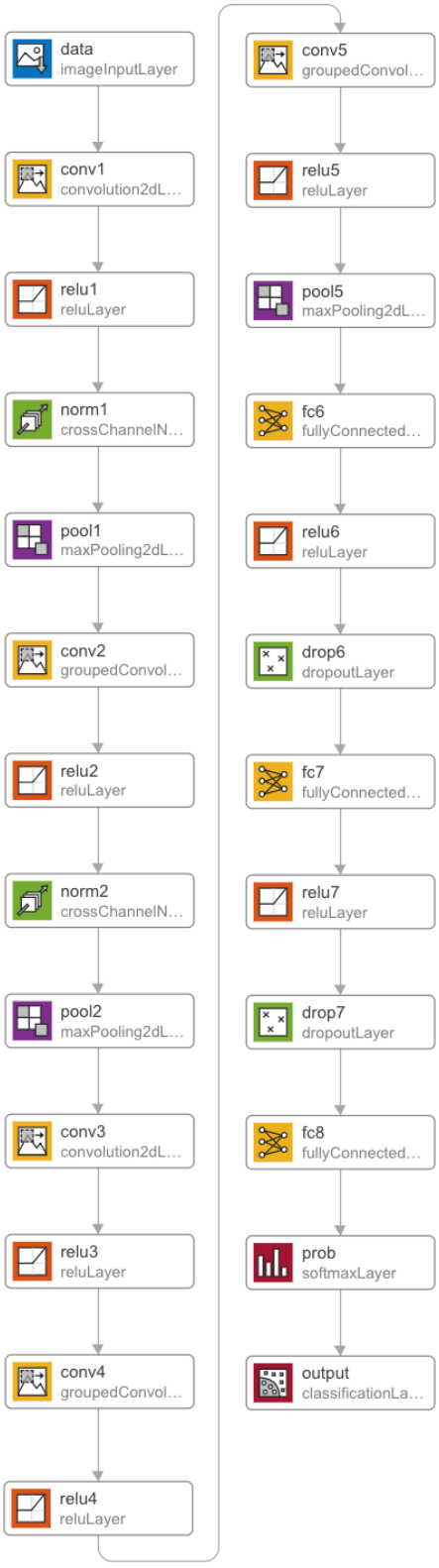


Figure 5. Typical architecture of an AlexNet.

Source: authors

The AlexNet network (8 layers) was the first to be tested, and it is one of the first networks to popularise CNN in the field of computer vision. Its architecture is shown in Figure 5. The newer network that has been tested is the GoogLeNet (22 layers). In this network, there has been a rapid reduction in parameters compared to AlexNet.

A "Fully Annotated Database of Face Thermal Imaging" was used to train convolutional neural networks [21]. It is one of the few large databases of thermal images of the face. It contains a total of 2 907 thermal images of faces from 74 men and 14 women. The division into categories (man, woman) was done manually. Since there are fewer images of women's faces (424) in the database than images of men (2 483), image transformations were used for images of women to make the numbers of images in each category similar. Specifically, image transformations were used - mirror image rotation, image rotation in the range (-20 – 20), magnification. Furthermore, testing of neural networks was performed on thermal images obtained using UAVs. It contains 203 images from 5 men and 5 women. Examples of images obtained using UAV are shown in Figure 6. Thermal images of people's faces from both databases were resampled using the nearest neighbour interpolation method to sizes (227 x 227 x 3 for AlexNet and 224 x 224 x 3 for GoogLeNet); these are the required input sizes for given convolutional neural networks.

Figure 6. Examples of images obtained in the infrared part of the electromagnetic spectrum using UAVs in the first line are images of a man, in the second line, images of a woman.

Source: authors

First, both convolutional neural networks were trained on training data. The backpropagation algorithm was used for learning, which uses stochastic gradient descent with a momentum optimisation algorithm to minimise errors. Due to the computing technology on which the implementation was performed and the size of the dataset, the batch size was set at 32. The epoch was set to 20. Finding the network parameters that ensured the learning error value kept below the required limit is the result of many different test scenarios. The Learning Rate (0.001) and Momentum (0.95) parameters were optimised. Then, the proposed model was verified using test data, and the accuracy of the system was determined. In the end, a comparison and comparison with existing approaches were made.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

V. RESULTS

Figure 7 shows the activations of the first convolutional layer of the trained AlexNet network. You can see in the figure which functions are extracted in this layer.

Figure 7. Features visualisation of the first convolutional layer

Source: authors

Table 2 shows the results – accuracy of classification using individual neural networks. When learning networks, the network achieved GoogLeNet, but AlexNet achieved better results by 82.3% when tested.

TABLE II. TABLE ACCURACY OF CNNS. SOURCE: AUTHORS.

CNN	*Training Accuracy*	*Test Accuracy*
AlexNet	84,9%	82,3%
GoogLeNet	85,1%	81,6%

Table 3 compares with related works. Because, to the best of our knowledge, this is the first paper to deal with gender recognition using UAV thermal images. It is not possible to compare the accuracy of the classification with the method that uses similar images, similarly obtained. Compared to conventional methods, where visible images are used for gender recognition. Our method has good classification accuracy. Methods using visible images have higher classification accuracy but deteriorates with deteriorating lighting conditions. Our method also has good results compared to the use of thermal images or a combination of thermal images with visible images.

TABLE III. COMPARISON WITH RELATED WORKS. SOURCE: AUTHORS.

Sources	*Techniques Applied for Recognition*	*Size of the Database*	*Accuracy*
Gao & Ai (2009) [33]	ASM/ Adaboost	Private, 1 300 **visible images**	92,89%
Cao et al., (2011) [34]	Metrology/ SVM	MUCT, 276 **visible images**	86,83%
Shan (2012) [35]	LBP/SVM	More databases, 17 814 **visible images**	94,81%
Chen & Ross (2011) [36]	LBP/SVM LBP/Random forest	Private, 1 003 **thermal images**	90,41% 85,55%
Wang (2015) [37]	thermal statistical temperature and LBP /Bayesian Networks	NVIE 532 and Equinox 1 269 **visible and thermal image**	83,3% 86,2%
Proposed method	CNN - AlexNet	Annotation database [21] 2 557 + Private UAV 203 **thermal image**	82,3%
Proposed method	CNN - GoogLeNet	Annotation database [21] 2 557 + Private UAV 203 **thermal image**	81,6%

VI. CONCLUSION

We have proposed using UAVs to acquire thermal images of persons in the outdoor environment, where it is not possible to obtain thermal images of persons with fixed thermal cameras. Subsequently, we detected a face in each image.

CNNs (AlexNet and GoogLeNet) were optimised for gender recognition. Compared to similar studies, the results of optimised networks have shown that the proposed CNNs have good classification performance. Convolutional neural networks achieving classification accuracy of 81.6 %% (GoogLeNet) and 82.3% (AlexNet). When using a freely available database [21] for learning networks and own databases (images obtained using a thermal camera connected to a UAV) for networks testing.

Due to the GDPR, we decided not to publish the database of face images created by us.

Future solutions are an improvement in the acquisition of input data using more suitable UAVs to acquire thermal images. A suitable choice is the DJI Mavic 2 Enterprise ADVANCED, which has a built-in thermal camera with a resolution of 640x512 pixels. This drone is one of the more expensive in its category, but thanks to its higher resolution, it allows obtaining images from higher flight levels while maintaining image quality.

Research work can be extended to the task of detecting people with elevated body temperature, which is one of the symptoms of viral diseases; recently, the biggest problem is the viral disease COVID-19. Recognising people with fever in public places could be one way to fight the disease. CNNs of similar architectures can be used for this classification.

ACKNOWLEDGMENT

This article was supported by grant No. SGS_2021_08, No. SGS_2021_011, and No. SGS_2020_018 of the Student Grant Competition.

REFERENCES

[1] J. Wayman, "A Definition of "Biometrics". Collected works" US National Biometric Test Center, San José State University. August, 2003, pp. 20-23.

[2] AWK. Kong, D. Zhang and M.S. Kamel, "An Analysis of IrisCode," IEEE Transactions on Image Processing. 2010, vol. 19, no. 2, pp. 522-532.

[3] D. Nguyen, T. Pham, Y. Lee and K. Park, "Deep Learning-Based Enhanced Presentation Attack Detection for Iris Recognition by Combining Features from Local and Global Regions Based on NIR Camera Sensor," Sensors, 2018, vol. 18, no. 8.

[4] S. M. Lajevardi, A. Arakala, S. A. Davis and K. J. Horadam, "Retina Verification System Based on Biometric Graph Matching," IEEE Transactions on Image Processing. 2013, vol. 22, no. 9, pp. 3625-3635.

[5] D. Wang, H. Lu and M.-H. Yang, "Kernel collaborative face recognition," Pattern Recognition, 2015, vol. 48, no. 10, pp. 3025-3037.

[6] D. Frumkin, A. Wasserstrom, A. Davidson and A. Grafit, "Authentication of forensic DNA samples. Forensic Science International: Genetics," 2010,vol. 4, no. 2, pp. 95-103.

[7] "Voice biometrics growth likely," Biometric Technology Today. 2009, vol. 17, no. 6, pp. 3-4.

[8] T. Scheidat, M. Kalbitz and C. Vielhauer, "Biometric authentication based on 2D/3D sensing of forensic handwriting traces," IET Biometrics. 2017, vol. 6, no. 4, pp. 316-324.

[9] T. Hafs, L. Bennacer, M. Boughazi and A. Nait-Ali, "Empirical mode decomposition for online handwritten signature verification," IET Biometrics, 2016, vol. 5, no. 3, pp. 190-199.

[10] T. Lee, M. Belhatir and S. Sanei, "A comprehensive review of past and present vision-based techniques for gait recognition" Multimedia Tools and Applications, 2014, vol. 72, no. 3, pp. 2833-2869.

[11] Z. Yang and H. Ai, "Demographic classification with local binary patterns," ICB, pp. 464-473, 2007.

[12] A. Dantcheva, P. Elia, A. Ross, "What else does your biometric data reveal? A survey on soft biometrics," Transactions on Information Forensics and Security, 2016, vol. 11, no. 3, pp. 441-467.

[13] E. Mkinen, R. Raisamo, "Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces," IEEE Trans. Pattern Analysis and Machine Intelligence, 2008, vol. 30, no. 3, pp. 541-547.

[14] H. D. Benitez-Restrepo, A. C. Bovik and C. G. Rodriguez Pulecio, "Image quality assessment to enhance infrared face recognition," IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 805-809.

[15] Infectious diseases. World Health Organization. 2020. Retrieved from: https://www.who.int/topics/infectious_diseases/en/

[16] X. Mei, Y. Zhang, H. Zhu, Y. Ling, Y. Zou, Z. Zhang, et al. "Observations about symptomatic and asymptomatic infections of 494 patients with COVID-19 in shanghai, china," American Journal of Infection Control, 2020, vol. 48, no. 9, pp.1045-1050.

[17] Unmanned Aircraft. Civil Aviation Authority Czech Republic. Retrieved from: https://www.caa.cz/en/flight-operations/unmanned-aircraft/

[18] S. R. Loth, M. Y. Iscan, "Sex Determination Encyclopedia of Forensic Sciences," San Diego, CA, USA:Academic, 2000, vol. 1.

[19] V. Carletti, A. Greco, G. Percannella, and M. Vento. "Age from faces in the deeplearning revolution," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, vol. 42, no. 9, pp. 2113-2132.

[20] K. Prihodova, and J. Jech, "Low-Cost System for Gender Recognition Using Convolutional Neural Network," 34th International-Business-Information-Management-Association (IBIMA) Conference, Madrid, Spain, 2019, pp. 6316-6322.

[21] M. Kopaczka, R. Kolk and D. Merhof, "A fully annotated thermal face database and its application for thermal facial expression recognition," IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Houston, TX, USA, 2018, pp. 1-6.

[22] I. Rafique, A. Hamid, S. Naseer, M. Asad, M. Awais and T. Yasir, "Age and Gender Prediction using Deep Convolutional Neural Networks," 2019 International Conference on Innovative Computing (ICIC) , Lahore, Pakistan, 2019, pp. 1-6.

[23] K. Khan, M. Attique, I. Syed and A. J. S. Gul, "Automatic gender classification through face segmentation", vol. 11, no. 6, pp. 770, 2019.

[24] A. Geetha, M. Sundaram and B. Vijayakumari, "Gender classification from face images by mixing the classifier outcome of prime distinct descriptors", soft computing, vol. 23, no. 8, pp. 2525-2535, 2019.

[25] D. T. Nguyen and K.R. Park, "Body-based gender recognition using images from visible and thermal cameras," Sensors, 2016, vol.16, no. 2.

[26] D. T. Nguyen and K.R. Park, "Enhanced gender recognition system using an improved histogram of oriented gradient (HOG) feature from quality assessment of visible light and thermal images of the human body," Sensors, 2016, vol. 16, no. 7, pp. 1134.

[27] N. Davis, F. Pittaluga and K. Panetta, "Facial recognition using human visual system algorithms for robotic and UAV platforms," In: 2013 IEEE Conference on Technologies for Practical Robot Applications (TePRA). IEEE, 2013. pp. 1-5.

[28] K. Prihodova and J. Jech, "Prevention of the Spread of Viral Disease Using Artificial Intelligence from Data Obtained by UAVs," In: SHS Web of Conferences. EDP Sciences, 2021. pp. 01042.

[29] M. Afifi, "11K Hands: gender recognition and biometric identification using a large dataset of hand images," Multimedia Tools and Applications, 2019, vol. 78, no. 15, pp. 20835-20854.

[30] K. Prihodova, and M. Hub, "Hand-Based Biometric System Using Convolutional Neural Networks," Acta Informatica Pragensia, 2020, vol. 9, no. 1, pp. 48-57.

[31] A. Karpathy, J. Johnson and L. Fei-Fei, CS23. 1n: Convolutional Neural Networks for Visual Recognition [online]. 2020 [Webpage]. Retrieved from http://cs231n.github.io/convolutional-networks/

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, vol. 39 no. 6, pp. 1137-1149.

[33] W. Gao and H. Ai, "Face gender classification on consumer images in a multiethnic environment" International Conference on Biometrics. Springer: Berlin, Heidelberg. 2009 pp. 169-178.

[34] D. Cao, C. Chen, M. Piccirilli, D. Adjeroh, T. Bourlai, and A. Ross, "Can facial metrology predict gender" 2011 International Joint Conference on Biometrics (IJCB), IEEE, 2011 pp. 1-8.

[35] C. Shan, "Learning local binary patterns for gender classification on real-world face images," Pattern recognition letters. 2012, vol. 33 no. 4, pp. 431-437.

[36] C. Chen and A. Ross, "Evaluation of gender classification methods on thermal and near-infrared face images," 2011 international joint conference on biometrics (IJCB), IEEE, 2011 pp. 1–8.

[37] S. Wang, Z. Gao, S. He, M. He and Q. Ji, "Gender recognition from visible and thermal infrared facial images," Multimedia Tools and Applications, 2016, vol. 75, pp. 8419–8442 .

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Long-wave infrared remote sensing data spatial resolution enhancement using modulation transfer function fusion approach

S. A. Stankevich, M. S. Lubskyi, A. R. Lysenko
Scientific Centre for Aerospace Research of the Earth, Institute of Geological Science, National Academy of Sciences of Ukraine
Kiev, Ukraine

Abstract— **Remote sensing imagery in the long-wave infrared band represents land surface thermal radiance which is crucial for many studies related to urbanization development, geological activity, climate and landscape changes, etc. The main problem of long-wave infrared data application is insufficient spatial resolution relative to visible and short-wave infrared data. For detailed surface temperature mapping, which can be derived from the long-wave infrared data the resolution enhancement needs to be applied. The resolution enhancement technique based on the imaging system's modulation transfer functions (MTF) processing is proposed. As a basic, the high-resolution visible band image has been taken. Fusion of the low-resolution (long-wave infrared) and the high-resolution image (visible) was conducted by inverse filtering within the frequency domain. The more sharpened land surface temperature (LST) distribution without technical improvements of imaging systems is provided.**

Keywords — remote sensing; long-wave infrared radiance; spatial resolution enhancement; frequency domain; discrete Fourier transform; modulation transfer function

I. INTRODUCTION

Satellite remote sensing applications need a trade-off between the spatial and the temporal resolution of the data acquired. Deriving highly detailed image data requires a significant focal length of the imaging system, therefore, significantly reducing the field of view (FOV) and satellite revisit period correspondingly. These performance features are essential for both long and short-term Earth surface monitoring. For example, the WorldView-2 satellite sensor with 13.3 m focal length and 1.8 m spatial resolution (multispectral) has 1.28° FOV meanwhile Landsat-8 OLI satellite sensor with 0.886 m focal length and 30 m spatial resolution in visible, near-infrared (VNIR), and short-wave infrared (SWIR) spectral bands has 15° FOV [1]. Approximately the same FOV provides a long-wave infrared (LWIR) TIRS sensor, but with a resolution of 100 m. It is obvious that an additional problem lies in the low spatial resolution of the LWIR data. It arises from lower quantum energy and higher wavelength of LWIR radiance relative to VNIR spectra.

Spatial resolution enhancement is the relevant challenge in the remote sensing imagery processing. It resolves the problem of deriving additional information from the image without imager hardware improvement. Frequency-domain processing is one of the most common approaches for the imagery resolution enhancement.

The proposed technique of remote sensing data spatial resolution enhancement considers image's MTFs estimating and further separating of images' frequency components. High-frequency components of the high-resolution VNIR data are extracted for fusion with the low-resolution LWIR data through MTFs fusion. Fused frequency-domain data is transformed back into the spatial-domain data with enhanced spatial resolution using inversed discrete Fourier transform (DFT).

II. STATE OF THE ART

The main idea of frequency-domain processing is the signal splitting onto the set of harmonics of different frequencies. Obtained in this way frequency spectrum can be subdivided into several spectral components with different amplitude and phase, which could be processed separately [2]. High-frequency components represent high-contrast edges in the image. This part of the frequency spectrum is derived from detailed images and could be joined with the low-frequency spectrum of the same scene. Such superresolution approach is adopted for color imagery enhancement [3] by extracting the high-frequency spectrum from the monochromatic broadband image. This high-frequency spectrum is used to replace the same frequency components in the raw low-resolution monochromatic data to enhance the resolution and turn it back into a multiband image.

Many satellite multispectral imaging systems contain a panchromatic band. It overlaps several visible bands by spectrum and provides little spectral information but possess a higher spatial resolution. The panchromatic band is typically used for the spatial resolution enhancement. In the frequency domain, high-frequency components of the low-resolution image could be extracted from the panchromatic band with further fusion with a low-frequency component of the multispectral low-resolution image [4].

One of the main benefits of combining spatial-domain and frequency-domain processing techniques is the ability to avoid noise and frequency components losses in transition from one domain to other. The Fourier spectrum contains complete information about the input image.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Fourier-wavelet transform [5] is an efficient superresolution technique for the set of an aliased images based on splitting images onto small tiles with further estimating of the subpixel motion of the tiles, applying suboptimal multi-frame Wiener Filter for fusing and deblurring, as well as a wavelet analysis for resulting image denoising. Long-wave infrared data spatial resolution enhancement requires not only edge sharpening and scene details extraction but also radiometric consistency between input image and enhanced one. Thus, LWIR satellite imagery processing with multi-frame techniques is restricted by several conditions. Processed imagery has to contain invariant data to avoid radiometric errors.

Other known discrete-wavelet transform techniques [6] due to interpolation allow data degradation compensating within texture edges. This approach implies input image decomposing into several high-frequency subbands. Thus, a NEDI (new-edge directed interpolation) technique used for high-frequency subbands estimating with further image high-resolution reconstruction is proposed.

A. Surface temperature estimation

The low-resolution long-wave infrared data is typically interpolated and doesn't contain a high-frequency component for high-contrast edges extraction. It is necessary to include in processing procedure data, which could be a source of high-frequency component for enhanced surface temperature image calculated from the low-resolution radiance data. According to the Planck's law, temperature is estimated from the LWIR radiance data as following [7]:

$$T = \frac{c_2}{\lambda \ln\left(\frac{\varepsilon(\lambda) c_1}{\lambda^5 L(\lambda, T)} + 1\right)}, \quad (1)$$

where T – land surface temperature, $L(\lambda, T)$ – surface long-wave infrared radiance obtained from the satellite data, $\varepsilon(\lambda)$ – spectral emissivity of the land surface, c_1 and c_2 – radiation transfer constants, λ – radiance median wavelength.

Land surface emissivity, which describes thermal emission capability, as a rule, is unknown. It could be determined from the VNIR satellite imagery [8]. This is the key data for high-frequency components extraction for the resolution enhancement.

It is not possible to enhance the resolution without the intrusion of the external information into the infrared image. Such information is contained in additional images [9], in the scene objects' spectra [10], in the land cover classes' topology [11], or elsewhere [12]. Obtaining additional information that is necessary for infrared image resolution enhancement is also possible in the frequency domain [13].

B. Modulation transfer function

The modulation transfer function is the most important characteristic of the imaging system, which determines the resolution and overall quality of the images acquired. MTF fully describes the reproducing properties of the imaging system. It is objectively measurable and is based on well-known mathematics [14].

The imaging system's response is the smallest imagery detail that an optical system could form with perceptible contrast relative to the background. Impulse response could be represented as a surface irradiance (W/cm2) distribution as a function of position. In the imaging system's modeling process the radiance distribution $E(x, y)$ is expressed through the convolution of the impulse response $h(x, y)$ and the ideal input image $E_0(x, y)$ [15]:

$$E(x, y) = E_0(x, y) \otimes h(x, y) \quad (2)$$

The ideal image is performed as an irradiance distribution in the image if the imaging system has the perfect reproducing quality with preservation of all details.

The MTF determines the contrast difference that could be reproduced by the imager relative to the contrast of the original object that is submitted as an ideal image. For demonstration of the functional relation between the spatial distribution of the image content and the irradiance power, image might be transformed into frequency domain [16]. MTF represents the precision of the obtained imagery's frequency content relative to the original scene, captured by the imager. It depends on the imaging system's spatial resolution, the distance to the land surface, signal-to-noise ratio (SNR), system's sensitivity, and radiometric resolution.

MTF generalizes three other imaging system's features that are represented as functions as well [17]:

- point-spread function (PSF);
- line-spread function (LSF);
- edge-spread function (ESF);

These functions characterize the imaging system's capability of the main three corresponding imagery elements imaging in the frequency domain (Fig. 1).

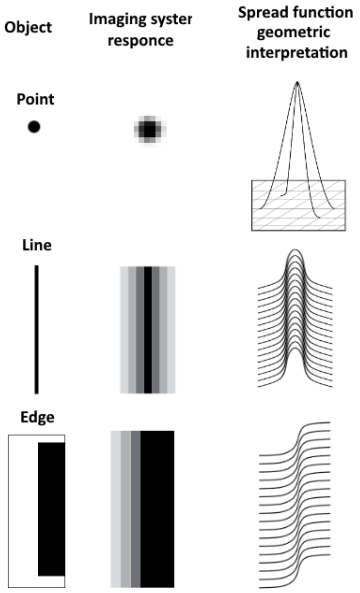


Figure 1. Functional interpretation of the three main elements of the image: PSF, LSF and ESF (top-down)

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

As it is obvious form Fig. 1, in the most commonly case the one-dimensional point spread function can be approximated by a Gaussoid:

$$h(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-x_0)^2}{2\sigma^2}} \quad (3)$$

where x_0 is the PSF location and σ is the Gaussoid's parameter.

The LSF expression is estimated by the superposition of the PSFs (3) for all the line points, and ESF is estimated as the integral of LSF [17]. LSF and PSF are appropriate for systems' experimental spatial resolution measurement and determination of the difference in modulation that represents an ideal image and an obtained image. Modulation K_0 is defined as:

$$K_0 = \frac{|E_0 - E|}{E_0 + E} \quad (4)$$

where E_0 and E are the optical signal values in the object on the image and surrounding background. The modulation transfer ratio T could be expressed as following:

$$T = \frac{K}{K_0} \quad (5)$$

where K is the modulation within the obtained image (Fig. 2).

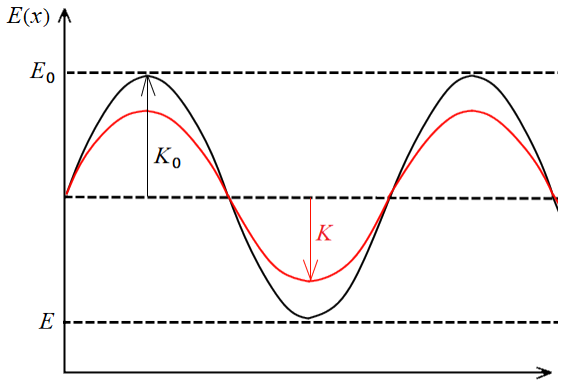


Figure 2. Modulation of the ideal input image (K_0) and image, obtained by imaging system (K)

From Fig. 2 it can be assumed, that MTF characterizes, how well the imaging system is transferring the modulation of the object. MTF $T(\nu)$ of the imager, taking into account (2), is defined as the Fourier transform of the PSF [18]:

$$T(\nu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(x)\, e^{i2\pi\nu x}\, dx \quad (6)$$

where ν is the spatial frequency, x is the spatial distance.

Higher spatial distance provide higher modulation due to the lower quantity of transition, and on the contrary, high frequencies cause blurred image due to the impossibility of transferring the modulation on the short spatial distance. Thus, the typical MTF can be represented as shown in Fig. 3.

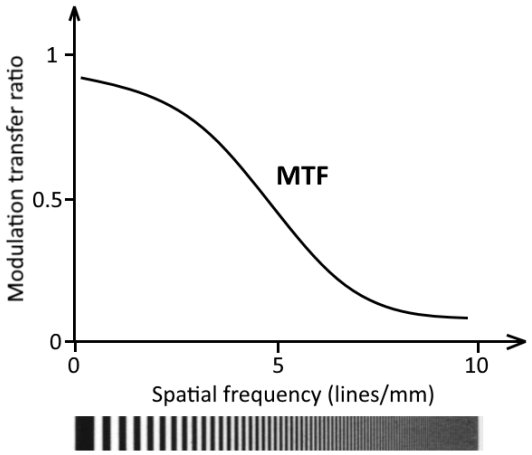


Figure 3. Modulation transfer ratio of the optical image

An increment of modulation transfer ratio could be reached by mixing the high-frequency components derived from the Fourier spectrum of the high-resolution image, and low-frequency components of the input low-resolution image.

III. Materials and Methods

The optical signal path, starting from the target, through the environment and the imager up to the image formation, can be described by a sequence of physical links. The MTF benefit is the ability to estimate the overall effect by the simple multiplication of individual link's MTFs.

The main links of the optical signal path in remote sensing usually address the target, the atmosphere, the optical lens system, and the photodetector sensor array, as shown in Fig. 4. Each link is described by its MTF. Sometimes the optical path includes additional links for the image motion, image processing, and image display, but in modern imaging systems such links can be neglected [19].

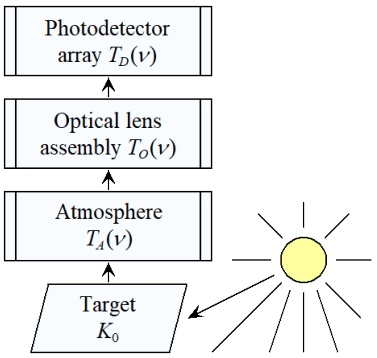


Figure 4. The optical path links

The optical signal in the image $E(\nu)$ is determined through the links' MTFs $T_i(\nu)$ within the optical path as

$$E(\nu) = E_0(\nu) \prod_i T_i(\nu) \quad (7)$$

where $E_0(\nu)$ is the input optical signal of the system, ν is the spatial frequency.

The MTF of the individual links within the optical path Fig. 1 are known from the frequency analysis theory. The target is usually characterized by its constant contrast K_0.

The precise calculation of the MTF of the atmosphere is quite complex, but a simplified model can be applied for a preliminary assessment [20]:

$$T_A(\nu) = K_A e^{-2(\pi\sigma_A f\nu)^2} \tag{8}$$

where K_A is the light scattering coefficient, σ_A is the atmospheric turbulence factor, f is the lens focal distance.

The optical lens system's MTF mainly describes the effect of diffraction on aperture and aberrations [21]:

$$T_O(\nu) = T_D(\nu)\, T_B(\nu) \tag{9}$$

where

$$T_D(\nu) = 1 - 1.32\lambda \frac{f}{D}\nu \tag{10}$$

is the diffraction MTF,

$$T_B(\nu) = e^{-B\frac{D^2}{f}\operatorname{tg}\beta\nu} \tag{11}$$

is the aberrations MTF. In equations (10) and (11) B is the aberration coefficient, λ is the operating wavelength of the optical signal, D is the aperture diameter, and β is the half of the field of view.

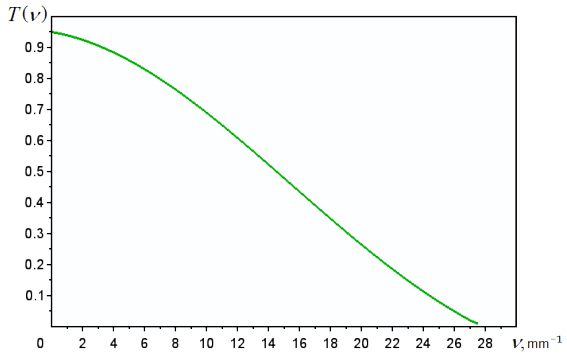
MTF of the discrete sensor array in the simplest case of square photodetectors of $d \times d$ size with uniform sensitivity and infinitely narrow boundaries takes the form of [22]:

$$T_S(\nu) = \frac{\sin \pi d\nu}{\pi d\nu} \tag{12}$$

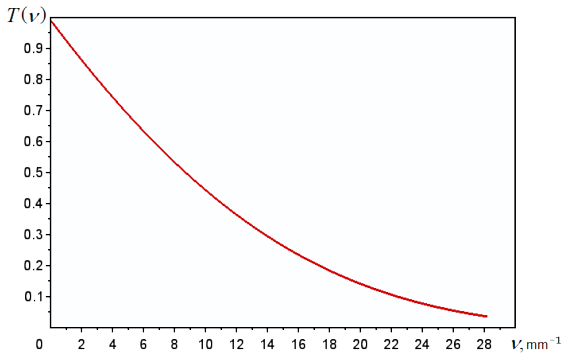
Finally, the complete MTF of the imaging system is described by the multiplication of (8) – (12):

$$T(\nu) = T_A(\nu)\, T_D(\nu)\, T_B(\nu)\, T_S(\nu) \tag{13}$$

Theoretical MTFs of on-board imagers of the Landsat-8 remote sensing satellite are shown in Fig. 5.



a



b

Figure 5. Theoretical MTFs of the Landsat-8 satellite: *a* – OLI visible, near and short-wave infrared bands imager. *b* – TIRS long-wave infrared bands imager

Theoretical MTFs Fig. 5 were calculated using an equation (13). Partial MTFs of atmosphere, optical lens assemblies, and photodetectors arrays were considered for both the OLI [23] and the TIRS [24] imagers.

MTF-based apparatus can be used for a wide range of remote sensing image simulations, taking into account the properties of the imaging systems. Among other things, the MTF tool allows conversion of images of one sensor into another one, as well as the elimination of the unwanted distortions in the images, under the condition that the Fourier transform of the distorting driver is known. The latter mentioned approach is called the inverse filtering because in the simplest case it consists of multiplication by a composite, which is inversed to the disturbance [25].

With inverse filtering, the long-wave infrared image can be rearranged to match one's frequency-response to the visible band image. To perform this, the infrared Fourier transform EIR(ν) must be modified as follows:

$$E_{IR-V}(\nu) = E_{IR}(\nu)\frac{T_V(\nu)}{T_{IR}(\nu)} \tag{14}$$

where $E_{IR\text{-}V}(\nu)$ is the visible-modified Fourier transform of the input long-wave infrared image, $T_V(\nu)$ is the visible imager MTF, and $T_{IR}(\nu)$ is the infrared imager MTF.

Since the frequency properties of the visible imager are much better than the infrared, the method (8) applying will enhance the spatial resolution of the long-wave infrared image without the engagement of any other image. All information required for this is already contained in the MTF models. Also, suchlike virtual fusion will be appropriate in the further joint analysis of the infrared and visible images.

IV. Results and Discussion

There are plenty different notations of the Fourier transform. Thus, before the method is described, we should make it clear what is meant by the two-dimensional Fourier transform.

Let E be an image of size $M \times N$. Let the two-dimensional Fourier transform $E(u, v)$ be:

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

$$E(u,v)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}E(m,n)\exp\left[-i2\pi\left(\frac{mu}{M}+\frac{nv}{N}\right)\right] \quad (15)$$
$$u = 0, ..., M–1, v = 0, ..., N–1$$

Having the source image E of size $M\times N$, the MTF of the visible imager $T_V(u, v)$ and the MTF of the infrared imager T_{IR}(u, v), the high-frequency inverse filter $H(u, v)$ can be described as follows:

$$H(u,v)=\frac{T_v[s(u,v)]}{T_{IR}[s(u,v)]} \quad (16)$$

where

$$s(u,v)=\frac{r(u,v)\,v^*}{\sqrt{\left(\frac{M}{2}\right)^2+\left(\frac{N}{2}\right)^2}}$$

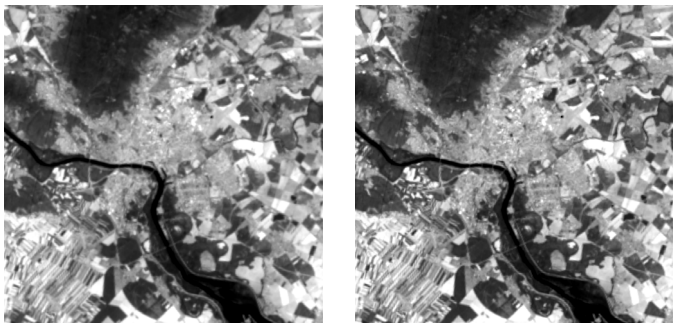
$$r(u,v)=\sqrt{\left(\frac{M}{2}-u\right)^2+\left(\frac{N}{2}-v\right)^2}$$

$\left(\frac{M}{2},\frac{N}{2}\right)$ is the center of the Fourier transform of the input image, $s(u, v)$ is the MTF's argument, $r(u, v)$ is the distance from the center of the Fourier spectrum to the processing pixel (u, v), and v^* is the cut-off frequency of the MTF.

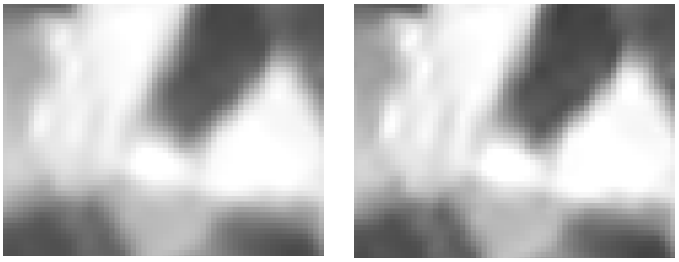
The described above filter is applied as follows:

- the Fourier transform $E(v)$ of the input image (15) is calculated;
- the zero-frequency component of $E(v)$ is being shift to the center of the spectrum;
- the shifted Fourier transform is then element-wise multiplied by $H(u, v)$ filter;
- the inverse Fourier transform of the previous step is done;
- the result is the absolute value of the inverse Fourier transform.

This algorithm was applied on the Landsat-8 image of the capital of Slovakia Bratislava (ID LC08_L1TP_189027_20200801) captured on August 1, 2020. Images from OLI (30 m ground resolution) and TIRS (100 m ground resolution) imagers are available. Fig. 6 include the thermal image fragments acquired by the TIRS long-wave infrared imager – both the input and the enhanced images.



a *b*

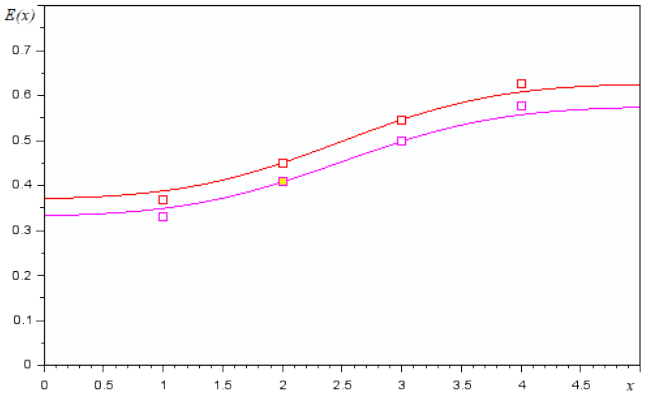


c *d*

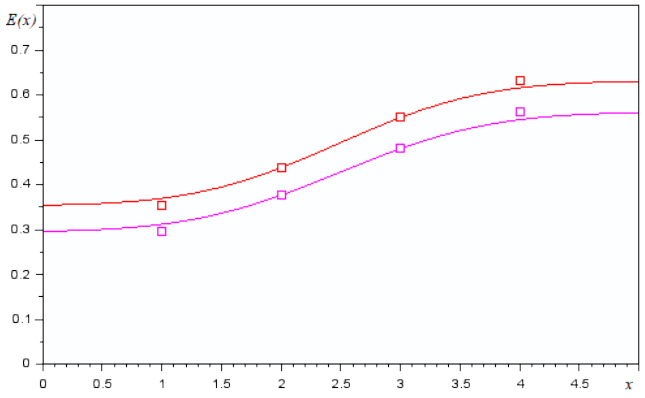
Figure 6. Input satellite image of Bratislava in long-wave infrared band: *a* – input image, *b* – enhanced resolution image, *c*, *d* – zoomed fragments

The actual spatial resolution of the images was estimated objectively by the MTF's Gaussoid approximation acquired from the bidirectional ESF of the image [26]. A special technique for automatic bidirectional ESF extraction from the digital image has been developed and applied for this purpose.

The ESFs of the input and enhanced resolution images can show reached improvement.



a



b

Figure 7. Comparing of biderectional ESFs of the input image (*a*) and enhanced image (*b*)

According to the obtained ESFs parametrization, the isotropic resolution of the input image is 3.89 pixels and the isotropic resolution of the enhanced one is 3.64 pixels. Thus, the enhanced image gave an improvement in spatial resolution by 6.94%.

V. CONCLUSIONS

In this paper, the infrared imagery resolution enhancement technique based on the MTF fusion approach has been described. This technique supposes the input infrared image transformation inside the frequency domain so that its frequency response will be matched to a better image of the visible band. The inverse filtering with an ad hoc determined high-frequency filter (16) performs this.

The resolution increment is insignificant, but the main advantage of this approach is a reduction of edge blurring. Such effect is achieved without engaging additional visible-band image or any other one. Furthermore, long-wave infrared data resolution enhancement itself is not a final stage of the resolution enhancement technique. Further refinement of the long-wave infrared image or the high-level derived products could be reached through the land surface emissivity estimated using data with higher spatial resolution. In addition, a variety of knowledge-based techniques, including artificial intelligence, can be applied to enhance the infrared imagery sharpness [27].

Spatial resolution enhancement techniques for satellite remote sensing data remain relevant even during technical progress, which leads to scanning systems improvements. Firstly, freely available remote sensing data, like Landsat imagery, have incomparable low resolution, compared to data provided by commercial satellites. Secondary, there is a big amount of low-resolution remote sensing data accumulated during the past decades, which is quite useful for lots of research, like long-term data time series processing.

Future work should be focused on the improvement of the MTFs theoretical models, on integrating the proposed approach with the rest ones applicable for the infrared imagery resolution enhancement, as well as on optimizing the inverse filtering computational procedures.

VI. ACKNOWLEDGEMENT

The results of the presented research are published with the support of The Faculty of Management Science and Informatics of the University of Žilina.

REFERENCES

[1] A. Das. Guide to Signals and Patterns in Image Processing. Foundations, Methods and Applications, Cham, Springer, 2015.

[2] F. Cutu, J. Eilertsen, "Super-resolution image reconstruction using high-frequency band extraction," US Patent 10147167B2, December 4, 2018.

[3] E. J. Knight, G. Kvaran, "Landsat-8 Operational Land Imager Design, Characterization and Performance," Remote Sensing, vol. 6, no. 11, pp. 10286-10305, October 2014.

[4] R. Zhang, X. Zhang, Z. Gong, S. Luo, "Fusion image quality assessment based on modulation transfer function," International Symposium on Image and Data Fusion (ISIDF 2011), Yunnan, IEEE, pp. 132-137, August 2011.

[5] M.D. Robinson, C.A. Toth, J.Y. Lo, S. Farsiu, "Efficient Fourier-wavelet super-resolution," IEEE Transactions on Image Processing., vol. 19, no. 10, pp. 2669-2681, October 2010.

[6] W. Witwit, Y. Zhao, K. Jenkins, Y. Zhao, "Satellite image resolution enhancement using discrete wavelet transform and new edge-directed interpolation," Journal of Electronic Imaging, vol. 26, no. 2, 023014, March-April 2017.

[7] H. Tang, Z.-L. Li, Quantitative Remote Sensing in Thermal Infrared: Theory and Applications, Berlin, Springer-Verlag, 2014.

[8] Z.-L. Li, H. Wu, N. Wang, S. Qiu, J. A. Sobrino, Z. Wan, B.-H. Tang, G. Yan, "Land surface emissivity retrieval from satellite data," International Journal of Remote Sensing, vol. 34, no. 9-10, pp. 3084-3127, October 2013.

[9] S. A. Stankevich, S. V. Shklyar, V. N. Podorvan, N. S. Lubskyi, "Thermal infrared imagery informativity enhancement using sub-pixel co-registration," International Conference on Information and Digital Technologies (IDT 2016), Rzeszów, IEEE, pp. 245-248, July 2016.

[10] S. Stankevich, E. Zaitseva, I. Piestova, P. Rusnak, J. Rabcan, "Satellite imagery spectral bands subpixel equalization based on ground classes' topology," International Conference on Information and Digital Technologies (IDT 2019), Žilina, IEEE, pp. 424-427, June 2019.

[11] S. A. Stankevich, I. O. Piestova, K. Y. Sukhanov, S. Cao, "Multispectral satellite imagery spatial resolution enhancement with reference spectra database," III Scientific Conference "Aerospace Technologies in Ukraine: Problems and Prospects", Kiev, NSFCTC, pp. 35-36, September 2019.

[12] S. A. Stankevich, I. O. Piestova, M. S. Lubskyi, "Remote sensing imagery spatial resolution enhancement," In: I.B. Abbasov (Ed.), Recognition and Perception of Images: Fundamentals and Applications, Beverly, Scrivener Publishing, pp. 327-368, February 2021.

[13] S. A. Stankevich, M. S. Lubskyi, A. Forgac, "Thermal infrared satellite imagery resolution enhancement with fuzzy logic bandpass filtering," International Conference on Information and Digital Technologies (IDT 2019), Žilina, IEEE, pp. 428-432, June 2019.

[14] S. Simon, "Modulation transfer function for optimum performance in vision systems," PhotonicsViews, vol. 16, no. 1, pp. 72-76, February-March 2019.

[15] R.L. Lagendijk, J. Biemond, "Basic methods for image restoration and identification," In: A.C. Bovik (Ed.), The Essential Guide to Image Processing, Burlington, Academic Press, 2009.

[16] D. Solimini, Understanding Earth Observation: The Electromagnetic Foundation of Remote Sensing, Cham, Springer Nature, 2016.

[17] S. N. Ahmed. Physics and Engineering of Radiation Detection, San Diego, Elsevier, 2015.

[18] T. L. Williams, The Optical Transfer Function of Imaging Systems, London, CRC Press, 1998

[19] J. Li, Z. Liu, "High-resolution dynamic inversion imaging with motion-aberrations-free using optical flow learning networks," Scientific Reports, vol. 9, 11319, August, 2019.

[20] R. A. Schowengerdt, Remote Sensing. Models and Methods for Image Processing, Burlington, Academic Press, 2007.

[21] W. G. Rees, Physical Principles of Remote Sensing, Cambridge, Cambridge University Press, 2012.

[22] G. D. Boreman, Modulation Transfer Function in Optical and Electro-Optical Systems, Bellingham, SPIE Press, 2001.

[23] J. Storey, M. Choate, K. Lee, "Landsat 8 operational land imager on-orbit geometric calibration and performance," Remote Sensing, vol. 6, no. 11, pp. 11127-11152, November 2014.

[24] B. N. Wenny, D. Helder, J. Hong, L Leigh, K. J. Thome, D. Reuter, "Pre- and post-launch spatial quality of the Landsat 8 thermal infrared sensor," Remote Sensing, vol.7, no. 2, pp. 1962-1980, February 2015.

[25] S. A. Stankevich, "Evaluation of optical transfer functions and restoring digital aerospace images by inverse filtering," Journal of Automation and Information Sciences, vol. 38, no. 6, pp. 39-47, June 2006.

[26] S. A. Stankevich, "Evaluation of the spatial resolution of digital aerospace image by the bidirectional point spread function parameterization," In: S. Shkarlet, A. Morozov, A. Palagin (Eds.), Advances in Intelligent Systems and Computing, vol. 1265, Cham: Springer Nature, pp. 317-327, September 2020

[27] S. Stankevich, N. Lubskyi, I. Piestova, A. Lysenko, "Knowledge-based multispectral remote sensing imagery superresolution," International Workshop on Reliability Engineering and Computational Intelligence (RECI 2020), Žilina, University of Žilina, p. 14, October 2020.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Advanced Genetic Algorithm for the Embedded FPGA Logic Diagnostic

Ekaterina Y. Danilova
Perm State National Research University
PSU
Perm, Russia
ket-eref@yandex.ru

Alexey Y. Gorodilov
Perm State National Research University
PSU
Perm, Russia
gora830@yandex.ru

Dmitry A. Kovylyaev
Perm State National Research University
PSU
Perm, Russia
354981@mail.ru

***Abstract*— FPGAs' logic is diagnosed to detect failures, finding a place of its occurrence, and determining its type. Diagnostic sequences are often used for diagnostics. They include several input sets. It is determined which failure occurred, according to the model of failures and the results of the function on the input sets.**

IEEE standards require systems on a chip (SoC) and FPGAs to contain built-in diagnostics. It is advisable to use genetic algorithms for this purpose. However, there is a problem of limited memory of the built-in diagnostic microcontroller that implements GA.

This article describes the construction of a diagnostic sequence using a genetic algorithm with intermediate coding of individuals. Estimates of the efficiency of the algorithm are given, as well as estimates of the required memory for its execution. The simulation of the GA execution on an embedded microcontroller with limited memory has been carried out.

Keywords*— *FPGA, single stuck-at failures, genetic algorithm, diagnosing, microcontroller, built-in diagnostics.

I. INTRODUCTION

Reliability [1] is one of the most important indicators of the quality of objects. To provide it, control and diagnostics are widely used.

Field-Programmable Gate Array (FPGA) is an electronic component used to create digital integrated circuits. Unlike regular digital microcircuits, the logic of FPGA operation is not determined during manufacture but is set through programming [2].

FPGA diagnostics is the process of detecting a failure, its type, and location [3]. FPGAs are widely used in various fields of human activity, while failures may occur regardless of the field of application. Therefore, a diagnostic is an important task, and many works are devoted to it [3-17].

The IEEE P1500 standard presupposes the development of FPGAs with built-in diagnostic tools. This obliges the developers of modern microcircuits both to embed additional elements into the circuit that allow diagnosing and to develop new algorithms designed to accelerate the diagnosis.

One of the ways to carry out diagnostics is to apply a special diagnostic sequence to the input. Based on the results of the execution of the diagnostic sequence and the failure model, it is possible to determine whether there is a failure, what kind of it, and where it occurred.

Genetic algorithms (GA) have proven themselves well in constructing diagnostic sequences [6,11,16]. But they, as a rule, require a significant amount of memory for storing the population, and therefore there is a problem of limited memory of the built-in diagnostic microcontroller that implements GA.

To solve this problem, a new GA with intermediate coding was developed. It significantly reduces the amount of memory required for storing GA data. The developed GA was used to find single constant failures.

A single stuck-at failure is a failure in which one of the inputs, instead of the right signal, turns to a stuck-at 0 or a stuck-at 1. If f_0 is the original function of n arguments then, in the event of a single stuck-at failure, one of the functions f_1 – f_{2n} can be obtained, where f_{2k-1} is a function obtained from f_0 by turning the input x_k to 0 and f_{2k} – by turning the input x_k to 1, $k = \overline{1..n}$. [17]

II. GA WITH INTERMEDIATE CODING

The failure model consists of $2n + 1$ functions, where n is the number of inputs. The principle of constructing functions of the failure model for a function of three arguments is presented in Table 1.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE I. EXAMPLE OF AUXILIARY FUNCTIONS FOR A FUNCTION OF 3 ARGUMENTS.

x_2	x_1	x_0	f_0	f_1	f_2	f_3	f_4	f_5	f_6
0	0	0	α_0	α_0	α_1	α_0	α_2	α_0	α_4
0	0	1	α_1	α_0	α_1	α_1	α_3	α_1	α_5
0	1	0	α_2	α_2	α_3	α_0	α_2	α_2	α_6
0	1	1	α_3	α_2	α_3	α_1	α_3	α_3	α_7
1	0	0	α_4	α_4	α_5	α_4	α_6	α_0	α_4
1	0	1	α_5	α_4	α_5	α_5	α_7	α_1	α_5
1	1	0	α_6	α_6	α_7	α_4	α_6	α_2	α_6
1	1	1	α_7	α_6	α_7	α_5	α_7	α_3	α_7

The diagnostic sequence is several input sets that will be carrying out to the FPGA inputs. Each set represents n zeros and ones. If we assume that a set is a binary notation of some number, it can be represented in decimal form. For example, a set from Table 1 **011** can be transformed like this:

$011_2 = 3_{10}$

Thus, to obtain the first level of GA coding, all input sets are taken and represented in decimal form. It turns out a sequence of non-negative integers from 0 to $2^n - 1$.

Such coding has already been considered in [16], but the efficiency of GA operation with such coding is not satisfactory. During crossing and mutation, individuals with duplicate sets may arise, which leads to the instant death of the resulting offspring. The crossing of individuals without intermediate coding is shown in Fig. 1.

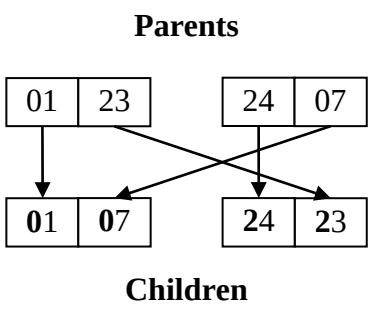


Figure 1. Crossing of individuals without intermediate coding.

To avoid the premature death of individuals, intermediate coding was introduced.

Let $p = 2^n$ be the number of input sets, where n is the number of function inputs; a_0, a_1, ..., a_l – coding of an individual; b_0, b_1, ..., b_l – intermediate coding of the individual; m_0, m_1, ..., m_{2^n-1} are auxiliary sequence M, where $m_k = k$, $k = \overline{0..2^n - 1}$.

The intermediate coding of an individual will be composed as follows: the next code element a_i is selected, its number (start from 0) in the sequence M is found, and is written into the intermediate coding. The number a_i is deleted from the sequence M. This continues until all the elements of a have been traversed.

The reverse transformation is carried out in the same way: a sequence M of numbers from 0 to 2^n–1 is written out. The next element of the intermediate coding is the number of the number in this sequence. The found number is written into the individual's code and deleted from the sequence.

Let there be a function of three arguments, and you need to encode the diagnostic sequence A = {4, 6, 1, 3}, respectively $a_0 = 4$, $a_1 = 6$, $a_2 = 1$, $a_3 = 3$. Fig. 2 shows the encoding process.

coding	sequence M		intermediate coding
4, 6, 1, 3	0, 1, 2, 3, **4**, 5, 6, 7	→4	
6, 1, 3	0, 1, 2, 3, 5, **6**, 7	→5	4
1, 3	0, **1**, 2, 3, 5, 7	→1	4, 5
3	0, 2, **3**, 5, 7	→2	4, 5, 1
	0, 2, 5, 7		4, 5, 1, 2

Figure 2 – Coding example.

As a result of coding, a set B = {4, 5, 1, 2} was obtained. For a given set, Fig. 3 shows the decoding process. As you can see, the result coincides with the set A.

intermediate coding	sequence M	coding
4, 5, 1, 2	0, 1, 2, 3, **4**, 5, 6, 7	
5, 1, 2	0, 1, 2, 3, 5, **6**, 7	4
1, 2	0, **1**, 2, 3, 5, 7	4, 6
2	0, 2, **3**, 5, 7	4, 6, 1
	0, 2, 5, 7	4, 6, 1, 3

Figure 3 – Decoding example.

The fitness function is implemented in such a way that the lower its value, the more adapted the individual is. Before calculating the fitness function, the same "slices" are calculated for the functions f_0, f_1, ..., f_{2n}. The "slice" is a sequence of bits of length l, obtained as a result of sequential input of the values of the diagnostic sequence of the individual for which the fitness function is searched. For counting purposes, the "slice" is stored as a decimal number. Let c_0, c_1, ..., c_h be the number of repetitions of each unique "slice", where h is the number of unique "slices". Then the value of the fitness function is calculated by the formula (1).

$$f_f = \sum_{i=0}^{h-1} 2^{c_i - 1} + (2 \cdot n + 1 - h) \quad (1)$$

For an optimal solution, all "slices" will be different, and the value of the fitness function will be equal to $2n + 1$ – the number of functions in the model.

The presented type of intermediate coding makes it possible to implement crossing and mutation in the classical form.

The algorithm used single-point crossing. The breakpoint can only pass between numbers. It was taken randomly in the range from 1 to $l - 2$, where l is the length of the diagnostic sequence. After that, there was a simple exchange of genes. An example of crossing is shown in Fig. 4.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

During the mutation, a random gene (integer) was selected, it was changed to a new value, which had to satisfy the condition of intermediate coding. The new value must be in the range from 0 to $2^n - i - 1$, where n is the number of function inputs, i is the number of the element to be changed, starting from zero.

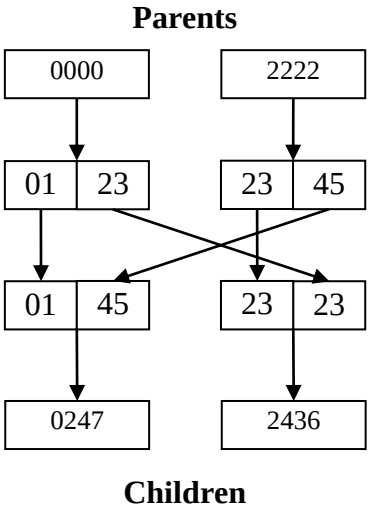


Figure 4. Crossing of individuals with intermediate coding.

Individuals for crossing and mutation are selected using the roulette method.

All individuals of the population and new individuals-children are stored in an array sorted by the value of the fitness function. It allows selection by simply discarding the "tail", i.e., all individuals whose numbers in the array exceed N – the size of the population.

In the classic GA, after the creation of the initial population in the cycle, the operations of crossing, mutation, and selection are applied. A diagram of a classical GA is shown in Fig. 5.

In the classical GA, the length of all individuals should be the same, and the construction of a diagnostic sequence of the minimum length assumes different lengths of individuals. To solve this problem, an outer loop is used that traverses all possible lengths of the diagnostic sequence until a solution is found.

The minimum length of the diagnostic sequence can be found by the formula (2).

$$\min = \lceil \log_2(2 \cdot n + 1) \rceil \quad (2)$$

The maximum length can be found by the formula (3)

$$\max = 2 \cdot n + 1 \quad (3)$$

The general scheme of the algorithm is shown in Fig. 6.

III. The Results of the GA

The presented genetic algorithm was first tested on a PC, and then it was emulated on a microcontroller.

Testing has shown that the algorithm works more efficiently than the algorithms presented in [16].

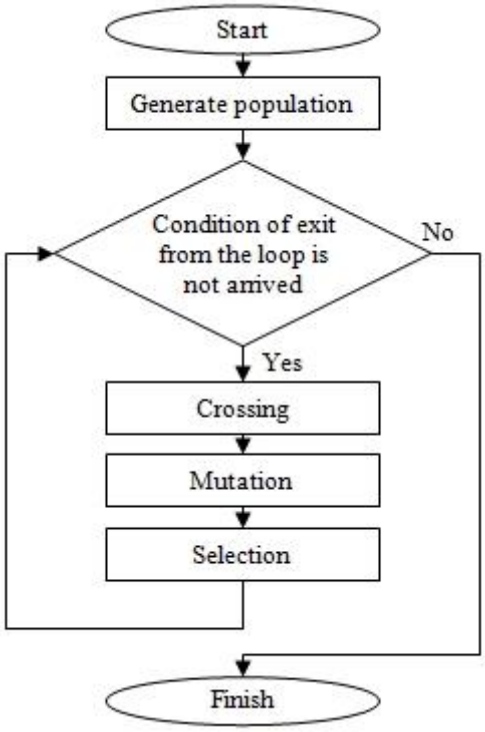


Figure 5. Classic genetic algorithm

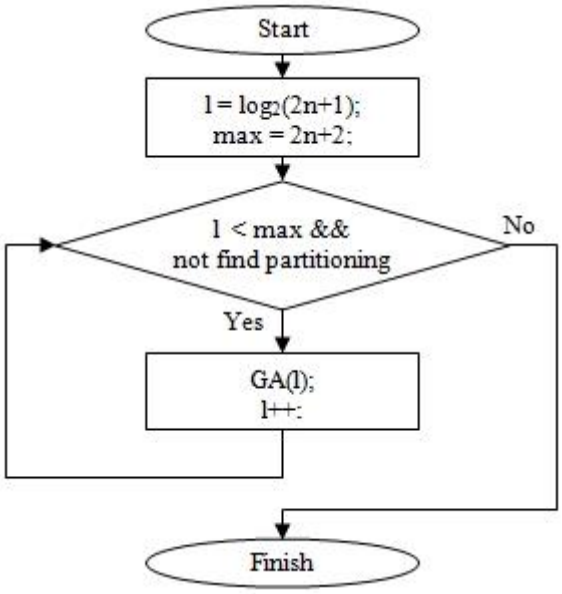


Figure 6. GA outer loop

As a result of testing, the characteristics of the GA, which are optimal for its operation, were identified. These characteristics depend on the number of inputs to the function. In formulas (4-9) n is the number of inputs.

Formula (4) for calculating the optimal population size:

$$N = 5\cdot(n-1) \quad (4)$$

Formula (5) for calculating the optimal number of crossing pairs:

$$S = 5\cdot(n-2) \quad (5)$$

Formula (6) for calculating the optimal number of mutating individuals:

$$M = 2\cdot(n-2) \quad (6)$$

Even though the work is devoted to optimizing the algorithm by memory, and the number of generations does not affect the amount of memory spent, this indicator is also important. Formula (7) for the optimal number of generations:

$$P = 50\cdot(n-1) \quad (7)$$

When the optimal parameters of the genetic algorithm are used, a diagnostic sequence of the minimum possible length is built in 60-100% of cases, in the remaining cases, the sequence has a length of 1 more than the minimum.

It is not difficult to estimate the amount of memory required to maintain a population of the described size (8). It should also be noted that the achieved value of the length of the diagnostic sequence may exceed the minimum by 1 or 2, and the answer obtained as a result of the GA may differ from the optimal one by 1 or 2 upwards. Thus, it can be assumed that even in the worst case, the obtained length of the diagnostic sequence will not exceed the minimum possible (2) by more than 4. The required memory is calculated in bytes.

$$\begin{aligned} m &= \left(2\cdot 2\cdot\left(\left\lceil \log_2(2n+1)\right\rceil + 4\right) + 2\cdot 4\right)\cdot \\ &\cdot\left(5\cdot(n-1) + 2\cdot 5\cdot(n-2) + 2\cdot(n-2)\right) + 4\cdot 4 = \\ &= \left(4\cdot\left\lceil \log_2(2n+1)\right\rceil + 24\right)\cdot(17n-49) + 16 \end{aligned} \quad (8)$$

Additional constants are memory spent on some entire fields of the classes "individual" and "population".

In addition, memory is spent on storing the failure model. By using packed bitsets, memory usage is minimized. The amount of memory used for storing the failure model can be calculated using the formula (9).

$$m2 = \left((n-3)+4\right)\cdot(2n+1) + 6\cdot 4 = 2n^2 + 3n + 25 \quad (9)$$

This algorithm is more efficient than previously developed algorithms in terms of memory. The algorithms presented in [16] had the following optimal characteristics and were tested only on functions with the number of arguments not exceeding 6:

$N = 50$;

$S = 20$;

$M = 10$;

$P = 200$.

IV. Emulation in Proteus

The idea of testing the operability of the genetic algorithm on a microcontroller is as follows: compiling the executable code for the microcontroller, using any possibilities for displaying the work results by the microcontroller, and creating an emulated circuit in Proteus. The STM32F401RE microcontroller was chosen for emulation. STM32F401RE can work with USART protocol. Thus, in Proteus, a circuit is created in which the microcontroller is connected to a receiver, which outputs messages sent by the microcontroller to the terminal. The resulting circuit is shown in Fig. 7.

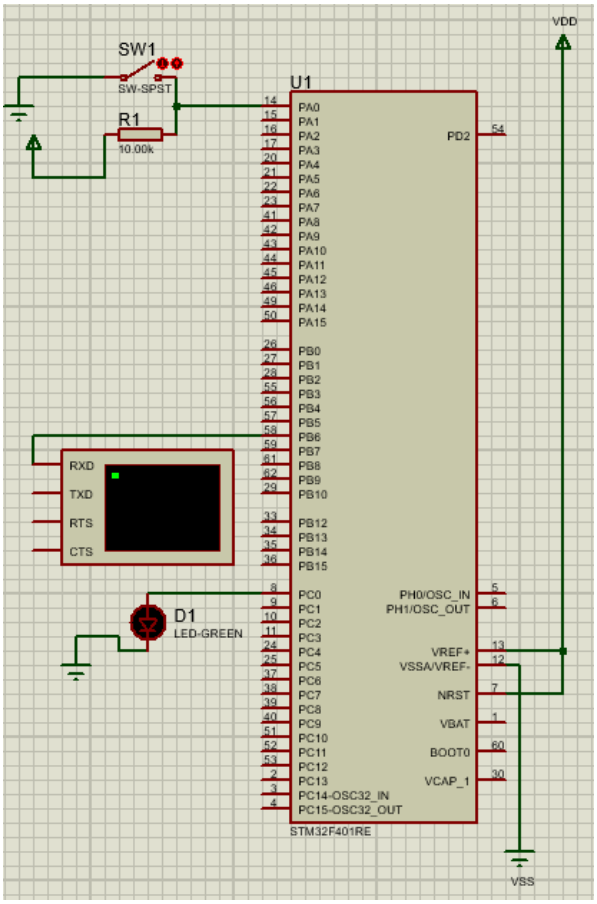


Figure 7. Scheme for running the GA.

The RXD pin of the terminal is receiving a signal. The TXD pin can send a signal from the console during simulation.

USART (Universal Synchronous-Asynchronous Receiver / Transmitter) is a universal synchronous-asynchronous transceiver, an interface for transferring data between digital devices [18].

In addition to the actions taken, in the STM32CubeMX project, you must specify the minimum stack size of 8192 bytes and the minimum heap size of 86016 bytes, which borders on the maximum 96 KB of memory.

Fig. 8 shows two iterations of the algorithm executed in the emulator on a microcontroller for the function $F(x_1, x_2, x_3) = (1100\ 1011)$.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

As you can see from the results, the launched program successfully found the optimal set of lengths equal to 4.

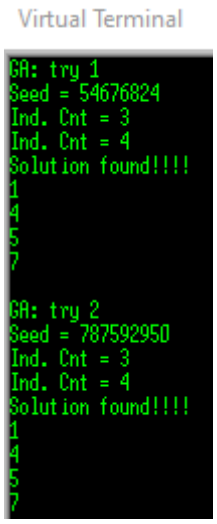


Figure 8. Emulation results

V. CONCLUSIONS

The article presents the new genetic algorithm with intermediate coding, the efficiency of which is higher than that of previously developed algorithms for finding single constant failures. The estimates of the required amount of memory for storing the GA population and the failure model are given. The optimal characteristics for launching the developed GA are found.

The work of the developed GA was emulated on the STM32F401RE microcontroller in the Proteus environment. Emulation showed that it is possible to find the optimal diagnostic sequence directly on the FPGA microcontroller. At the same time, developing a program for a microcontroller requires more attention to memory.

REFERENCES

[1] State Standard 27.002–2015. Nadezhnost' v tekhnike Osnovnye ponyatiya. Terminy i opredeleniya [Tekst]. – Vved. 2017–03–01. – M.: Standatinform, 2016. – 23 p.

[2] C. Carmichael, Triple Module Redundancy Design Techniques for Virtex FPGAs, Available at: https://www.xilinx.com/support/documentation/application_notes/xapp197.pdf (accessed: 03.05.2019).

[3] A.Y. Gorodilov Genetic algorithm of digital devices diagnosis // Vestnik PNRPU Electrical engineering, information technology, control systems, 2013. №7. [Electronic resource] Access mode: https://cyberleninka.ru/article/v/geneticheskiy-algoritm-diagnostirovaniya-tsifrovyh-ustroystv, free.

[4] A.Y. Gorodilov, E.Y. Danilova Built-In Diagnosis of The FPGA Logic Element // International Journal of Mechanical Engineering and Technology, 9(7), 2018, pp. 1347–1357. [Electronic resource] Access mode: http://www.iaeme.com/ijmet/issues.asp?JType=IJMET&VType=9&IType=7

[5] S.F. Tyurin, A.Y. Gorodilov, E.Y Danilova Diagnosing the logic element DC LUT FPGA // Inženernyj vestnik Dona (Rus) [Electronic resource] Access mode: http://www.ivdon.ru/uploads/article/pdf/IVD_19_Tyurin.pdf_2313.pdf, free.

[6] A. Gorodilov, Automatic synthesis of combinational circuits set for the purposes of FPGA reconfiguration within the model of partial failures of logic elements // Proceedings of the 2015 IEEE North West Russia Section Young Researchers in Electrical and Electronic Engineering Conference (2015 ElConRusNW). – IEEE, 2015. – Pp. 196-197.

[7] S.F. Tyurin, S.V. Ermakov, A.Y. Gorodilov External fault models comparison of FPGA PLIC logic gate multiplexer with respect to covering test sets, Vestnik PNRPU Electrical engineering, information technology, control systems, 2013. №7. [Electronic resource] Access mode: https://cyberleninka.ru/article/v/sravnenie-modeley-vneshnih-otkazov-elementarnogomultipleksora-logicheskogo-elementa-plis-fpga-otnositelno-pokryvayuschih-testovyh, free

[8] S.F. Tyurin, O.A. Gromov A residual basis search algorithm of fault-tolerant programmable logic integrated circuitsRussian Electrical Engineering- DOI: 10.3103/S1068371213110163

[9] S.F. Tyurin, LUT's Sliding Backup. IEEE transactions on device and materials reliability. Volume: 19 Issue: 1 Pages: 221-225 Published: MAR 2019.DOI: 10.1109/TDMR.2019.2898724

[10] Y.A. Skobtsov, V.Y. Skobtsov Logical modeling and testing of digital devices - Donetsk, 2005 - 436 p.

[11] K.V. Nikiforova, E.Y. Danilova Building a genetic algorithm with feedback for diagnosing FPGA // Mathematics and interdisciplinary research - 2018, 2018. - pp. 76-79

[12] O.Drozd, I.Perebeinos, O.Martynyuk, K.Zashcholkin, O.Ivanova, M.Drozd: Hidden fault analysis of FPGA projects for critical applications // In: IEEE International Conference TCSET. Paper 142, Lviv-Slavsko, Ukraine, (2020) doi: 10.1109/TCSET49122.2020.235591

[13] J. Drozd, A. Drozd, S. Antoshchuk, A. Kushnerov, V. Nikul Effectiveness of Matrix and Pipeline FPGA-Based Arithmetic Components of Safety-Related Systems // The 8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Warsaw, Poland, 2015, pp. 785–789. DOI: 10.1109/IDAACS.2015.7341410

[14] O. Drozd, V. Antoniuk, V. Nikul, M. Drozd, Hidden faults in FPGA-built digital components of safety-related systems // Proc. of the 14th International Conference “TCSET’2018 Conference “Modern problems of radio engineering, telecommunications and computer science”, Lviv-Slavsko, Ukraine, 2018, pp. 805-809. DOI: 10.1109/TCSET.2018.8336320

[15] S. F. Tyurin Hyper redundancy for super reliable FPGAs // Radioelectronic and computer systems No 1 (2021) PP.119-132, Available at: http://nti.khai.edu/ojs/index.php/reks/article/view/reks.2021.1.11/1449

[16] E.Yu Danilova, “Comparison of the genetic algorithms to build a diagnostic tree for diagnosing single stuck-at failures,” Proceeding of The Intenational Conference on “Information and Digital Technologies 2019”. Zilina, 2019. – pp. 93-97

[17] S.F. Tyurin, S.V. Ermakov, A.Y. Gorodilov External fault models comparison of FPGA PLIC logic gate multiplexer with respect to covering test sets, Vestnik PNRPU Electrical engineering, information technology, control systems, 2013. №7. [Electronic resource] Access mode: https://cyberleninka.ru/article/v/sravnenie-modeley-vneshnih-otkazov-elementarnogomultipleksora-logicheskogo-elementa-plis-fpga-otnositelno-pokryvayuschih-testovyh, free

[18] USART stm32 HAL, Available at: https://istarik.ru/blog/stm32/120.html.

Requirements Gathering for Specialized Information Systems in Public Administration

Stanislava Simonova
Institute of System Engineering and Informatics
University of Pardubice
Pardubice, Czech Republic
Stanislava.Simonova@upce.cz

***Abstract*— Identification of quality requirements is a necessary prerequisite for the development of a functional information system, collection, and analysis of data requirements and data functionalities are the initial steps in the general process of information system development. The quality of the process of collecting and evaluating requirements is a fundamental prerequisite for the created information system to accurately support the work needs of users. Users are therefore interested in participating in this requirements collection process. However, the situation is different for information systems in public administration, the aim of which is regulation and restriction. Here, the reluctance of users to cooperate and pass on the information needed for further development of the system can be assumed. This can then be reflected in the malfunction of the information system in the sense of incorrect support in the enforcement of regulation or restriction by the public administration. The article deals with the process of collecting and evaluating requirements for the restriction targeted information systems.**

Keywords— information system; voice of user; requirements gathering; support tools

I. INTRODUCTION

An information system is composed of people/users, hardware, software, database(s), application programs, and business processes. An information system is designed to facilitate the transformation of data into information and to manage both data and information [1]. Company management perceives information systems and information services generally as a necessary part of the business processes and expects their continuous performance. The degree of requirements on information environment in an organization is directly proportional to the economic and operational needs of the company; at the same time, it is determined by progress and abilities of information and communication technologies. The main goals are [2]:

- to ensure high functionality of information system; this means not only functions of keeping records and transactions but also analytic, functions for decision support and control functions;
- to achieve a high rate of application and technological availability, i.e. security, accessibility, reliability, and flexibility;
- to monitor continuously minimization of the cost compared to economic and non-economic effects.

Information systems in public administration have a specific position. The introduction of information systems into public administration activities in the Czech Republic is defined in Act No. 365/2000 Coll., (Act on information systems of public administration and amending certain other acts, as amended) [3] [4]. The Information Strategy of the Czech Republic focuses on digitization in the area of the exercise of official authority at the national level. It sets out the main objectives concerning the building of public administration information systems and also sets out general principles of the administration and operation of public administration information systems. The information system in public administration is determined to serve as support for ensuring the performance of public administration [5]. The information system in public administration is to some extent specific because it participates in the fulfillment of obligations arising from the powers of public administration bodies and at the same time stores data and information on the fulfillment of these obligations.

II. FORMULATION OF THE PROBLEM

The development of each information system takes place by the defined requirements, resp. development significantly depends on how comprehensively and correctly the requirements have been defined.

A requirement is any assumption or capability of a system that can help a participant (system user) solve a problem or achieve a goal. At the same time, the requirement is the assumption or capability of the system that the system must meet to achieve the required standard, security, certification, or legal and contractual conditions. Furthermore, all requirements must be documented.

Different types of requirements information can be distinguished, such as [6]:

- Business requirement: A high-level business objective of the organization that builds a product or of a customer who procures it.
- Business rule: A policy, guideline, standard, or regulation that defines or constrains some aspect of the

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

business. Not a software requirement in itself, but the origin of several types of software requirements.

- Constraint: A restriction that is imposed on the choices available to the developer for the design and construction of a product.
- External interface requirement: A description of a connection between a software system and a user, another software system, or a hardware device.
- Feature: One or more logically related system capabilities that provide value to a user and are described by a set of functional requirements.
- Functional requirement: A description of behavior that a system will exhibit under specific conditions.
- Non-functional requirement: A description of a property or characteristic that a system must exhibit or a constraint that it must respect.
- Quality attribute: A kind of non-functional requirement that describes a service or performance characteristic of a product.
- System requirement: A top-level requirement for a product that contains multiple subsystems, which could be all software or software and hardware.
- User requirement: A goal or task that specific classes of users must be able to perform with a system, or the desired product attribute.

Software requirements include three distinct levels: business requirements, user requirements, and functional requirements. In addition, every system has an assortment of nonfunctional requirements. The model in Figure 1 illustrates a way to think about these diverse types of requirements [6].

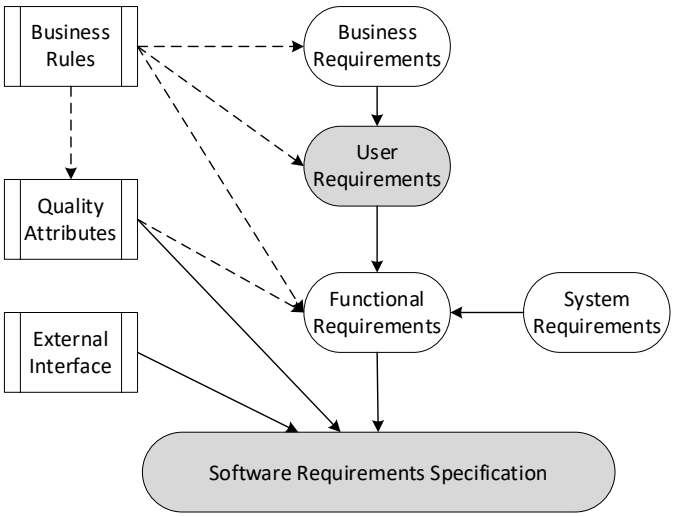


Figure 1. Relationships among several types of requirements information; source: own, prepared based on [6]

Identification of user requirements is the main concern of this text. The users' requirements should be essential for information system development. This is a basic assumption and there are recommendations for getting the users' requirements. However, the crucial questions are - Are the identified requirements really major?; Did the user identify really the essential requirements? Didn't the user forget about some important requirements? Thus, the relevance of the acquired user requirement set can be determined and checked?

Identification of requirements for the information system forms an important process in the development of the information system. The cooperation of the users of the emerging system in this process is necessary and crucial, and directly affects the quality and functionality of the system in terms of supporting the work of employees. The identification of requirements is generally considered to be one of the weakest points in the development of an information system. Methods and procedures for determining requirements are theoretically recommended and practically used, future users of the emerging information system are interested in cooperating in requirements identification, however, verification of the correctness of the requirements often only becomes apparent after the system has been completed, when users can verify whether the information system contains the essential requirements that they need to support them in carrying out their work activities. This is not a rare situation, which this situation occurs even though users are interested in working together to identify requirements.

It is necessary to take into account the situation that it is a special type of information system, respectively determined for a special type of user who is not interested in cooperating and even who tends to conceal functionalities. However, even here we need to get a complete set of system requirements. This type of system and users is described in Chapter IV. Therefore, our interest is divided into two sentences, firstly how to generally obtain a complete set of requirements, and secondly how to approach the analysis of special systems.

III. USER REQUIREMENTS – VOICE OF USER

User requirements generally relate to a product, whether the product is a product or service or an information system. User requirements describe goals or tasks the users must be able to perform with the product/system that will provide value to someone [7]. The domain of user requirements also includes descriptions of product attributes or characteristics that are important to user satisfaction.

A. Weaknesses in Identifying User Requirements

User requirements characterize what the users need to do with the system from their working point of view. However, it is not possible to assume and rely on users to determine a comprehensive and correct set of information system requirements.

Typical situations when identifying and analyzing requirements are [6]:

- The project's business objectives, vision, and scope were never clearly defined.
- Users were too busy to spend time working with analysts or developers on the requirements.

- The team could not interact directly with representative users to understand their needs.
- Users claimed that all requirements were critical, so they didn't prioritize them.
- Developers encountered ambiguities and missing information when coding, so they had to guess.
- Communications between developers and stakeholders focused on user interface displays or features, not on what users needed to accomplish with the software.
- The users never approved the requirements, or the customers approved the requirements for a release or iteration and then changed them continually.
- The project scope increased as requirements changes were accepted, but the schedule slipped because no additional resources were provided and no functionality was removed.
- Requested requirements changes got lost; no one knew the status of a particular change request.
- Users requested certain functionality and developers built it, but no one ever uses it.
- At the end of the project, the specification was satisfied but the customer or the business objectives were not.

Ways to represent user requirements include use cases [8], user stories, and event-response tables [9] [10]. However, the information system is a product intended for its customer, i.e. for the user. Therefore, it is suitable to use support tools, as various business methods use the Voice of Customers modeling, i.e. in this case the Voice of Users modeling.

B. Methods for Voice of Customers Identification

The Voice of the customer can be obtained and identify in a variety of ways: interviews, surveys, customer specifications, observation, warranty data, field reports, etc. [11]. All business process methods recommend the use of support tools for identifying and describing the voice of the customer. These include such as Cause and effect diagram, Kano model, 5xWhy, Critical to Quality method, Flowchart, Mental map, etc. Key quality criteria are the measurable characteristics of a process or service from the customer's point of view. When using the Critical to Quality method, it transforms the voice of the customer into these critical characteristics. These characteristics are already applicable since they mostly define specific limits of the information system [12].

CTQ tree is a graphical tool for transforming a user's voice to critical values where user needs are gradually decomposed into individual measurable parameters (see Figure 2) [13]. Causes and Critical to Quality indicators are analyzed with the help of the Cause and Effect diagram, definition and effects have to capture the root cause.

The Kano model provides a useful method how to evaluate whether the set of requests is correct, complete, without mistakes. The requests can deal with a product or an information system. Traditional ideas about quality have often assumed that customer satisfaction was simply proportional to how functional the product or service was. This would mean that an information system with more fulfilled functionalities is more satisfying for users, and an information system with fewer implemented functionalities less satisfies users. But this is not true, because it depends on which functionalities are implemented and also which functionalities in the IT service are missing. The Kano Model offers a way of understanding and categorizing the types of Customer/Users Requirements (or potential features) for new products/information services [14]. Requirements are evaluated according to categories and this expresses their importance or, conversely, insignificance.

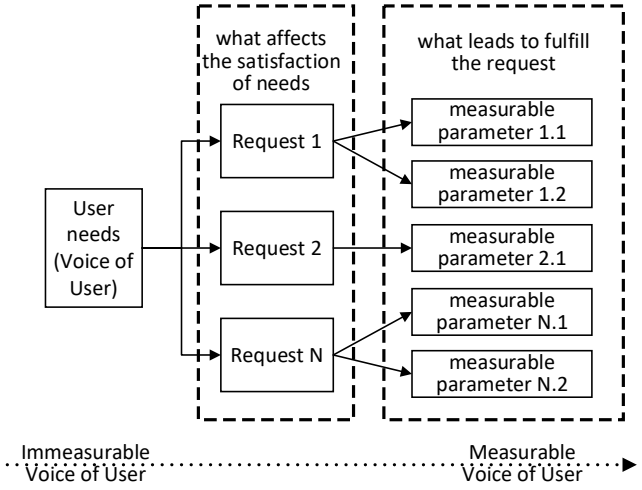


Figure 2. CTQ tree; source: own

The Must-be (basic) requirements are the most important, the user considers them as a matter of course requirements, however, are usually not explicitly demanded by the user. The One-dimensional (standard) requirements are important, are usually explicitly demanded by the user. The Indifferent or Reversal requirements are useless and the user doesn't even need them. (see Figure 3).

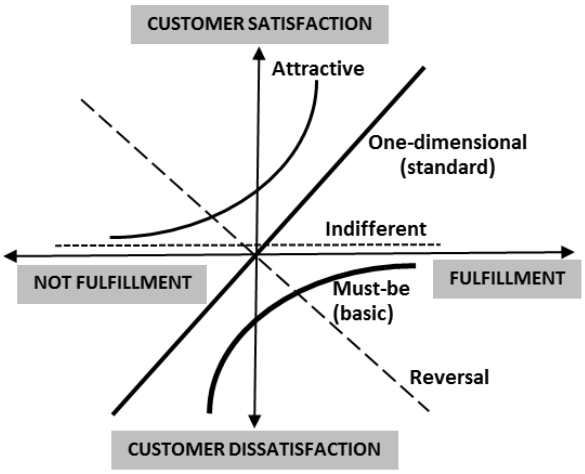


Figure 3. Categorization of the requirements by the Kano model; source: own, prepared based on [15]

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

C. *Evaluation of Identified Requirements*

Evaluation of requirements, their prioritization, and finding out the significance of requirements is the content of other auxiliary methods, which include, for example, CRUD and MoSCoW methods. The MoSCoW method deals with the scheme of four possible priority classifications for the requirements in a set [16]. The Must-requirement must be satisfied for the solution to be considered a success. The Should-requirement is important and should be included in the solution if possible, but it's not mandatory to succeed. The Could-requirement represents a desirable capability, but it could be deferred or eliminated, in other words, it is implemented only if time and resources permit. The Won't-requirement indicates a requirement that will not be implemented at this time but could be included in a future release.

Finding missing requirements is an important next step in requirements analysis. Missing requirements constitute a common type of requirement defect. The following techniques can help detect previously undiscovered requirements [6]:

- Decompose high-level requirements into enough detail to reveal exactly what is being requested.
- Ensure that all user groups have provided input.
- Trace system requirements, user requirements, event-response lists, and business rules to their corresponding functional requirements to make sure that all the necessary functionality was derived.
- Check boundary values for missing requirements.
- Represent requirements information in more than one way.
- Create a checklist of common functional areas to consider for your projects.
- A data model can reveal missing functionality.

The previous section defined a general approach to obtaining and evaluating user requirements. It is assumed that the user defines his requirements for the system with the intention that he is interested in providing as many requirements as possible and as well characterized as possible. Support tools and methods then help to verify the properties of the requirements, i.e. correctness, completeness, feasibility, necessity, unambiguity, verifiability, etc.

IV. SPECIALIZED INFORMATION SYSTEMS OF PUBLIC ADMINISTRATION – RESTRICTIONS TARGETED SYSTEM

The situation is different or more complicated with specialized information systems in public administration.

Public administration manages and regulates the social system through its performance, both by creating support, creating legal norms, or by exercising control (collecting information). The control requires the existence of a controlled system and a control system that controls the function of the controlled system. Control and regulation of the system are possible only if the correct transfer of information between the elements of the system. A similar way of driving appears in the work environment [17].

Public administration, with the support of information systems, carries out controls and appropriate countermeasures (restrictions). As part of the performance of public administration with the use of the information system, a situation may arise in which a defensive function, such as lying or concealment, follows when social norms are violated or when any harmful situation arises [18].

Generally speaking, people consciously and unconsciously avoid management and restrictions because they do not want to be controlled. During control, it is necessary to ensure the acquisition of information about the actual state between the controlled and control element. If an information system is to be created for management and restriction, it is necessary to assume the reluctance of users to cooperate and the reluctance to pass on all information necessary for the further development of the system during the design and especially the management of system requirements (see Figure 4).

Especially in the phase of analysis and collection of requirements for this information, testing of requirements must be performed.

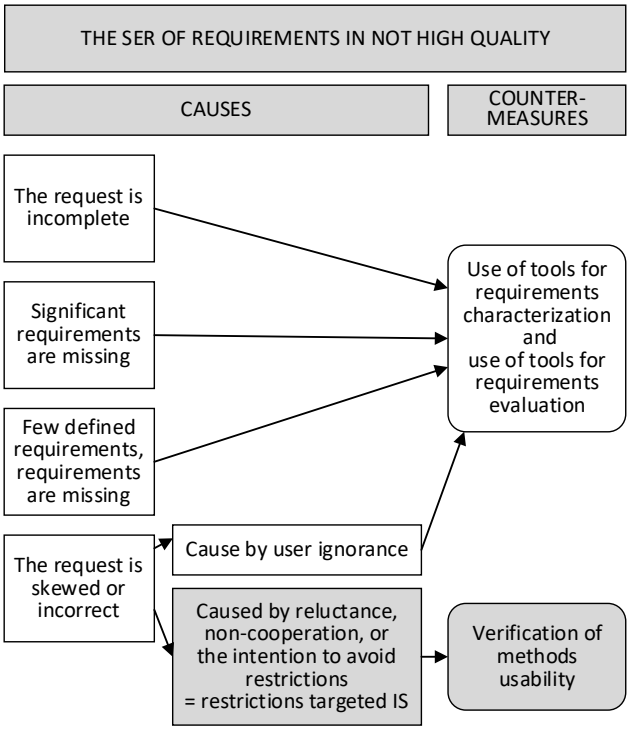


Figure 4. Causes of insufficient identification of requirements and countermeasures; source: own

A. *The restrictions targeted information system - requirements collection*

When creating a restriction targeted information system, one of the sources can be reports of misdemeanors and criminal offenses, which are registered by individual public

administration organizations. An evaluation of the number of such violations of legal regulations against the target group can determine appropriate activities and processes in public administration in which the new system solution will bring the greatest benefit.

As another source of requirements, it is possible to contact professional organizations or trade unions to protect workers in the field of interest. These organizations have knowledge and understanding of the activities in which the supervisory activities of public administration organizations are carried out [19].

Probably the most demanding in terms of time and cost are personal interviews with all interested participants in the project and the subsequent system, further group interviews, facilitation workshops, and the use of group creativity or questionnaires and surveys. If the project modifies an existing information system, it is possible to use observation of current processes or the record of reported problems from technical support [20].

When collecting requests within personal interviews, it is a matter of extracting as much information as possible relative to the future information system. The subject of interest is a specific information system, i.e. restriction targeted, the users of the system at the same time will be the subject to a restriction, therefore their primary interest is not to cooperate in the collection of requirements, or they directly want to avoid this restriction. The method of questioning must be in the form of a structured interview, such as interrogation in forensics.

The interrogation serves to establish the facts of the case without there being reasonable doubts about the established situation. A dialogue is conducted to identify erroneous information and to remind of missing facts. In case a different statement is found for more persons/users, it is possible to use a confrontation of these persons with each other during the interrogation [21].

Setting the main requirements for the system is possible to achieve using the Cause and effect diagram (see Figure 5), which shows the causes and consequences and aims to find the most likely cause of the problem.

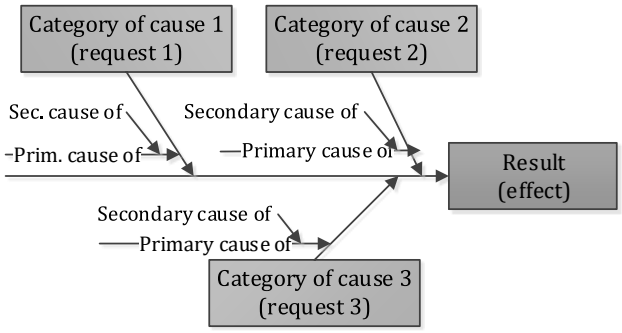


Figure 5. Causes of insufficient identification of requirements and countermeasures; source: own

B. The restrictions targeted information system - requirements verification

Requirements verification focuses on ensuring the properties of a set of requirements, such as completeness, accuracy, feasibility, necessity, unambiguity, and verifiability [22]. For these restriction targeted systems, considerable interest is focused on the completeness, respectively, incompleteness because:

- users of these types of systems will not be interested or willing to provide certain facts,
- most requirements are written in natural language, which does not have a fixed form to prevent incompleteness.

Requirements expressed by natural language sentences can be analyzed using the following phases:

- Syntactic analysis: nouns and verbs are distinguished in the sentences, which are further worked on in identifying classes and their relationships as follows - nouns forming the subject are identified as classes or attributes, verbs are identified as relationships between classes. The main purpose is to identify classes, their attributes, and relationships, respectively identification of classes without attributes, which means the request was incomplete (see Figure 6).

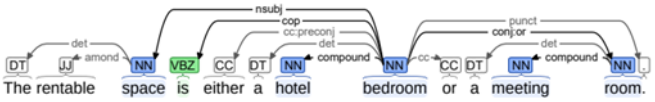


Figure 6. Example of word mapping in requests; source [23]

- UML class diagram is created, which shows the found objects in the sentences as separate classes. Attributes from the found attributes are assigned to individual classes, relationships among classes are created based on verbs and identified objects.
- Subsequently, it is possible to analyze the incompleteness by checking the created class diagram. The classes are checked whether the class has at least one attribute and whether each class has at least one link to another class.

The requirements are checked to the correctness, resp. error detection. The following methods can be used:

- Creation of Checklists: for each type of request (functional, non-functional, user, and business) a list of common errors is written. During the check, all requests are scanned and errors are searched according to the list. Checklists are created mainly within one organization and are updated over time with the current and most common errors.
- Requirements inspection: it is a process that looks for errors. Participants in the inspection consist of the authors of the requirements documentation, people who will create a system based on the identified requirements, and also people who will be responsible

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

for the outputs of the inspection and managing versions of requirements during changes. The inspection consists of several meetings, where all requirements are gradually reviewed. Requests are recited by one of the participants, and inspectors look for errors that are written down immediately. Some errors can be corrected immediately, but for some solutions, it will be necessary to contact some stakeholders again.

The procedure is planned to be used in the design of a specialized tool, namely a system for performing inspections of obligations for spirits and for mapping the distribution of spirits. The system is to be used to restrict persons handling alcohol, as well as to restrict control staff who will be the main users and will use the system for their activities.

V. CONCLUSION

Identification of requirements for the information system, their collection, and analysis, all form an important process in the development of the information system. The cooperation of the users of the emerging system in this process is necessary and crucial, and directly affects the quality and functionality of the system in terms of supporting the work of employees. The identification of requirements is generally considered to be one of the weakest points in the development of an information system. Methods and procedures for determining requirements are theoretically recommended and practically used, future users of the emerging information system are interested in cooperating in requirements identification, however, verification of the correctness of the requirements often only becomes apparent after the system has been completed, when users can verify whether the information system contains the really essential requirements that they need to support them in carrying out their work activities. This is not a rare situation, but again let us remind this situation occurs even though users are interested in working together to identify requirements.

The situation is more complicated with special information systems, where users may have reasons not to cooperate. More precisely, these are systems where there are two types of users, the state (checking person) and the checked person. These are the information systems applying the control obligation of the state, i.e. they are, for example, intended to restrict persons and at the same time to restrict control staff. In this text, they have been referred to like the restrictions targeted systems. Users of the developed system will be subject to restriction, so it is not their primary interest to cooperate in the comprehensive identification of requirements, or they are directly trying to avoid restriction. Again let us remind that a discrepancy in requirements may occur even if users are interested in cooperating in identifying the requirements. In this case, the main focus is on helping users to identify complete requirements. Here, the situation is complicated by the fact that the information system under construction is to serve two types of users, on the one hand, the state, on the other hand, the checked/restricted person. The second type of user is not interested in cooperating or it is even possible to encounter intentional concealment of facts. Therefore, this process is complex and requires the application of identification methods and procedures from other fields, so that the result is a complete list of the correct requirements for the planned information system.

REFERENCES

[1] C. Coronel, S. Morris, P. Rob, and K. Crockett. Database principles. Fundamentals of Design, Implementation, and Management. 2013. Centage Learning, 2013, p. 866.

[2] L. Gala, J. Pour, and Z. Sediva. Podniková informatika: počítačové aplikace v podnikové a mezipodnikové praxi. Praha: Grada Publishing. Management v informační společnosti, 2015, p. 240.

[3] Act No. 365/2000 Coll., Act on information systems of public administration and amending certain other acts, as amended. [online]. [cit. 2021-04-08]. Available from WWW <http://aplikace.mvcr.cz/ sbirka-zakonu/>.

[4] S. Simonova, and M.Y. Amare. "Aspects of Digital Documents Archiving by the Organizations in the Czech Republic in Context of the EU eGovernment". In the International Conference on Information and Digital Technologies (IDT), 2019. pp. 4865-4872.

[5] MPO, Ministry of Industry and Trade of the Czech Republic. Digital CZ. [online]. [cit. 2021-04-02] Available from <https://mpo.cz/en /business/digital-society/digital-czech-republic--243601/>.

[6] K. E. Wiegers, and J. Beatty. Software Requirements. Washington: Microsoft Press,U.S., 2013, p. 673.

[7] R. Sherman. Business Intelligence Guidebook. From Data Integration to Analytios. Elsevier, Morgan Kaufmann. 2020. p. 525.

[8] D. Kulak, and E. Guiney. Use Cases: Requirements in Context, Boston: Addison-Wesley. 2012, p. 272.

[9] M. Cohn. User Stories Applied: For Agile Software Development. Boston: Addison-Wesley. 2004, p. 304.

[10] C. N. Khaflic. Storytelling with data, a data visualization guide for business professionals. Wiley, 2015. p. 288.

[11] iSixSigma: Voice Of the Customer [online]. [cit. 2021-03-04]. Available from WWW <https://www.isixsigma.com/ dictionary/voice-of-the-customer-voc/>

[12] S. Simonova, and N. Foltanova. "Implementation of Quality Principles for IT service Requirements Analyse". In. Proceedings of the Conference on Information and Digital Technologies 2017, pp. 365-372.

[13] Critical to Quality (CTQ) Trees. Mind Tools [online]. [cit. 2017-03-22]. Available from WWW <https://www.mindtools.com/pages/article/ctq-trees.htm>

[14] What is the Kano Model? [online]. [cit. 2021-04-08]. Available from WWW:< https://www.kanomodel.com/>

[15] G. J. Goddard, G. Raab, R. A. Ajami and V. B. Gargeya. Customer Relationship Management: A Global Perspective. USA: Gower Publishing Company, 2012. p. 216.

[16] IIBA, International Institut of Business Analysis. A Guide to the Business Analysis Body of Knowledge (BABOK Guide). Toronto: International Institute of Business Analysis. 2015. p. 514.

[17] J. Koubek. Řízení lidských zdrojů: základy moderní personalistiky. Praha: Management Press. 2015. p. 400.

[18] Z. Vybiral. Lži a pravda v lidské komunikaci. Praha: Portál. 2015. p. 175.

[19] J. Rowley. "E-Government stakeholders—Who are they and what do they want?" in International Journal of Information Management, vol. 31, Issue 1, 2011, pp. 53-62.

[20] K. Schwalbe, Kathy. Information Technology Project Management. Cengage Learning. 2015. p. 672.

[21] J. Musil, Z. Konrad and J. Suchanek. Kriminalistika. Praha: C.H. Beck. Beckovy mezioborové učebnice. 2004. p. 583.

[22] V. Bóka. Návrh informačního systému s restriktivním přístupem. Pardubice: Univerzita Pardubice. 2021, in press. p. 59.

[23] D. Senkyr and P. Kroha. „Patterns in Textual Requirements Specification" in Proceedings of the 13th International Conference on Software Technologies, 2018, pp. 197-204.

Self-Timed FPGA Design Perspectives

Sofya K. Chikarenko
Department of Automation and Telemechanic
Perm National Research Polytechnic University
JSC "Perm Scientific-Industrial Instrument Making Company"
Perm, Russia
sophia-chikarenko@yandex.ru

Kseniya M. Ivanova
Department of Automation and Telemechanic
Perm National Research Polytechnic University
JSC "Perm Scientific-Industrial Instrument Making Company"
Perm, Russia
ivanovakseniyamich@yandex.ru

Alexandra Yu. Skornyakova
JSC "Perm Scientific-Industrial Instrument Making Company"
Perm, Russia
juris-plot@mail.ru

Sergey F. Tyurin
Department of Automation and Telemechanic
Perm National Research Polytechnic University
Department of Software Computing Systems
Perm State University
Perm, Russia
tyurinsergfeo@yandex.ru

Abstract — **Self-timed (ST) circuits (STC) proposed by D. Muller in the 50s of the twentieth century are the perspective direction of the digital technology, despite the known difficulties and problems. This direction is especially important in connection with the development of nanoelectronics, since quantum effects can lead to the impossibility of global synchronization. This paper discusses further ways of the developing ST devices oriented on FPGAs. The authors in their work focus on the design of the Self-Timed communication switch, ST adaptive logic gate LUT-DC LUT, and consider the possibility of expanding the applicability of previously developed optimization algorithms for the choosing optimal sets of ST logic gates.**

Keywords—FPGA; Self-Timed; Communication Switch; Layout Simulation; Adaptive Logic Gate LUT-DC LUT

I. INTRODUCTION

Now, the development of self-timed circuits [1 – 3] is still at the research stage. According to the design principle, self-timed circuits can be divided into two types. The first type is quasi-self-timed. Such circuits are characterized by the use of a combinational circuit model. The most common method for design such circuits is the maximum delay model method. In practice, such circuits do not have great advantages over their synchronous analogs. The second type is strictly self-timed circuits (SST). The principle of operation of such circuits is that the moment of the end of the transient process in each block of the circuit is determined. For the design of SST, the most optimal is the use of the paraphase coding method with a spacer. In the work of the SST there are the working and the spacer phases. In the working phase functional transformations take place, and in the spacer phase preparation for the next transformations takes place. Usually, easily distinguishable sets of signals are used as a spacer: either all 0 or all 1. To determine which phase the device is in at a given time, an indicator of the end of the transient is used. There are prototypes of ST devices that have been implemented [4 – 5]. They show advantages over synchronous countertypes [7]. The design of self-timed devices has been studied from various angles, for example, the problems of developing fault-tolerant and energy-efficient self-timed circuits were considered in [7 – 9]. However, these are all examples of STC ASIC implementation or STC ULA implementation. FPGA [10, 11] implementations of STC devices are important. There are examples of the development of ST FPGAs, for example [12, 13], however, analysis shows that they are not strictly ST. Advanced synchronic logic gates for FPGAs were proposed in [14, 15], and in [16, 17] there was investigated the combination of the self-timed approach and programmable logic. However, these studies were aimed only at the implementation of logic functions in various bases. To create self-timed programmable logic integrated circuits, you need not only blocks that calculate functions, but also various transfer circuits, switching arrays, etc.

II. CONNECTION SWITCHES

One of the next directions of work in the field of self-timed circuits FPGAs can be the development of self-timed signal switching arrays. There are two possible variants: a unitary switch (based on the unitary connection code, One Hot) and a switch based on the previously developed ST LUT (with the usual binary connection coding, minimal bit), which implements the repeating function. LUT (Look-Up Table) is the main logic element for design FPGA. It is a multiplexer in the form of a tree of transmitting transistors. It calculates the logical functions specified in the DNF (disjunctive normal form, in boolean logic it is a canonical normal form of a logical

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

formula consisting of a disjunction of conjunctions). Only one Boolean function can be evaluated on one LUT.

Fig. 1 shows the proposed unitary switch, which is a unitary n-1 multiplexer in fact, where the configuration (tuning of signal communication) written to the cells of the static random access memory SRAM determines one of *n* paraphase communications to the *q0*, *q1* outputs. The indication procedure is carried out by the usual way for ST circuits using of G triggers and elements AND on the inputs of the communications, however, due to the fact that the configuration settings select only one of n communications, both in the main and in the dual channels, the inputs of the signal communication blocks are not indicated, the analysis is carried out by outputs (indicator I).

Let us rate the number of transistors required for one channel of the signal communication selection tree by the unitary code, considering the cells of the configuration RAM SRAM (six-transistor cells):

$$L_1 = 2^{\lfloor n \rfloor} + 6n \qquad (1)$$

To implement a switch with minimal coding, it is proposed to use LUT in the main and dual communication channels (Fig.2).

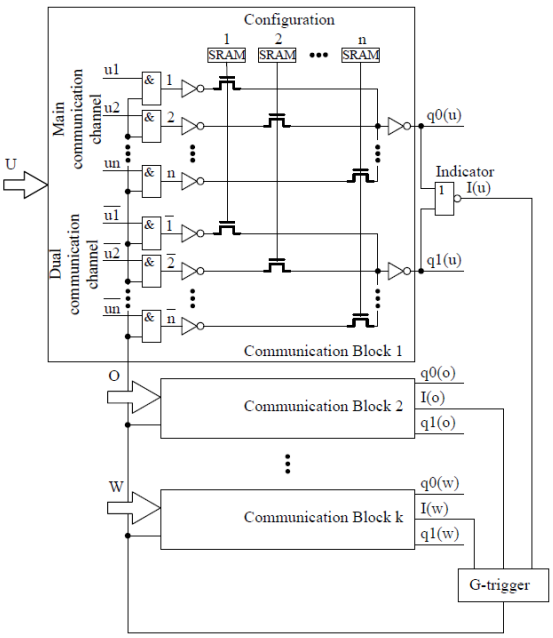


Figure 1. Proposed ST one hot communication switch

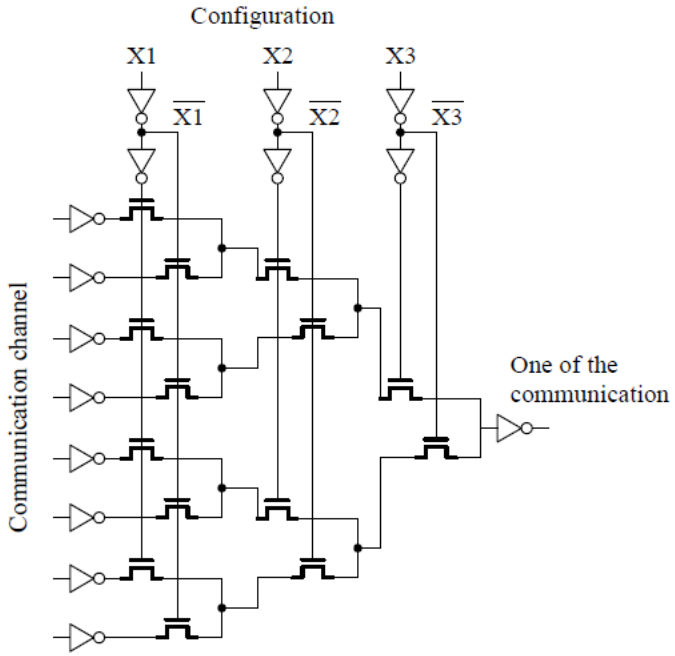


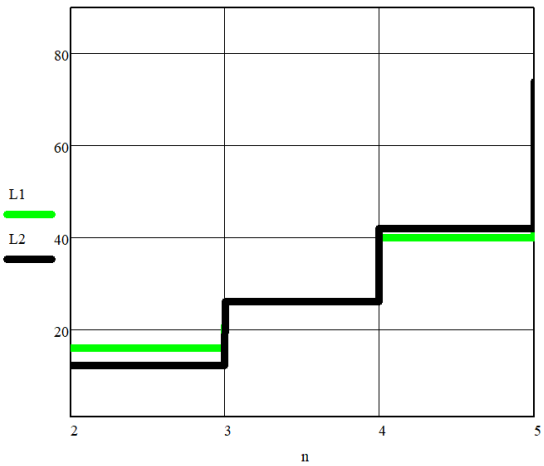
Figure 2. Switch based on the previously developed ST LUT (with normal communication binary coding, Minimal Bit), only the main channel

Configuration settings are applied to inputs *X*, one of *n* communications become selected. Therefore, the required number of transistors for one channel is determined by the expression:

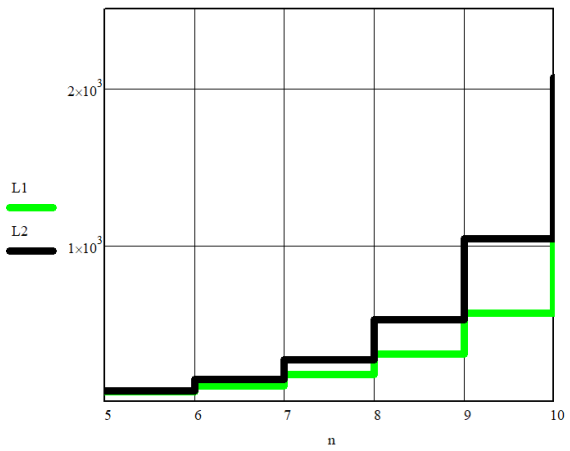
$$L_2 = 2^{\lfloor n \rfloor + 1} - 2 + 6\log_2(n). \qquad (2)$$

Comparison shows that the gain in complexity of the unitary decoder relative to the decoder with the minimum coding will appear for $n > 4$ (Fig. 3).

The gain will be much more, taking into account the cost of tree decomposition when formation a switch based on a LUT for a large number of inputs. However, this requires further research, including specific technologies for creating FPGAs switching arrays.



a)



b)

Figure 3. Complexity of the number of transistors as a function of number of inputs for a switch with unitary coding (L1) and a switch based on LUT (L2): a) for the range $n = 2 \dots 5$; b) for the range $n = 5 \dots 10$

Feedback switching *1-n* can also be carried out using a unitary decoder based on the developed DC LUT-ST gate, while configuration constants are supplied to the inputs of variables, and instead of constants to the inputs of the main and dual trees are supplied the main and dual communications. DC-LUT - decoder Look-Up Table - is a reverse tree of LUT transmitting transistors. It calculates the logical functions specified in the PDNF (perfect disjunctive normal form, a special case of a DNF function in which there are no elementary conjunctions, each term has no repeated variables and contains all the variables on which the Boolean function depends). On this element, you can calculate several logical functions at once. The setting for calculating functions takes place in separate setting blocks.

Let us consider the simulation of the communication switch, shown in Fig. 1, using CAD "Multisim" [19] without taking into account indicators and G-triggers. As an example, a communication block is used for $n = 2$. The spacer phase at this stage is implemented by a separate input G. If the G input has a logical zero, then this means that the communication block in the spacer phase, at the g0 and g1 outputs, is zero, regardless of the input variables. If G = 1, then the circuit is in the working phase state.

Fig. 4 shows the communication block in the spacer phase. Transistors Q1 - Q6 and elements U1A - U4A implement the main communication channel. Transistors Q7 - Q12 and elements U5A - U8A implement a dual communication channel. Keys X0 - X3 are used as communications. Memory cells are implemented on keys S1 - S2. Inverters U15A - U16A serve for the restoring signal levels.

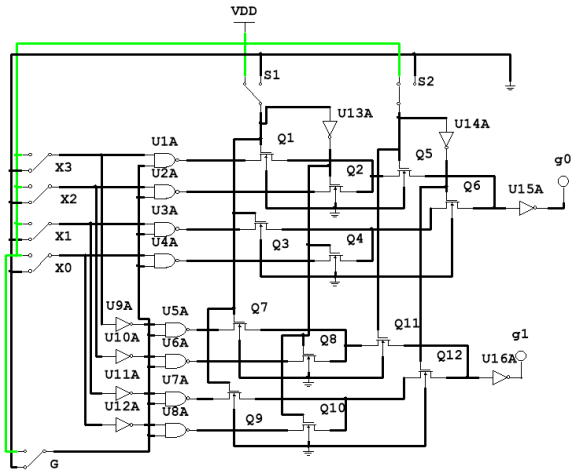


Figure 4. Communication block

Let S1 = 1, S2 = 0, transistors Q1, Q3, Q6, Q7, Q9 and Q12 are open. However, the signals to the g0 and g1 outputs will not pass through the open transistors Q1 and Q7, since Q5 and Q11 are closed. In this case, the essential communication is X1, which value of is transmitted through the open transistors Q3 and Q6, Q9 and Q12 to the outputs g0 and g1.

Fig. 5 shows that the circuit is working correctly, the main communication channel has a logical high, and the dual channel has zero. Fig. 6 illustrates the correct work of circuit when the main communication channel has a logical zero, and the dual channel has logical high.

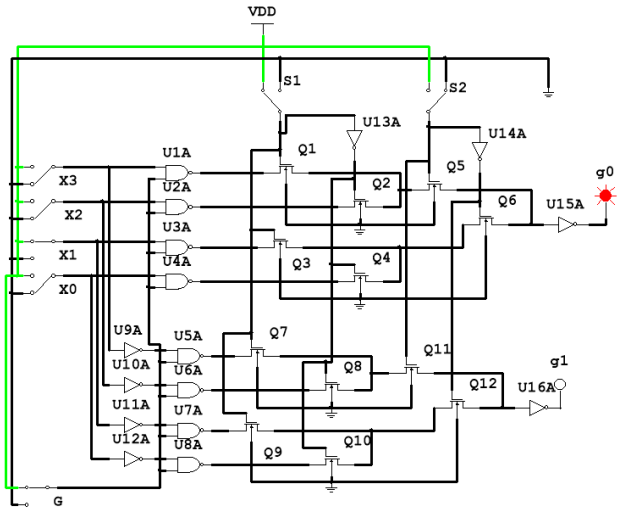


Figure 5. Working phase at S1 = 1, S2 = 0, X1 = 1

Other results of the work of the connections block are summarized in Table 1.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

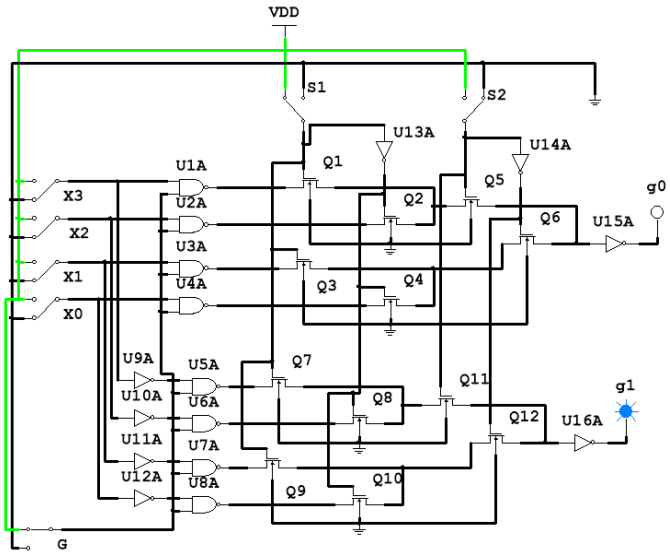


Figure 6. Working phase at S1 = 1, S2 = 0, X1 = 0

TABLE I. COMMUNICATION BLOCK TRUTH TABLE

S1	S2	X0	X1	X2	X3	G	g0	g1
0	0	0	~	~	~	1	0	1
0	0	1	~	~	~	1	1	0
0	1	~	~	0	~	1	0	1
0	1	~	~	1	~	1	1	0
1	0	~	0	~	~	1	0	1
1	0	~	1	~	~	1	1	0
1	1	~	~	~	0	1	0	1
1	1	~	~	~	1	1	1	0

To confirm the correct working of the communication block we simulated it at the layout level. For this, we use CAD "Microwind" [19] and design rules 90 nm. The timing diagrams are shown in Fig. 7.

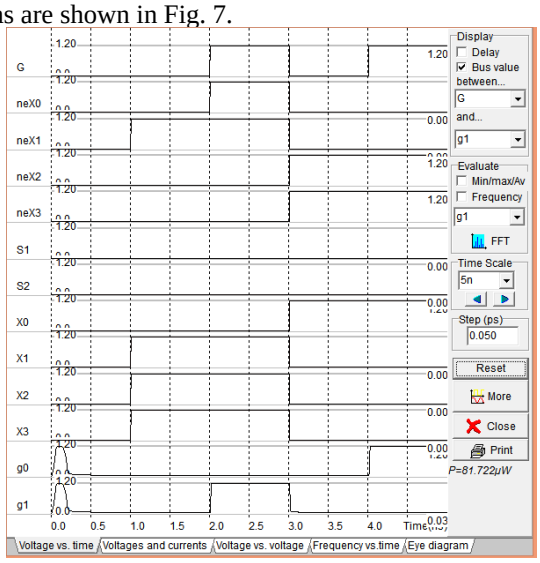


Figure 7. Timing diagrams S1=S2=0

Fig. 9 shows the result of simulation the set S1 = S2 = 0 with alternation of working and spacer phases. It can be seen that for G = 0 the outputs g0 = g1 = 0, for X0 = 0 - g1 = 1, and g0 = 0, for X0 = 1 - g1 = 0, and g0 = 1. The layout simulation is shown in Fig. 8.

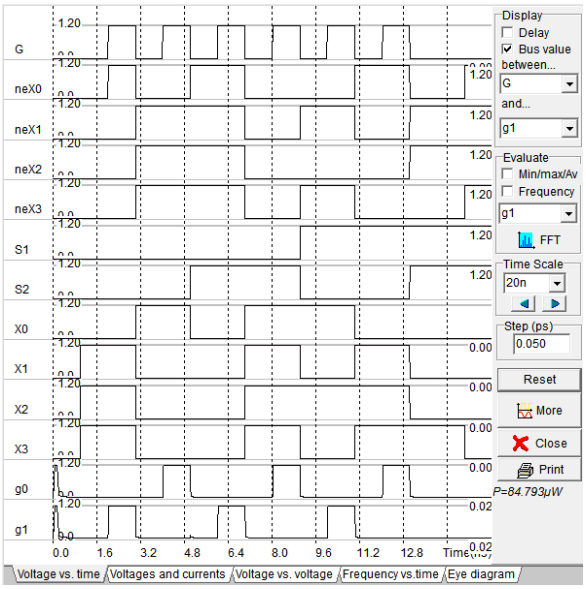


Figure 8. Timing diagrams with alternation of working and spacer phases

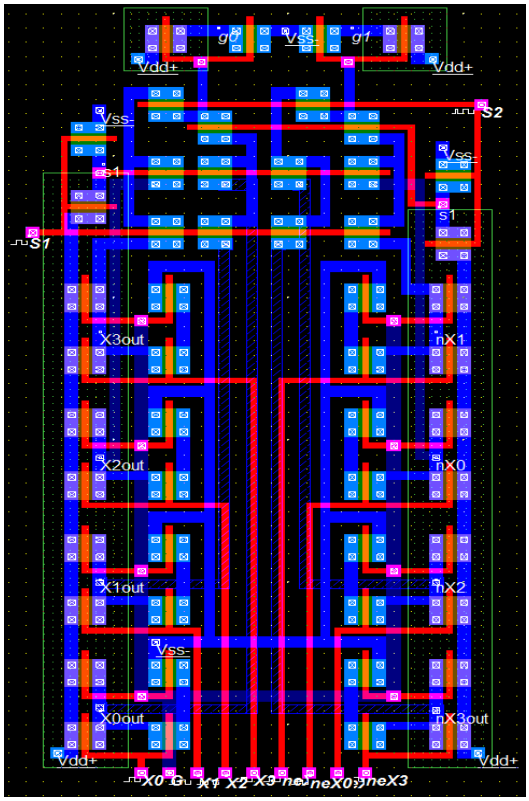


Figure 9. Layout simulation of communication block

The obtained result of layout simulation coincides to the truth table 1, which confirms the correct working of the connection block.

III. LOGIC GATES FOR UNITARY SETS OF VARIABLES

At present, the number of logic gates in advanced FPGAs reaches tens of millions with a total number of transistors of several billions. Analysis of FPGA development trends, for example, Intel's "hyperflex" and "hyperoptimization" technologies, allows us to conclude that further improvement of FPGAs is possible by reducing time delays in calculating logic functions by the sharp increasing the cost of logic elements, for example, by using the Shannon decomposition. That is why there is proposed the logic gate for the unitary sets of variables LUT-oh (Fig. 10).

The unitary code of a set of variables activates one of the *n* transistors and then the corresponding value of the logic function written in the configuration memory is fed to the output of the gate. In this case, the costs of switching such "long" sets of variables also increase. However, the delay in calculating the logic function is minimal. The decomposition issues of formation multi-bit elements from elements of lower bit depth and self-timed implementation of such elements require further research.

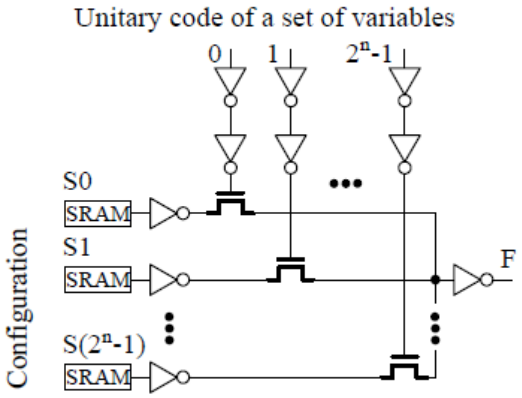


Figure 10. Logic gate for unitary sets of variables

IV. DEVELOPMENT OF ADAPTIVE LUT / DC LUT GATES AND THE USING OF FINFET TRANSISTORS

Authors propose to consider in further research the possibility of creating self-timed circuits based on an element with a higher level of configuration: it is configured to either LUT ST or DC-LUT ST, depending on the tasks. Such an element in a synchronous version was proposed in the Ph.D. dissertation by R.V Vikhorev [15,16] (Fig. 11). The layout model of such an adaptive logic element ADC-LUT for one variable with pull-up resistors at the DC outputs has been developed using CAD “Microwind” [20] (free version 3.1) (Fig. 12). The proposed element, depending on the tuning, can implement both LUT and DC-LUT. The tuning is performed by the signal of the configuration cell (CC): with CC = 1 the developed element implements the logical element LUT (Fig. 13), with CC=0 it implements DC-LUT (Fig. 14).

In the LUT mode when input signals *In0 LUT* = 0 and *In1 LUT* = 1, the output signal *OutLUT* is inverse to *x* (Fig. 13, a). When *In0 LUT* = 1 and *In1 LUT* = 0, it repeats *x*. In DC mode the element splits the signal into two outputs Out0 DC-LUT and Out1 DC-LUT.

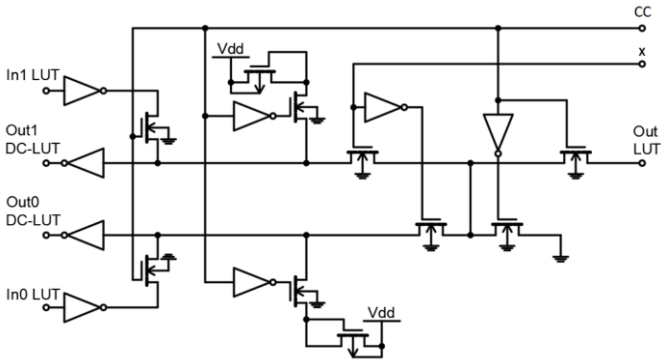


Figure 11. Adaptive logic gate ADC-LUT

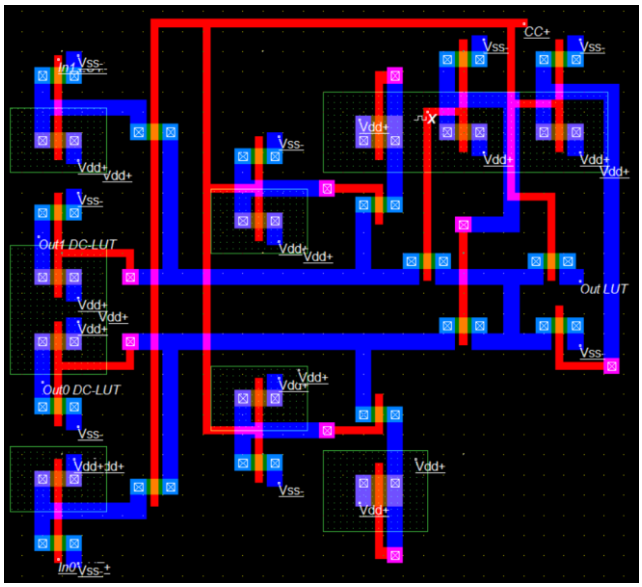
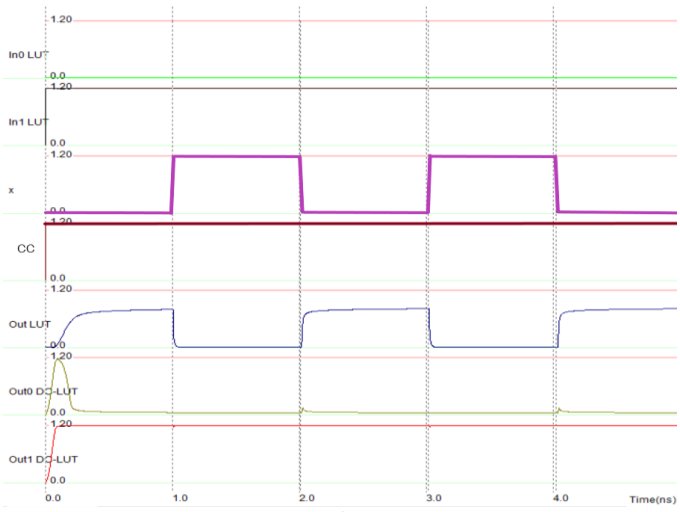
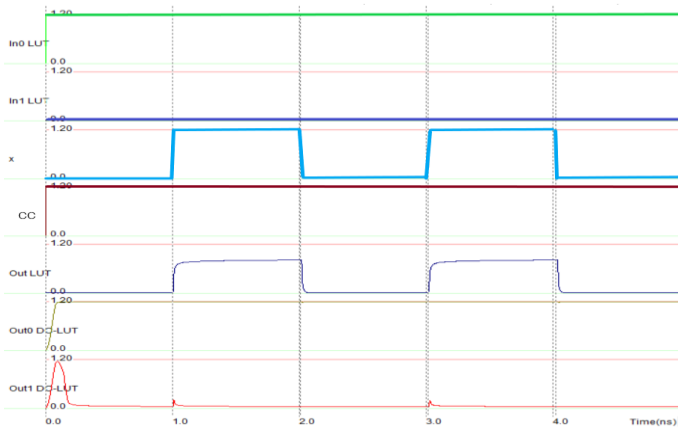


Figure 12. Layout model of ADC-LUT



a)

978-1-6654-3692-2/21/$31.00 ©2021 IEEE



b)

Figure 13. Simulation results of ADC-LUT in LUT mode:
a) inversion of x;
b) repeat of x

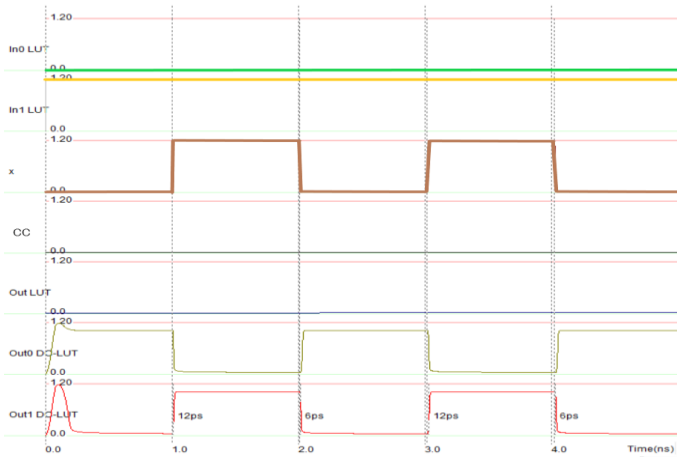


Figure 14. Simulation results of ADC-LUT in DC mode

The proposed model is implemented on the basis of standard MOS transistors. In the future, it is planned to develop a self-timed adaptive logic gate based on the FinFET technology that became possible in CAD "Microwind" version 3.8.5 [19]. FinFET transistors, due to their design features, have a number of indisputable advantages over traditional MOS transistors, namely: leakage currents are reduced multiple times, dynamic power consumption is reduced, and performance is significantly increased. Therefore, their application in the field of self-timed technology seems to be very promising.

V. Improvement of Algorithms for Choosing Optimal Sets of Logic Gates

It is planned to move to a more general formalization based on the developed algorithm for choosing the optimal set of configurable ST elements, which implements multicriteria optimization, forming a Pareto-optimal set. First, this is the possibility of entering a non-empty set, the elements of which are *any* logical elements, i.e., for example, using not only a self-timed approach, but also a synchronous one. A list of parameters for formalization will be offered for each element: implementation property of the element (how many logic functions and systems it implements); the number of conjunctions; types of systems of functions: DNF or PDNF; formulas for evaluation quality parameters. Secondly, at this time the number of admissible elements in the set is limited, in the future it is planned to exclude this parameter, and the number of elements in the configurable set will depend on the formalization of the element types. Thirdly, in addition to calculating complex optimization characteristics, such as: the number of transistors, the area occupied on the chip, speed and power, one more parameter will be added - reliability. For the reliability parameter, you will be able to select the type of reservation from the following options:

1.No Redundancy;
2. Triple Modular Redundancy (TMR);
3.Deep TMR;
4. Transistor's Quadding.

Using the required combination of optimization criteria and a reliability parameter, the designer can choose the configuration of the elements according to a given parameter, for example, according to the probability of no-failure operation, which will make it possible to obtain a more reliable system.

After the changes, the algorithm can be used in the design of various FPGAs for the layout of configurable logic blocks consisted of logic elements that can be formalized with the proposed parameters, to determine the optimal set taking into account various cardinal characteristics that can be selected depending on the need.

Conclusion

It is advisable to direct further research in the field of self-timed FPGAs to the development of communication switches for switching arrays. Perspective approaches are focused on the using of unitary sets of input variables and corresponding logic gates, the development of adaptive logic gates that can be configured for implementing both a LUT and DC-LUT, which is just required for unitary calculations.

Despite the increased complexity of ST elements, due to the unification of the implementation of systems of logical functions, there are new opportunities for using the self-timed approach in the field of nanoelectronics and quantum, reversible computations, where, according to some estimations, there are no alternatives to ST circuits. To solve the problems of import substitution, it makes sense to work with domestic developers and manufacturers of FPGAs in order to organize the relevant R&D.

References

[1] D. E. Muller, W.S. Bartky, "A theory of asynchronous circuits", Proceedings of an International Symposium on the Theory of Switching. Harvard University Press, pp. 204–243, April 1959.

[2] V. I. Varshavskij, A. G. Astanovskij, V. B. Marakhovsky, V. A. Peschanskij, L. Ya. Rozenblyum, N. A. Starodubtsev, R. L. Finkel'shtejn, B. S. Tsirlin, "Aperiodic Automata". Moscow, Nauka, p. 304, 1976.

[3] V. B. Marakhovsky, A. V. Surkov, “Globally asynchronous system of interactive Moore state machines”, IET Computers and Digital Techniques, vol. 10, issue 4, pp. 186–192, July 2016.

[4] A. Yakovlev, “Energy-modulated computing,” Design, Automation & Test in Europe Conference & Exhibition (DATE), vol. 39, issue 5, pp. 952–965, April 2011.

[5] Y. A. Stepchenkov, V. N. Zakharov, Y. V. Rogdestvenski, Y. G. Diachenko, N. V. Morozov, D. Y. Stepchenkov, “Speed-independent floating point coprocessor,” 2015 East-West Design & Test Symposium (EWDTS), pp. 29–36, September 2015.

[6] Y. A. Stepchenkov, D. Y. Stepchenkov, Y. V. Rogdestvenski, Y. I. Shikunov, Y. G. Diachenko, “Energy efficient speed-independent 64-bit fused multiply-add unit,” 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic, pp. 1709–1714, January 2019.

[7] Y. A. Stepchenkov, A. N. Kamensky, Y. G. Diachenko, Y. V. Rogdestvenski, D. Y. Diachenko, “Fault-tolerance of self-timed circuits,” 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), pp. 41–44, June 2019.

[8] I. A. Danilov, M. S. Gorbunov, S. G. Bobkov, A. I. Shnaider, A. O. Balbekov, Y,B, Rogatkin, “On board electronic devices safety provided by dice-based Muller C-elements,” Acta Astronautica, vol. 150, pp. 28–32, November 2018.

[9] B. Hollosi, M. Barlow, G. Fu, C. Lee, J. Di, S. C. Smith, H. A. Mantooth, M. Schupbach, “Delay-insensitive asynchronous ALU for cryogenic temperature environments”, 51st Midwest Symposium on Circuits and Systems, pp. 322–325, September 2008.

[10] O. Drozd, O. Ivanova, K. Zashcholkin, V. A. Romankevich, J. Drozd, “Checkability Important for Fail-Safety of FPGA-based Components in Critical Systems,” Intelligent Information Technologies & Systems of Information Security (IntelITSIS), pp. 471–480, March 2021.

[11] O. Drozd, I. Perebeinos, O. Martynyuk, K. Zashcholkin, O. Ivanova, M. Drozd, “Hidden fault analysis of FPGA projects for critical applications,” 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Paper 142, Februrary 2020.

[12] S. C. Smith, “Design of an FPGA Logic Element for Implementing Asynchronous NULL Convention Logic Circuits,” IEEE Transactions on very large scale integration (VLSI) systems, vol. 15, no. 6, pp. 672–683, June 2007.

[13] (2021, June) Speedster22i Configuration User Guide. [Online]. Available: https://www.achronix.com/sites/default/files/docs/Speedster22i_Configuration_User_Guide_UG033_v1.3.pdf

[14] R. V. Vikhorev, “Universal logic cells to implement systems functions,” Proceedings of the 2016 IEEE North West Russia Section Young Researchers in Electrical and Electronic Engineering Conference, pp. 373–375, February 2016.

[15] R. V. Vikhorev, “Improved FPGA logic elements and their simulation,” Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, pp. 259–264, January 2018.

[16] S. F. Tyurin, A. Yu. Skornyakova, Y. A. Stepchenkov, Y. G. Diachenko, “Self-timed Look up Table for ULAs and FPGAs,” Radio Electronics, Computer Science Control, vol. 1, issue 2, pp. 36–45, March 2021.

[17] A. Yu. Skornyakova, R. V. Vikhorev, “Self-Timed LUT Layout Simulation,” Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, pp. 176–179, January 2020.

[18] (2021, June) National Instruments. [Online]. Available: http://www.ni.com/multisim/

[19] (2021, June) Microwind. [Online]. Available: https://www.microwind.net

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

On Polynomial Convergence Rate for Extended Reliability Standby System

Galina Zverkina*
*V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
65 Profsoyuznaya street, Moscow 117997, Russia
zverkina @ gmail . com

***Abstract*—Obviously, in a reliability system consisting of two restorable elements, the distributions of work and repair times are exponential. And obviously, the switching between operating mode and repair mode and vice versa is instantaneous. In this paper, we consider the case when the behaviour (intensity of wear-out or repair) of both elements depends on each other, and the switching can be delayed. The time of such switching can be random, yet we suppose that it is limited.**

The random time of work and repair of elements is determined using intensities. The work and repair intensities depend on the *full state of the system*, i.e. on the status (element work or no) of each element and on its elapsed times in their statuses.

If the distribution of work or repair time of at least one element is non-exponential, the random process describing the behaviour of such a system is not regenerative.

Sufficient conditions for the ergodicity of such a process are formulated.

Also, sufficient conditions for the possibility of calculating the upper polynomial bound for the rate of convergence of the numerical characteristics of the system under consideration are proposed.

***Index Terms*—Convergence rate, generalized Lorden's inequality, strong upper bounds, dependent alternating processes, reliability theory.**

I. Introduction

The reliability system consisting of two dependent restorable elements is studied in this paper. Many studies of reliability systems assume that the time for the operation of each element of the system and that the repair time of the restorable elements has an exponential distribution. It is assumed that the time for switching between operating and repair modes and vice versa occurs instantly.

Using an exponential distribution for work time has some technical rationale. However, in real reliability systems, the operating time may have a non-exponential distribution, for example, the Erlang distribution, etc. The repair time distribution can be continuous or discrete. For example, a repair can be a simple replacement of a failed item. For example, a repair can be a simple replacement of a failed item in a fixed amount of time.

In addition, switching between the states of an element from working to being repaired and vice versa can occur in some (random) limited time.

For the restorable element number j of the pair of restorable elements, we put $\xi_i^{(j)}$ are the times of the work of j-th element, and $\eta_i^{(j)}$ are the times of its repairs.

Thus, the behaviour of both elements describes by alternating renewal process (see, e.g.,[1]).

The behaviour of both elements (i.e. the distributions of $\xi_i^{(j)}$ and $\eta_i^{(j)}$) will be described by *intensities* of failures and repairs.

A. Intensities

Recall, that the intensity of the distribution of a continuous positive random variable (r.v.) ξ with distribution function (d.f.) $F(s)$ is the function $\lambda(s) \stackrel{\text{def}}{=} \dfrac{F'(s)}{1-F(s)}$, and $\mathbf{P}\{\xi \in (s, s+\Delta) | \xi > s\} = \lambda(s)\Delta + o(\Delta)$.

The function $\lambda(s)$ defines d.f. $F(s)$:

$$F(s) = 1 - \int_0^s e^{-\int_0^u \lambda(v)\,\mathrm{d}v}\,\mathrm{d}u. \tag{1}$$

For mixed distributions, where d.f. is a sum of continuous function and step function, we put (see [2]):

$$f(s) = \begin{cases} F'(s), & \text{if } F'(s) \text{ exists} \\ \\ 0, & \text{otherwise;} \end{cases}$$

$$\lambda(s) \stackrel{\text{def}}{=} \frac{f(s)}{1-F(s)} - \sum_i \delta(s-a_i)\ln(F(a_i+0) - F(a_i-0)),$$

where $\{a_i\}$ is the set of discontinuity points of $F(s)$, and $\delta(s)$ is a standard δ-function;
The formula (1) is true with these notations.

So, the distribution of continuous positive, discrete and mixed r.v.'s can be defined by intensities.

B. Conditions for generalized renewal process with renewal times ξ_i

Conditions I. *The following conditions for the (generalized) intensities $\lambda_i(s)$ are assumed:*

Ia. *The (generalized) measurable non-negative functions $\varphi(s)$ and $Q(s)$ exist such that for all $s \geqslant 0$, $\varphi(s) \leqslant \lambda_i^{(n)}(s) \leqslant Q(s)$;*

Ib. $\displaystyle\int_0^\infty \varphi(s)\,\mathrm{d}s = \infty$, *and*

$$\int_0^\infty x^{k-1} \exp\left(-\int_0^x \varphi(s)\,\mathrm{d}s\right)\mathrm{d}x < \infty \text{ for some } k \geqslant 2;$$

Ic. *$Q(s)$ is bounded in some neighbourhood of zero;*

Id. *There exists the constant $T \geqslant 0$ such that $\varphi(s) > 0$ a.s. for all $s > T$.*

Remark 1: Condition Ia holds: $G(s) = \mathbf{P}\{\zeta \leqslant s\} \geqslant F_i^{(n)}(s) = \mathbf{P}\{\xi_i^{(n)} \leqslant s\} \geqslant \Phi(s) = \mathbf{P}\{\widetilde{\zeta} \leqslant s\}$, or $\zeta \gtrsim \xi_i^{(n)} \gtrsim \widetilde{\zeta}$ are ordered in distribution. ▷

Remark 2: Condition Ib holds: there exists $\mathsf{E}\,\zeta^k < \infty \Rightarrow \mathsf{E}\,\widetilde{\zeta}^k < \infty$ and $\mathsf{E}\left(\xi_i^{(n)}\right)^k < \infty$. ▷

Remark 3: Condition Ic holds: $\mathsf{E}\,\zeta^2 > 0$. ▷

Remark 4: Condition Id holds: $\Phi'(x) > 0$ a.s. for $x > T$, i.e. we may consider delayed switching for considered reliability sysytem. ▷

Denote $\mu_i(s) \stackrel{\mathrm{def}}{=} \lambda_i(s) - \varphi(s)$. The intensity $\mu_i^{(n)}(s)$ corresponds the distribution function

$$\mathcal{F}_i^{(n)}(s) \stackrel{\mathrm{def}}{=} 1 - \exp\left(-\int_0^s \mu_i^{(n)}(u)\,\mathrm{d}u\right).$$

C. Conditions for reliability system under study

So, our reliability system consists on two restorable elements, with work times $\xi_i^{(j)}$ and repair times $\eta_i^{(j)}$ with intensities $\lambda_i^{(\xi,n)}$ and $\lambda_i^{(\eta,n)}$ accordingly.

The *full state* of the reliability system is the vector $X_t \stackrel{\mathrm{def}}{=} (n_1, x_1; n_2, x_2)$, where:

$$n_j = \begin{cases} 1, & \text{if } j\text{-th element is in working state;} \\ 0, & \text{if } j\text{-th element is in failure state;} \end{cases}$$

the variable x_j is the elapsed time of the stay of j-th element in the status n_j.

The behaviour of X_t, its distribution give the full information about this reliability system.

For example, the set $\mathcal{S}_{0,0} \stackrel{\mathrm{def}}{=} \{(0, x; 0, y),\ x, y \in \mathbf{R}_{\geqslant 0}\}$ corresponds to the failure state of the system.

Thus, if the characteristics of the distribution of the studied system are known, then the numeric characteristics of the studied system are known.

The intensities of failure (repair) of both element depends on the *full state of system* X_t and satisfy conditions **I** with (generalized) functions $\varphi^\xi(s)$, $Q^\xi(s)$ and $\varphi^\eta(s)$, $Q^\eta(s)$.

Conditions II.

IIa. *k-th moment of any repair or work time are bounded by the constants $C_j^{(\xi,n)}$ and $C_j^{(\eta,n)}$ ($k \geqslant 2$):*

$$\mathsf{E}\left(\xi_j^{(n)}\right)^k = \int_0^\infty s^k\,\mathrm{d}\mathcal{F}_j^{(n)}(s) \leqslant C_j^{(\xi,n)} < \infty;$$

$$\mathsf{E}\left(\eta_j^{(n)}\right)^k = \int_0^\infty s^k\,\mathrm{d}\mathcal{F}_j^{(n)}(s) \leqslant C_j^{(\eta,n)} < \infty.$$

Denote $\left\{C^\ell\right\} \stackrel{\mathrm{def}}{=} \left\{C_1^{(\xi,\ell)}, C_2^{(\xi,\ell)}, C_1^{(\eta,\ell)}, C_2^{(\eta,\ell)}\right\}$.

IIb. *The distributions $F_i^{(\xi,n)}(s)$ (or $F_i^{(\eta,n)}(s)$) are not-lattice, i.e. $\nexists\, a > 0$ such that $\mathbf{P}\{\xi_j^{(n)} \neq ma\} = 0$ for all $n, m \in \mathbf{N}$ (or, accordingly, $\nexists\, a > 0$ such that $\mathbf{P}\{\eta_j^{(n)} \neq ma\} = 0$ for all $n, m \in \mathbf{N}$).* For definiteness, we will assume that the distributions $F_i^{(\eta,n)}(s)$ are non-lattice.

Note, that for the situation, when there exist only polynomial moments, the *positive recurrence* was proved in [3].

II. Main results

Theorem 1: In conditions I and II are satisfied, then the process X_t is ergodic. ▷

The proof of this Theorem 1 based on the Lemma 1 and results of [2].

Lemma 1: If the Conditions I and Condition II, then the distribution of the sum of any two consecutive periods $\xi_i^{(n)}$ and $\eta_i^{(n)}$ satisfies Conditions I and it is non-lattice. ▷

Here we skip the technical proof of this Lemma.

Proof of Theorem 1: Denote $\theta_i^{(n)} \stackrel{\mathrm{def}}{=} \xi_i^{(n)} + \eta_i^{(n)}$. This is one cycle of work and repair of i-th element. This period can be defined by the intensity of distribution $F_i^{(n)} * \mathcal{F}_i^{(n)}(s)$ and its intensity depends on the full state of the process X_t.

Consider two dependent generalized renewal processes

$$N_i(t) \stackrel{\mathrm{def}}{=} \sum_{m=1}^{\infty} \mathbf{1}\left(\sum_{j=1}^{m} \theta_i^{(j)} < t\right).$$

And put $B_i(t) \stackrel{\mathrm{def}}{=} t - \sum_{j=1}^{N_t} \theta_i^{(j)}$.

The pair $(B_1(t), B_2(t))$ is a two-dimensional generalised Markov modulated Poisson process (see [2]), and it is ergodic. ■

Denote $\mathcal{P}_t$ – distribution of X_t: $\mathbf{P}\{X_t \in A\} = \mathcal{P}_t(A)$ for all $A \in \mathcal{B}(\mathcal{X})$. So, $\mathcal{P}_t \Longrightarrow \mathcal{P}$, where $\mathcal{P}$ is invariant probability measure.

Theorem 2: If the Conditions I and Condition II are satisfied, then for all $\ell \leqslant k-1$, and for all initial states of the process X_t, there exists the countable constant $K = K(\ell, X_0, \{C^1\}, \{C^2\}, \{C^\ell\}, \{C^{\ell+1}\}, \varphi(\cdot), Q(\cdot)) = K(\ell, X_0, \vec{C}, \varphi(\cdot), Q(\cdot))$ such that for all $t \geqslant 0$,

$$\|\mathcal{P}_t - \mathcal{P}\|_{TV} \leqslant \frac{K}{t^\ell},$$

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

where $\|\mathcal{P}_t - \mathcal{P}\|_{TV}$ is a distance in total variation metrics:

$$\|\mathcal{P}_t - \mathcal{P}\|_{TV} \stackrel{\text{def}}{=} \sup_{A \in \mathcal{B}(\mathcal{X})} |\mathcal{P}_t(A) - \mathcal{P}(A)|.$$

Schema of the proof. The proof of Theorem 2 is similar to the proof of the main result in [2].

The base of the proof is the coupling method (see, e.g., [4]).

In this paper, all intensities have been separated from zero. This fact made it possible to use Basic Coupling Lemma for continuous distribution (see, e.g., [5]).

We consider the times when the first element of the system finish the repair and starts working.

There are two situations.

1. At the time of completion of the repair of one element, the second element is in working state, then the intensities of both elements are greater than $c > 0$, and by Basic Coupling Lemma, it can create some Markov process with the same marginal distribution in such a way that the states of the elements will coincide with a nonzero probability at the end of the working period. Thus, this is a coupling epoch.

2. In the case when at the time of completion of the repair of one element, the second element is in repair state, it possible to use Lorden's inequality ([6]). This inequality gives an estimate of the expectation of residual time in a given state of each of the elements.

Then it can estimate by Markov inequality the probability of the event *"at the time of the repair of one element, the residual time of the stay in the current state of other elements is less than some constant"*. So, it can estimate the probability of event *"the working element will be in working state at the time when the other element will finish the repair"*.

This event leads to case 1, and it can estimate the probability of the coupling at the end of the current work period.

Theorem 2 is formulated in the situation when the intensities are not separated from zero.

Thus, it must use a modified Lorden's inequality given in [7]. This generalized Lorden's inequality estimates the expectation of residual time of work period of the element of studied reliability system under Condition I (see section I-B).

Applying generalized Lorden's inequality is similar to the description above.

But in the studied reliability system, the work time end repair time has not finite exponential moments. So, the calculation of expectation of coupling epoch is slightly different from [2].

The coupling epoch is bounded by the geometrical sum of independent conditional random variables given *"at the last period in this sum, there was a coupling"*.

The calculation of the bounds $K = K(\ell, X_0, \vec{C}, \varphi(\cdot), Q(\cdot))$ has some technical difficulties to be overcome.

The full proof is constructive and gives the algorithm of calculation of K, and subsequent improvement of estimates. □

III. About more complex reliability systems

The same methods can be applied to analyze more complex systems in reliability theory.

But now, the proposed techniques of the calculations of the bounds for the characteristics of reliability theory gives very great error in calculation (the numerical experiment is demonstrated in [8]).

IV. Conclusion

The bounds

$$\|\mathcal{P}_t - \mathcal{P}\|_{TV} \leqslant \frac{K}{t^\ell},$$

can be useful for an estimate the convergence rate of some parameters of the reliability system dependent on the distribution of the full state system.

For example, availability factor of studied system is $A_t = \mathbf{P}\{X_t \in \mathcal{R}\}$, where $\mathcal{R} = \{0 \times \mathbf{R}_+ \times 0 \times \mathbf{R}_+\} \subset \mathcal{X}$.

So, there exists $\lim\limits_{t\to\infty} A_t = A$, and $|A_t - A| \leqslant \frac{K}{t^\ell}$.

If the behaviour of the elements is not easy for studying (the distributions of work and repair periods are not exponential), it can estimate the convergence rate. Then, the estimation of stationary end non-stationary characteristics can be founded by simulation modelling.

Acknowledgments

The work is supported by RFBR, project No 20-01-00575 A.

References

[1] W. Smith, "Renewal theory and its ramifications," in *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 20, no. 2. Wiley, 1958, pp. 243–302.

[2] G. Zverkina, "A system with warm standby," in *Computer Networks (Proceedings of the 26th International Conference (CN 2019, Kamień Šląski, Poland)*. Springer, 2019, pp. 387–399.

[3] A. Veretennikov, "On polynomial recurrence for reliability system with a warm reserve, markov processes and related fields," in *Markov Processes and Related Fields*, vol. 25, 2019, pp. 745–761.

[4] T. Lindvall, "Lectures on the coupling method." Wiley, New York, 1992.

[5] K. Kato, "Coupling lemma and its application to the security analysis of quantum key distribution," in *Tamagawa University Quantum ICT Research Institute Bulletin*, vol. 4, no. 1, 2014, pp. 23–30.

[6] G. Lorden, "On excess over the boundary," in *The Annals of Mathematical Statistics*, vol. 41, no. 2, 1970, pp. 520–527.

[7] E. Kalimulina and G. Zverkina, "On some generalization of lorden's inequality for renewal processes," in *arXiv.org. 1910.03381*. Cornell university library, 2019, pp. 1–5.

[8] G. Zverkina, "On exponential convergence of availability factor," in *Upravleniye bol'shimi sistemami (About Large-scale Systems Control – in Russian)*, vol. 90, 2021, pp. 5–35.

Algorithms for calculation of logical derivatives for survival signature and their analysis

Patrik Rusnak
University of Zilina,
Faculty of Management Science and Informatics,
Zilina, Slovakia
patrik.rusnak@fri.uniza.sk

***Abstract*—Nowadays, it is necessary to examine the reliability of systems, especially when it comes to systems on the functionality of which great emphasis is placed. Such systems are often made up of components of the same type. For the needs of reliability analysis, it is therefore appropriate to use a survival signature approach, which is based on a structural function, but reduces its dimension and it is, among other things, suitable for comparing systems. When analyzing how a certain type of component affects the functionality of the whole system, it is appropriate to use direct partial logic derivatives. This article will present several algorithms that can be used for their calculation and then their analysis is performed in terms of the time required of these algorithms.**

Keywords—survival signature, structure function, logic differential calculus, algorithm

I. INTRODUCTION

The principal step in reliability analysis is the development of the mathematical description of the investigated system. Different types of the mathematical approaches are used for the system representation [1, 2]. Most often used mathematical representations in reliability engineering are structure function [3, 4], stochastic model (for example, Markov model) [5, 6], universal generating function [7, 8], Monte Carlo based model [7, 8], Petri Nets [9]. Every one model from those mathematical models has specifics for the application. The application of any of these models depends on the investigated system type, conditions of its functioning, goals of the reliability analysis, number of components and its connections/correlations.

The structure function one of the first developed mathematical representations of the system in reliability analysis [10]. According to the definition of the structure function, this function maps the system components states (which are interpreted as function variables values) to the system states (which is considered as the function value). Importance advantages of the structure function are possibility to be formed for the system of any structural complexity and simplicity of the analysis. Importance advantages of the structure function are (a) possibility to be formed for the system of any structural complexity and (b) simplicity of the analysis. The structure function in many researches is interpreted as Boolean function therefore the mathematical methods of Boolean algebra can be used for reliability evaluation of the system based on the structure function. These methods are developed well [3, 4, 11, 12].

One of often problems in the development of mathematical model in form of the structure function is dimensional of the mathematical representation [3, 12]. Need to note that the large dimensional is problem which needs decision for other mathematical representations too, for example, Markov model [5, 6]. Therefore, the development of new approaches and method which allows decreasing the dimensional of these mathematical representations or improve the efficiency of their evaluation is relevant problem in reliability engineering [13, 14].

This problem of reliability analysis of large dimensional system based on the structure function can be decided by the decomposition of initial system [13, 15] or development of special methods for lager dimensional system reliability analysis [16, 17, 18]. One of often used approach in decomposition is modular decomposition, which has effective application in development of fault tree [15], importance analysis [13] and system optimization [19]. Methods of Boolean algebra for analysis of large dimensional function are used in reliability analysis too. BDD based methods are developed for the reliability analysis of large dimensional system [17]. Such methods are developed for special system as network system [20], complex system [21], phase mission system [17]. BDD based methods are developed for the system reliability and availability calculation [17], sensitivity analysis [22], importance analysis [21]. But as known, the development of optimal BDD is complex problem [23]. Therefore, the transformation of the structure function into BDD is more complex problem that the reliability indices and measures calculation based BDD in some applications [23, 24].

There are one more approach for the large dimensional system reliability analysis which is based on the transformation of the structure function into survival signature [16]. The survival signature is probabilistic representation of the system functioning depending on the number of components of specified types. Because this representation is defined depending on the components type, it has less dimensional if components types are less than number of system components. The survival signature-based approach is developed for many problems in reliability analysis and one on them is importance analysis [25, 26]. In this paper alternative method based on survival signature for the system importance analysis is developed. This method is developed based on the application of Logical Differential Calculus, in particular, *Direct Partial Logical Derivative* (DPLD).

DPLDs application in importance analysis have been developed in studies [3, 4, 13, 27]. This derivative allows analyzing the influence of change of one component state into the system reliability change, if the system is represented by the structure function. DPLDs permit to compute the system critical states which are used in the importance measures calculation. But the application of DPLD in some cases has restriction, caused by the large dimensional of the structure function, which is exponentially increases depending on

[1]This work was supported by the Slovak Research and Development Agency under the contracts APVV SK-SRB-18-0002.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

number of the system components. Therefore, in this paper the method for the analysis of the system critical states by DPLDs and based on the system representation by the survival signature are considered. In the paper the detail analysis of the computational complexity of the proposed method is presented and discussed.

II. Theoretical Background

Several mathematical representations are used to describe systems in reliability engineering. One of them is the structure function [14, 21]. It describes the dependence between the states of system components and its state in point of view reliability of this system. This dependency can be expressed as follows [14]:

$$\phi(x_1, \dots, x_n) = \phi(\boldsymbol{x}): \ \{0,1\}^n \to \{0,1\}, \qquad (1)$$

where $\boldsymbol{x} = (x_1, \dots, x_n)$ represents the state vector, i.e., vector of Boolean variables, that represents states of the system components. It is important to point out that the structure function by default describes all the possibilities of functioning of system components and subsequent behavior of the system (functionality or non-functionality of the system). As a result, for n components, there are exactly 2^n different state vectors [14].

A simple illustrative example will now be shown. Consider a series-parallel system whose graphical representation in the form of the reliability block diagram can be seen in Figure 1. This system consists of four components that have one out of two different types. The state of individual components is represented by Boolean variables $x_1, \dots, x_4$. This system is in working state when components 1 and 4 work or when components 1, 2 and 3 work. Based on these facts, it is possible to create a structural function of such a system, which has the following form:

$$\phi(x_1, x_2, x_3) = x_1 \wedge (x_2 \wedge x_3 \vee x_4). \qquad (2)$$

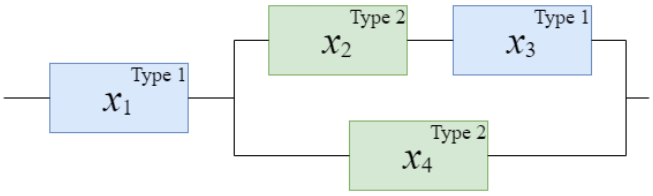


Figure 1. Reliability block diagram of the series-parallel system

In reliability engineering, the system can be understood as coherent and noncoherent [4]. A coherent system is a system, in which the state of each component affects the state of the system and it is impossible for the failure of a system component to cause the improvement of the state of the system [4]. If these conditions are not meet, then the system is incoherent. From this point onward, the coherent system will be considered. The series-parallel system represented by (2) is an example of a coherent system.

For binary-state systems, the definition of the structural function is identical to the definition of the Boolean function [21]. This fact can be used to permit the usage of mathematical methods for the Boolean function in the reliability engineering. One such mathematical method is Logical Differential Calculus (LDC), which can be used to determine how a change in the state of one or more components of a system affects its state [27]. There are several derivatives that are defined within LDCs, one of which is the partial logical derivative (PLD). This derivative examines a particular change in the state of a given component against a particular change in the state of the system, and the change may be reverse or direct (DPLD). For coherent systems, it is appropriate to calculate the DPLD and it can be calculated as follows:

$$\frac{\partial \phi(1 \to 0)}{\partial x_i(1 \to 0)} = \frac{\partial \phi(0 \to 1)}{\partial x_i(0 \to 1)} = \overline{\phi(0_i, \boldsymbol{x})} \wedge \phi(1_i, \boldsymbol{x}), \qquad (3)$$

where $\wedge$ denotes Boolean operation AND and $\bar{}$ is a Boolean operation NOT. For example, if we wanted to see how failure of the component 1 affects the system, we will get the following formula:

$$\begin{aligned} \frac{\partial \phi(1 \to 0)}{\partial x_1(1 \to 0)} &= \overline{(0 \wedge (x_2 \wedge x_3 \vee x_4))} \wedge (1 \wedge (x_2 \\ &\quad \wedge x_3 \vee x_4)) \\ &= 1 \wedge (x_2 \wedge x_3 \vee x_4) \\ &= x_2 \wedge x_3 \vee x_4 \,. \end{aligned} \qquad (4)$$

It is possible to see that in cases where components 2 and 3 are working or component 4 is working, component 1 is crucial for working system, which corresponds with his depiction.

Reliability engineering examines systems whose components are not always different but they are often the same types of components. For such systems, it is more appropriate to use another approach instead of the structural function, which is based on the structural function, but also takes into account the types of components. Such a mathematical approach is known as the Survival Signature (SS) [16] and represents the probability that a system is in functional state if exactly $l_1, \dots, l_K$ components of types $1, \dots, K$ are in functional state. The Survival Signature can be calculated as follows:

$$\Phi(l_1, \dots, l_K) = \left[\prod_{k=1}^{K} \binom{n_k}{l_k}^{-1}\right] * \sum_{\boldsymbol{x} \in S_{l_1,\dots,l_K}} \varphi(\boldsymbol{x}), \qquad (5)$$

where $S_{l_1,\dots,l_K}$ is set of all state vectors for the whole system, for which $\sum_{i=1}^{n_k} x_i = l_k$, $k = 1, \dots, K$ and $\boldsymbol{x} = (x_1, \dots, x_n)$ is binary state vector.

Computed survival signature values for the series-parallel system for each $l_1 \in \{0,1,2\}$ and $l_2 \in \{0,1,2\}$ can be seen in Table I. For example, in case of $l_1 = 1$ and $l_2 = 1$, there are four state vectors $(1,1,0,0)$, $(1,0,0,1)$, $(0,1,1,0)$ and $(0,0,1,1)$ that represent the situation with exactly one working component of type 1 and type 2. However, the series-parallel system is in working state only for one state vector $(1,0,0,1)$ and therefore $\Phi(l_1, l_2) = 0.25$.

Table I
Survival signature of the Series-Parallel system

Number l_1 of working components of type 1	Number l_2 of working components of type 2	$\Phi(l_1, l_2)$
0	0	0
0	1	0
0	2	0
1	0	0

1	1	0.25
1	2	0.5
2	0	0
2	1	1
2	2	1

III. DPLD FOR SURVIVAL SIGNATURE

DPLDs are an excellent tool for finding critical states of systems represented by a structural function. This tool finds its application not only in the reliability analysis of binary-state systems but also in the reliability analysis of multi-state systems [3,4,13]. In [28], three new DPLDs used for Survival Signature are shown. This DPLDs represent a linking of the usefulness of DPLDs in analyzing system reliability and simplification of the description of a system that contains components of the same type using Survival Signature.

The first DPLD for survival signature shows a possibility of a system failure for a fixed number of working components of specific types if one of the components of type k fails and it can be computed as follows:

$$\frac{\partial \Phi(l_1,\dots,l_K)\downarrow}{\partial l_k(a\to a-1)} = \begin{cases} 1, \Phi(l_1,\dots,a_k,\dots,l_K) > \Phi(l_1,\dots,a_k-1,\dots,l_K) \\ 0, otherwise \end{cases} \quad (6)$$

where $a \in \{1,\dots,n_k\}$ represents count of working components of type $k \in \{1\dots,K\}$.

Taking into account the series-parallel system, his first DPLDs for each type and each possible number of working components are shown in Table II. From those DPLDs it is possible to see, that the least critical change in system survivability is when there are two working components of type 2 and one will fail. As for the rest, if all components of other type are in failed state, then the change in number of working components of analysed type is not significant.

TABLE II

THE FIRST DPLDS FOR THE SERIES-PARALLEL SYSTEM

l_1	l_2	$\Phi(l_1,l_2)$	$l_1(2\to1)$	$l_1(1\to0)$	$l_2(2\to1)$	$l_2(1\to0)$
0	0	0	-	-	-	-
0	1	0	-	-	-	0
0	2	0	-	-	0	-
1	0	0	-	0	-	-
1	1	0.25	-	1	-	1
1	2	0.5	-	1	1	-
2	0	0	0	-	-	-
2	1	1	1	-	-	1
2	2	1	1	-	0	-

The second DPLD for survival signature shows a possibility of a system failure if one of the system components of specified type fails, which can be computed as follows:

$$\frac{\partial \Phi(l_1,\dots,l_K)\downarrow}{\partial l_k\downarrow} = \begin{cases} 1, \Phi(l_1,\dots,l_k,\dots,l_K) > \Phi\left(l_1,\dots,\widetilde{l_k},\dots,l_K\right) \\ 0, \text{otherwise} \end{cases} \quad (7)$$

where $l_k \in \{1,\dots,n_k\}$, and $\widetilde{l_k} = l_k - 1$. This derivation can also be computed from the first DPLD (6) as follows:

$$\frac{\partial \Phi(l_1,\dots,l_K)\downarrow}{\partial l_k\downarrow} = \sum_{a=1}^{n_k} \frac{\partial \Phi(l_1,\dots,l_K)\downarrow}{\partial l_k(a\to a-1)} \quad (8)$$

As for the series-parallel system, its second DPLDs can be seen in Table III. From those values we can see each type from a more holistic type of view, which can be used to further understand its criticality for system survivability. As we can see, type 1 more critical for system survivability than type 2, which is mostly to the fact, that the component 1 has this type.

TABLE III

THE SECOND DPLDS FOR THE SERIES-PARALLEL SYSTEM

l_1	l_2	$\Phi(l_1,l_2)$	$l_1\downarrow$	$l_2\downarrow$
0	0	0	-	-
0	1	0	-	0
0	2	0	-	0
1	0	0	0	-
1	1	0.25	1	1
1	2	0.5	1	1
2	0	0	0	-
2	1	1	1	1
2	2	1	1	0

Lastly, the third DPLD for survival signature describes a measure of the system failure if the one of the components of specified type fails, which can be computed as follows:

$$\frac{\partial \Phi(l_1,\dots,l_K)\Downarrow}{\partial l_k\downarrow} = \begin{cases} \xi, & \Phi(l_1,\dots,l_k,\dots,l_K) > \Phi\left(l_1,\dots,\widetilde{l_k},\dots,l_K\right) \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where $\xi = \Phi(l_1,\dots,l_k,\dots,l_K) - \Phi\left(l_1,\dots,\widetilde{l_k},\dots,l_K\right)$ for $l_k = 1,..,n_k$, $l_k > \widetilde{l_k}$ and $\widetilde{l_k} = l_k - 1$. This DPLD differs from the previous DPLDs, because it did not just indicate the change, but it describes the volume of a change of the system's survivability. This DPLD can be also computed from the structure function by using SS approach as follows:

$$\frac{\partial \Phi(l_1,\dots,l_K)\Downarrow}{\partial l_k\downarrow} = (n_k)^{-1}\cdot \sum_{x_i\in N_k} \Phi\left(\frac{\partial \phi(1\to0)}{\partial x_i(1\to0)}\right) \quad (10)$$

where $\Phi\left(\frac{\partial \phi(1\to0)}{\partial x_i(1\to0)}\right)$ is transformation of each DPLD $\frac{\partial \phi(1\to0)}{\partial x_i(1\to0)}$ based on the rules of the survival signature and N_k is a set of all components of type k.

In case of series-parallel system, third DPLDs can be seen in Table IV. From those values we can see that the most critical change is when there are 2 working components of type 1 and one component of type 2 and component of type 2 fails. This means that in system, there will be only components of type 1 and this means that the series parallel system will surely fail. On the other hand, the least critical changes are when there is one component of type 1 and one component of type 2 and one component will fail or when there is once component of type 1 and two components of type 2 and one component of type 2 fails.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE IV
THE THIRD DPLDS FOR THE SERIES-PARALLEL SYSTEM

l_1	l_2	$\Phi(l_1, l_2)$	$l_1 \Downarrow$	$l_2 \Downarrow$
0	0	0	-	-
0	1	0	-	0
0	2	0	-	0
1	0	0	0	-
1	1	0.25	0.25	0.25
1	2	0.5	0.5	0.25
2	0	0	0	-
2	1	1	0.5	1
2	2	1	0.5	0

IV. EVALUATION AND EXPERIMENTS

The new DPLDs defined for Survival Signature are useful for a reliability analysis based on Survival Signature and therefore we decided in this chapter to present the implementation of the proposed DPLDs for computer processing and computation. As a first step, it was necessary to create an algorithm that could compute these derivatives and to show their time and memory complexity. We create those algorithms from the assumption that we have a structural function defined as a vector of values, while it is possible to obtain a state vector from the position of a given value in the vector. We thought similarly about the representation of the Survival Signature.

Another thing that needs to be addressed before the algorithms themselves is the fact that the third DPLD can be used in the calculation of the second DPLD, so that if the value of the third DPLD is greater than zero, the value of the second DPLD will be 1, if its value is 0, then the second DPLD will have the value 0, and if the third DPLD is not defined for the given state vector, then the second DPLD will not be defined either. Similarly, it is possible to obtain the first DPLD from the second or third DPLD, while it is necessary to pay attention to the current number of functioning components of specific type and in case of obtaining the first DPLD from the second, values of the second DPLD are only copied, provided that the condition of the number of currently functioning elements of the given type holds.

The third DPLD for type k can be computed by using the Survival Signature (9) of by using the structure function (10). As for the first approach, its algorithm can be presented as follows:

- For each value of the SS at index i execute the following steps:
  - Compute the state vector from i;
  - If the number of working components of type k is at least one:
    - Obtain index j that represents the state vector, in which the number of working components of type k is decreased by one;
    - Save value of the third DPLD at i as difference between value of the SS at i and j;
  - Else:
    - Save value of the third DPLD at i as undefined.

As for the second approach that uses the structure function, if we have at our disposal vector map with mapping from the structure function to the SS its algorithm can be as follows:

- Initialize the temporary vector vec with number of elements as in case of SS and each element consists of two integers initialized to value zero. The first value sum represents sum of values of the DPLDs of the structure function for specific index and the second value all represents its count.
- For each component c of type k execute the following steps:
  - For each value at index i of the structure function, for which the component c is working execute the following steps:
    - Obtain index j that represents the state vector, in which the component c is in failed state;
    - Get index l from map;
    - If the value at index i is different from the value at index j;
      - Increase the value of the sum in vec at index l by one;
    - Increase the value of the all in vec at index l by one;
- For each element in vector vec at index i execute the following steps:
  - If
    - Save value of the third DPLD at i as a ratio between values sum and all;
  - Else
    - Save value of the third DPLD at i as undefined.

The second DPLD for type k can be computed by using the SS or by using the third DPLD. As for the first approach, its algorithm can be presented as follows:

- For each value of the SS at index i execute the following steps:
  - Compute the state vector from i;
  - If the number of working components of type k is at least one:
    - Obtain index j that represents the state vector, in which the number of working components of type k is decreased by one;
    - Save value of the second DPLD at i as one if value of the SS at i and j differs, otherwise save value zero;
  - Else:
    - Save value of the second DPLD at i as undefined.

As for the second approach that uses the third DPLD, its algorithm can be as follows:

- For each value of the third DPLD at index i execute the following steps:
    - If the third DPLD is defined and its value is greater than zero:
        - Save value of the second DPLD at i as one;
    - Else:
        - Save value of the second DPLD at i as value of the third DPLD at i.

The first DPLD for type k and number of working components l can be computed by using the SS, by using the third DPLD or by using the second DPLD. As for the first approach, its algorithm can be presented as follows:

- For each value of the SS at index i execute the following steps:
    - Compute the state vector from i;
    - If the number of working components of type k is l:
        - Obtain index j that represents the state vector, in which the number of working components of type k is decreased by one;
        - Save value of the first DPLD at i as one if value of the SS at i and j differs, otherwise save value zero;
    - Else:
        - Mark value of the second DPLD at i as undefined.

As for the second approach that uses the third DPLD, its algorithm can be as follows:

- For each value of the third DPLD at index i execute the following steps:
    - Compute the state vector from i;
    - If the number of working components of type k is l:
        - Save value of the first DPLD at i as one if value of the third DPLD at i is not zero, otherwise save value zero;
    - Else:
        - Mark value of the first DPLD at i as undefined.

As for the third approach that uses the second DPLD, its algorithm can be as follows:

- For each value of the second DPLD at index i execute the following steps:
    - Compute the state vector from i;
    - If the number of working components of type k is l:
        - Save value of the first DPLD at i as value of the second DPLD at i;
    - Else:
        - Mark value of the first DPLD at i as undefined.

All the above-mentioned algorithms were implemented in C ++ for the purpose of performing experiments. The experiments were performed on parallel and series systems, alternating between those two systems. Systems had from 15 to 20 components and the type ranged from 1 type to the maximum possible number of types. Each variant was performed 100 times and their values can be seen in Figure 2. In this article, we will be focusing on results with 20 number of components and with 14 to 20 different types of system components to better show attributes of each algorithm. Those results can be seen in Figure 3. From them we can see, how time for computation of the SS and DPLDs differs for each number of different types. It is needed to point out that each time is averaged from 100 different attempts and it is a sum for each type.

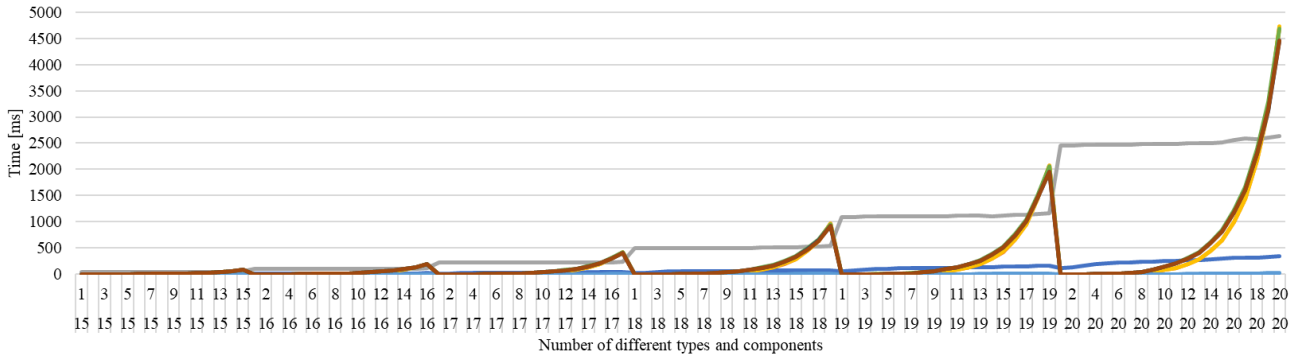


Figure 2. Experiment results for 15 to 20 components

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

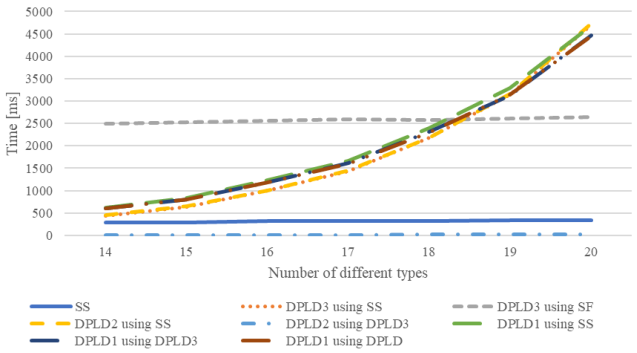


Figure 3. Results of the experiment with 20 components

In further analyses of the results, we divide them into three parts, depending on which DPLDs are partial results related to, while in each part there is also the time to calculate the SS for the purposes of mutual comparison between parts.

The results from the third DPLD can be seen in Figure 4. These results are very interesting, mainly because they nicely show how the calculation gradually becomes more time demanding in the case of using SS (orange dotted line), while in the case of using the structure function (grey dashed line) time demand is quite stable. This is based on the principle of operation of the respective algorithms. In the case of calculation from the SS, the increase in time is mainly due to an increase in the number of different types, which result in an increase in the number of SS values and the number of third DPLD calculations.

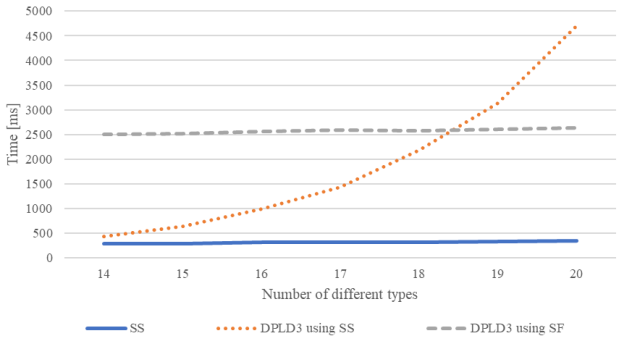


Figure 4. Results of the experiment for the third DPLD

In the case of the second DPLD, its results can be seen in Figure 5. Again, these are interesting results, where it can be seen that the calculation based on SS (dashed yellow line) is more time demanding than the calculation from the third DPLD (light blue dotted line). This is mainly due to the fact that in the case of the calculation from the third DPLD only its values are transformed, while in the calculation from the SS it is necessary to recalculate several transformations from and to the state vector for each DPLD calculation.

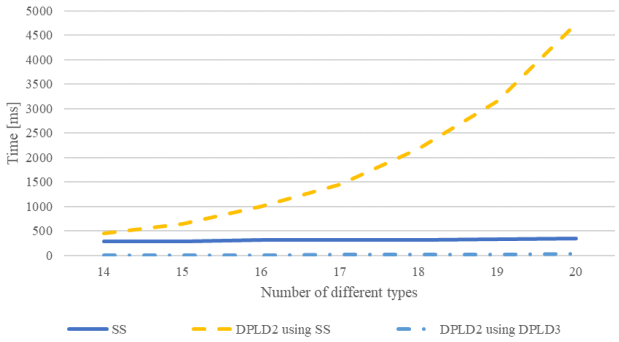


Figure 5. Results of the experiment for the second DPLD

Finally, we will turn our attention to the first DPLD, the results of which can be seen in Figure 6. Here we can see again interesting results because in the case of calculation from SS (green dashed line) the time complexity is a little greater than in the case of using the third DPLD (dark blue dotted-dashed line) or the second DPLD (brown double dotted-dashed line). This is mainly due to the fact that in each computation it is necessary to emphasize not only the types of components, but also the number of functional components of a given type.

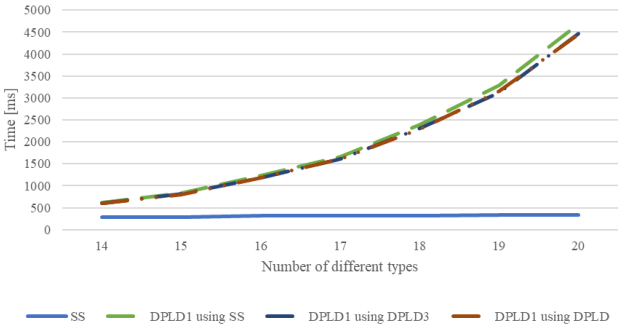


Figure 6. Results of the experiment for the first DPLD

CONCLUSION

Based on the results shown in previous part, it can be stated that an algorithm using SS is more suitable for calculating the third DPLD, and if the number of component types is close to the number of components, it is more appropriate to use an algorithm that uses a structure function. In the case of the second DPLD, it is more appropriate to use an algorithm based on the third DPLD, because it will mostly just be copying values from the third DPLD and there is no need to compute state vectors, which is apparent in incomparably faster computation time than for the algorithm using SS. And as for the first DPLD, here it is almost identical for all mentioned algorithms, although the algorithms using an already calculated DPLDs are a bit faster than the algorithm based on SS.

In the future, it is planned to perform a more detailed experiments for the detailed analysis of proposed algorithms and also to pay more attention to DPLD for SS in case of solving state change of several types of components and the possibility of calculating importance measures for systems described by SS.

REFERENCES

[1] D. W. Coit, E. Zio, "The evolution of system reliability optimization", *Reliability Engineering & System Safety*, vol. 192, 2019.

[2] *Reliability and Statistical Computing*. *Modeling, Methods and Applications*, H.Pham, Ed., Cham, Springer, 2020.

[3] E. Zaitseva, V. Levashenko, "Investigation multi-state system reliability by structure function", *Proc. of the Int. Conf. on Dependability of Computer Systems (DepCoS - RELCOMEX 2007)*, 2007, pp. 81-90.

[4] M. Kvassay, V. Levashenko, E. Zaitseva, "Analysis of minimal cut and path sets based on direct partial Boolean derivatives", *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 230, no. 2, 2016, pp. 147-161.

[5] I.A. Papazoglou, "Semi-Markovian reliability models for systems with testable components and general test/outage times", *Reliability Engineering & System Safety*, vol. 68, no. 2, 2000, pp 121-133.

[6] H. Yi, L. Cui, J. Shen, Y. Li, "Stochastic properties and reliability measures of discrete-time semi-Markovian systems", *Reliability Engineering & System Safety*, vol. 176, 2018, pp 162-173.

[7] A. Kumar, S. Tyagi, M. Ram, "Signature of bridge structure using universal generating function", *International Journal of Systems Assurance Engineering and Management*, vol. 12, no. 1, 2021, pp. 53-57.

[8] S. Bisht, S.B. Singh, R. Tamta, "Reliability evaluation of repairable weighted system using interval valued universal generating function", *International Journal of Quality and Reliability Management*, vol. 37, no. 7, 2020, pp. 957-981.

[9] D. Eisenberger, O. Fink, "Assessment of maintenance strategies for railway vehicles using Petri-nets", *Transportation Research Procedia*, vol. 27, 2017, pp. 205-214.

[10] R.E. Barlow, *Statistical theory of reliability and life testing: probability models*, Holt, Rinehart and Winston, 1974.

[11] W. G. Schneeweiss, "A short Boolean derivation of mean failure frequency for any (also non-coherent) system", *Reliab. Eng. Syst. Saf.*, vol. 94, no. 8, 2009, pp. 1363–1367.

[12] N. Brînzei, J.F. Aubry, "Graphs models and algorithms for reliability assessment of coherent and non-coherent systems". *Proceedings of the Institution of Mechanical Engineers Part O Journal of Risk and Reliability*, vol. 232, no. 2, 2018, pp. 201-215.

[13] M. Kvassay, E. Zaitseva, "Topological Analysis of Multi-state Systems Based on Direct Partial Logic Derivatives", A. Lisnianski, I. Frenkel, A. Karagrigoriou (ed.), *Recent Advances in Multi-state Systems Reliability*. *Springer Series in Reliability Engineering*, 2017, pp. 265-281.

[14] E. Zio, "Reliability Engineering: Old Problems and New Challenges", *Reliability Engineering and System Safety*, vol. 94, 2009, pp. 125–141.

[15] Z. Li, Y. Ren, L. Liu, Z. Wang, "Improving parallel modularization algorithm of large complex fault trees", *Proceedings - Annual Reliability and Maintainability Symposium*, 2016, pp. 1–6.

[16] F. P. A. Coolen, T. Coolen-Maturi, "Generalizing the signature to systems with multiple types of components", *Adv. Intell. Soft Comput.*, vol. 170 AISC, 2012, pp. 115–130.

[17] X. Zang, H. Sun, K. S. Trivedi, "A BDD-based algorithm for reliability analysis of phased-mission systems", *IEEE Trans. Reliab.*, vol. 48, no. 1, 1999, pp. 50–60.

[18] J. D. Andrews, S. J. Dunnett, "Event-tree analysis using binary decision diagrams", *IEEE Trans. Reliab.*, vol. 49, no. 2, 2000, pp. 230–238.

[19] W. Kuo, R. Prasad, "System Reliability Optimization: An Overview", R. Soyer, T.A. Mazzuchi, N.D. Singpurwalla (eds), *Mathematical Reliability: An Expository Perspective. International Series in Operations Research & Management Science*, vol 67, Springer, Boston, MA, 2004.

[20] R. Qiang, "Reliability analysis of wireless sensor networks based on fusion of binary decision diagram and fault tree", *Frontiers in Artificial Intelligence and Applications*, vol. 299, 2017, pp. 422-427.

[21] M. Kvassay, E. Zaitseva, V. Levashenko, J. Kostolny, "Binary Decision Diagrams in reliability analysis of standard system structures", *IDT 2016 - Proceedings of the International Conference on Information and Digital Technologies 2016*, 2016, pp. 164–172.

[22] I. Antoniano-Villalobos, E. Borgonovo, S. Siriwardena, "Which Parameters Are Important? Differential Importance Under Uncertainty," *Risk Anal.*, vol. 38, no. 11, 2018, pp. 2459–2477.

[23] R. Banov, Z. Šimić, D. Grgić, "A new heuristics for the event ordering in binary decision diagram applied in fault tree analysis", *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 234, no. 2, 2020, pp. 397-406.

[24] P. Rusnak, L. Cajka, M. Kvassay, "Software Tool for Manipulation with Decision Diagrams Used in Reliability Analysis", *Proc. Of the 16th IEEE International Conference on Emerging eLearning Technologies and Applications*, 2018, pp. 475-482.

[25] S. Eryilmaz, F. P. A. Coolen, T. Coolen-Maturi, "Marginal and joint reliability importance based on survival signature", *Reliab. Eng. Syst. Saf.*, vol. 172, 2018, pp. 118–128.

[26] G. Feng, E. Patelli, M. Beer, F. P. A. Coolen, "Imprecise system reliability and component importance based on survival signature", *Reliab. Eng. Syst. Saf.*, vol. 150, 2016, pp. 116–125.

[27] E. Zaitseva, V. Levashenko, J. Kostolny, "Importance analysis based on logical differential calculus and Binary Decision Diagram", *Reliab. Eng. Syst. Saf.*, vol. 138, 2015, pp. 135–144.

[28] P. Rusnak, E. Zaitseva, F. P. A. Coolen, M. Kvassay, V. Levashenko, "Logic Differential Calculus for Reliability Analysis Based on Survival Signature", *IEEE Transactions on Dependable and Secure Computing* (submitted)

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

On the Pragmatics of the Turing Test

Baptiste Jacquet
Laboratoire Cognition Humaine et Artificielle (CHArt-UP8)
Paris, France
& Université Paris 8
Saint-Denis, France
& Association P-A-R-I-S
Paris, France
baptiste.jacquet@paris-reasoning.eu

Frank Jamet
Laboratoire Cognition Humaine et Artificielle (CHArt-UP8)
Paris, France
& CY Cergy-Paris Université
ESPE de Versailles
Paris, France
& Association P-A-R-I-S
Paris, France
frank.jamet@paris-reasoning.eu

Jean Baratgin
Laboratoire Cognition Humaine et Artificielle (CHArt-UP8)
Paris, France
& Université Paris 8
Saint-Denis, France
& Association P-A-R-I-S
Paris, France
jean.baratgin@paris-reasoning.eu
Corresponding author

***Abstract*—The Turing Test was initially suggested as a way to give an answer to the question "Can machines think". Since then, it has been heavily criticized by philosophers and computer scientists both as irrelevant, or simply inefficient in order to evaluate a machine's intelligence. But while arguments against it certainly highlight some of the test's flaws, they also reveal the confusion that exists between thinking and intelligence. While we will not attempt here to define the concept of intelligence, we will instead show that such a definition becomes irrelevant if the Turing Test is instead considered to be a test of the humanness of a conversational partner instead, an experimental paradigm that can be used in order to investigate human inferences and expectations. We will review studies which use the Turing Test this way, not only in computer sciences where it is commonly used to evaluate the humanness of a chatbot but also its uses in the field of psychology where it can be used to understand human reasoning in conversation either with a chatbot or with another human.**

***Index Terms*—Turing Test, Chatbots, Cognitive Psychology, Pragmatics, Theory of Mind**

I. Introduction

With the recent advances of Artificial Intelligence, the idea that machines might become able to think for themselves sooner or later is making its way in the general population, helped by many movies like Spike Jonze's HER, Alex Garland's Ex Machina or Ridley Scott's Blade Runner among many others. But how close to this are we really? Can a machine really think? Do we have the tools necessary to evaluate it like Alan Turing [1] claimed with his famous Imitation Game or Turing Test (written TT from now on)?

This paper attempts to review some of the literature exploring the main issue that chatbots (the programs that can be evaluated by the TT) still face today: relevance, in other words, the ability to produce sentences that take into account the expectations of the users. Indeed, most of these programs still fail to sustain conversations of more than a couple successive sentences without relying on generic replies.

We will first describe what types of chatbots exist today and some of the technologies supporting them, before moving on to the ways that are used to evaluate them. We will then describe the TT along with some of its criticisms and the importance of conversational pragmatics in order to successfuly pass it. Finally we will review review some of the literature exploring the use of the TT as an experimental paradigm to study human reasoning and conversational pragmatics.

This paper was funded by the Carnot Institute, for Project LAC (Langages Alternatifs et Cognition).

II. Chatbots

Virtual conversational agents, often called chatbots (or chatterbots), stem from an idea initially suggested by Turing in his test for intelligence [1]. They are machines which process text messages, attempt to find a relevant answer before sending a text reply to the user.

Chatbots are all but new, with the first application meeting this definition being ELIZA [2]. Meant to behave like a psychotherapist focusing on empathy and repetitions, this program uses keywords in the user's messages to send questions back to them. The conversation is kept going by the chatbot giving the impression of trying to dig dipper into the user's mind and psyche. It does so by analyzing the content of its input text, giving each recognized word a value before reorganizing the order of the words according to predefined patterns to generate a reply in the form of a question. If the content of the input doesn't match a recognized pattern, predefined fallback answers are sent back. ELIZA was initially designed to show how superficial Human-Machine conversation was, and yet users still ended up sharing deep and private aspects of their lives with it, often being fooled and refusing to believe they had been talking to a machine. A simple conversation with a version of ELIZA is shown in Fig. 1.

ALICE [3] is another chatbot which generalizes the concept of pattern matching to send replies to the user without necessarily asking questions back like a psychotherapist might. Just like ELIZA, it still uses keywords to detect patterns in the input. Below is an example of a simple exchange with ALICE asking about her favorite color.

Human: Hello ALICE.
A.L.I.C.E: Hi there!
Human: What is your favorite color?

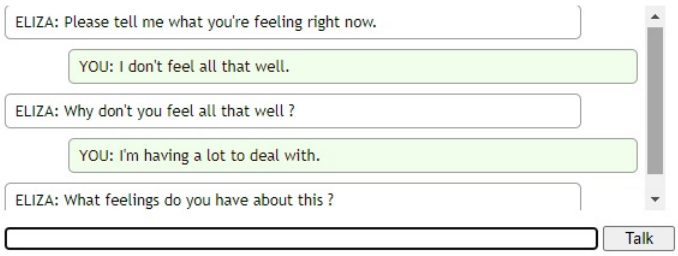


Fig. 1. Example of a conversation we had with ELIZA. This version of the chatbot can be found at https://www.eclecticenergies.com/psyche/eliza

A.L.I.C.E: I don't have a favorite colour. But my favorite book is Don't Read Me.
Human: What can you tell me about it?
A.L.I.C.E: Ambiguous: "Tell me about it." – colloquial or "Tell me about ".

We can see above that ALICE also uses humor in this conversation, making it seem a bit more human, although it then fails to recognize the input in the last message sent by the user in our example. ALICE is built around its own language called AIML (Artificial Intelligence Markup Language) which is today used by many chatbots for its simplicity, making ALICE a good example or what can be expected of chatbots today.

While the simplicity and predictability of pattern-matching chatbots makes them quite attractive, even the most advanced of these types of chatbots cannot (to this day) deal with implicit content in the input messages [4]. Indeed, as we can see in the previous conversation with ALICE, it fails to understand what the user is referring to in the sentence "What can you tell me about it?". Mitsuku (now called Kuki.AI), the latest chatbot in this category to have won the Loebner prize (international competition based on the TT to evaluate chatbots), is still incapable of responding in a relevant way to implicit content in messages unless follow-up questions are pre-programmed (like about favorite color, pets..., see in Fig. 2) [4].

Pattern-matching is not the only tool available to create chatbots though. With the rise of machine-learning techniques, including supervised and unsupervised learning algorithms (depending on whether they use human-labeled data or not), new opportunities for the development of chatbots are being created, potentially allowing chatbots to use contextual cues more efficiently. Unfortunately no chatbot based on these techniques is currently openly accessible in order to test their actual abilities, though platforms exist to create them such as IBM Watson, Amazon Lex, Facebook Messenger, Microsoft LUIS and others. They often use algorithms such as LSTM (long short term memory) [5], RNN (recurrent neural networks) [6], BRNN (bidirectional recurrent neural networks) [7] or CNTN (convolutional neural tensor networks) [8].

There are two main types of chatbots based on these machine-learning techniques: retrieval-based models and generative models. Retrieval-based models read the input sentence

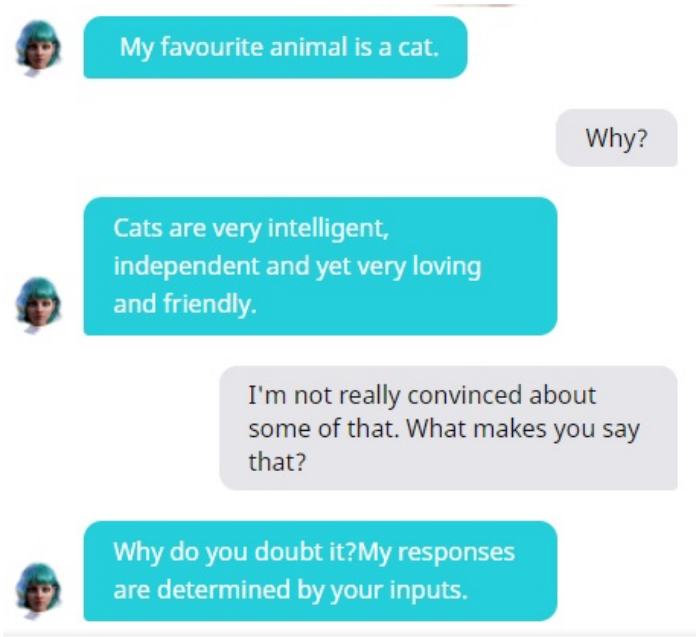


Fig. 2. Example of a conversation we had with Kuki.ai. Notice how it answers the first follow up question correctly, but fails with the second follow up question. This version of the chatbot can be found at https://chat.kuki.ai/

produced by the user to create a thought vector representing the meaning of the sentence (or intent). This thought vector is then compared to entries in a database containing the possible answers the chatbot can give. The entry that is closest to the generated thought vector is selected and sent to the user as the chatbot's reply. Generative models instead use the thought vector as the basis to generate new sentences word by word using the probabilities of a word appearing given the learnt probabilities in the general language and the thought vector that was created while reading the input sentence. Generative models are much more flexible than retrieval-based models as they can generate completely new sentences while retrieval-based models give more control to the owner of the chatbot as they can decide precisely what will be said and what will not. Hybrid models also exist combining both of these aspects. For example models that attempt to generate sentences but if they fail to do so fallback to retrieval methods.

The tools used in these models are based on machine translation using encoder and decoder systems to predict the next words in a sequence. Instead of translating a sentence from one language to another, they instead "translate" an input sentence into a reply. Indeed, much work has been done already in the field of machine translation and the tools used in this domain seem to give generally decent results when used to generate replies in conversations instead, though they are certainly not at the human-level yet and are generally unsatisfying [9].

III. Evaluating chatbots

Evaluating the quality of chatbots remains a rather controversial topic has no standard metric fits the following

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE I
EXAMPLE IN WHICH THE AUTOMATED BLEU OR METEOR METRICS WOULD GIVE FALSE-NEGATIVES.

Context of the conversation
Speaker A: Did you hear that the new Batman movie is coming soon?
Speaker B: Actually I did not. Do you think it's going to be good?
Ground truth
Speaker A: I'm not sure. They keep making new ones but they keep getting worse.
Chatbot
Speaker A: Yeah, I can't wait for it. I'm going to get tickets as soon as possible.

three criteria: automation, similarity to human judgment and precision. In this section we will review some of the methods that can be used today to evaluate conversational agents.

The TT remains a gold standard. Indeed, most users want to feel like they are conversing as easily with the chatbot as they would with a human [10]. In this case users are usually asked to evaluate how human-like the conversation felt. This method has shortcomings when it comes to being automated and does not have a good precision if no additional measures are not added. Indeed, while getting a high evaluation on the human-like aspect is the end goal, only asking the user once at the end of the conversation does not give a good indication on when during the conversation mistakes were made.

Other automated measures are often used: the task completion rate (TCR) which is especially valuable for goal-oriented chatbots which try to help users with a specific task, but cannot be applied to general purpose or chit-chat bots; the duration of the conversation can also be used, with the assumption that longer conversations mean more engagement and thus a more pleasant experience; the number of turns during the conversation, which gives another idea of the engagement and interest of the user in conversing with the chatbot. These are easy to measure but are not well correlated with the results of the TT. They also do not give insights in what went wrong when these measures give low numbers as they only inform on the general conversation rather than specific replies of the conversational agent.

Some measures instead give more specific information regarding the different turns in the conversation themselves rather than a global rating of the conversation. The most commonly used techniques are machine translation techniques like BLEU [11] and METEOR [12]. They assess how similar the generated replies are to an answer which would have been given by a human to the same question. These methods have the great advantage of being easy to automate, but have the disadvantage of not taking into account prior elements of the conversation. Besides, comparing the words being used can create false-negatives, as a perfectly intelligible and human-like response might go in an unexpected way that would be different to the sentences it would be compared to, and thus give a low score despite being perfectly valid (see an example in Table I) [13].

Artificial intelligence can also be used to evaluate the quality of chatbots. For example, RNN can be trained to mimic the evaluation of chatbots made by humans [14]. Scores given by the neural network were then significantly correlated to those given by humans on a scale of appropriateness, which the authors indicate to be the most consistent metric between human judges. Unfortunately the accuracy of such evaluation models tend to also depend on the context of the conversations (surely one would also appreciate the irony of having a chatbot emulating a human being evaluated by a similarly produced artificial judge emulating the evaluation of a human judge. It still remains an interesting first pass of evaluation). It is also possible to aggregate different metrics using trained models to emulate human judges rather than focusing on a single metric, such as engagement (captured with the number of turns or the median duration of the conversations), coherence, conversational depth, topical diversity and domain coverage [15]. The main issue here is that some ratings can be quite subjective and give a high variability. For example, the authors indicate that "A user might give a conversation 5 stars because he/she thought the socialbot was humorous, while another user might find it unknowledgeable". Thus it might be unfair to chatbot to expect them to be generally better at everything than other chatbots, while just like humans, some chatbots might be better suited than others to some tasks and not perform as well in others while remaining above an acceptable baseline.

Finally, an ideal metric would also include a rating of emotional aspects of the chatbot. Especially in conversations related to physical or mental health, having a robot show emotional skills such as empathy is an important aspect to improve how the users view and interact with the chatbot [16], [17]. These social skills would also likely be important to evaluate in the contexts of education and customer services.

Using human judges still remains the gold standard as ultimately these tools are meant to be interacting with humans. Despite the important part of subjectivity in human evaluations due to their individual expectations of a conversational partner, not all aspects of a normal human conversation are currently being encompassed by automated measures, and thus humans need to remain a part of the testing loop and the TT still has good days ahead of it before it can be fully replaced.

IV. THE TURING TEST AS A TEST OF HUMANNESS

While the TT was initially proposed to be a test of the intelligence of a machine [1], it has clearly shifted to being viewed as a test of humanness and is now used as such in the existing panel of evaluating metrics.

The TT, in its modern understanding of it, consists in having a human judge chat, through a text interface, with two other agents: a human and a machine. The goal for the human judge is to find which of the agent is the human, and which is the machine (or in some versions whether there is a machine at all). If after a five minute conversation the judge fails to identify the machine correctly in 50% of the trials then the machine is so much like a human that, according to Turing, it would be necessary to attribute thoughts to it in the same way

we do so with humans: assuming they have mental states the way we do because they behave the way we do.

Lassègue [18] indicates that there is also another entity which is important to consider in the Turing Test: what he calls the umpire, the experimenter or the arbiter who will stop the test after a specific amount of time and tell if the judge was right or wrong, who already knows the answer to the test. This is important because the amount of time required to pass a TT varies greatly, sometimes without much justification. Turing suggests 5 minutes, but why not 10? 7? 10 minutes and 30 seconds? Past that time the chatbot could potentially reveal itself quite clearly.

The TT historically received many critics when it was suggested to be testing intelligence. One of the most famous being Searle's Chinese room argument [19]. In summary, Searle is in a room in which he is given Chinese symbols that he must reply to with Chinese symbols, along with some instructions in English (called a program) to link one input list of symbol to one output list of symbols. Unable to understand Chinese himself, Searle claims that if he was able to fool Chinese people simply by following the instructions (program) given to him in making them believe that he was Chinese himself, he still would not understand Chinese at all, and would be mindlessly following these instructions.

It is important to point out here, that Searle only applied his objection to a specific kind of AI: formal AI, using formal rules to interact through text with the user, he did not say that machines would never be able to think, but that in order to do so we would need to understand the brain rather than abstracting its general functions without understanding how it is working. A machine able to pass the Turing Test thanks to the perfect use of the manipulation of symbols would not necessarily have a mind of its own, would not necessarily think, would not necessarily be intelligent. As others pointed out, these symbols need to be grounded in one way or another, to represent something to really mean anything, thus the need for a more sensori-motor development of AI [20] along with an understanding of how the brain works and understand objects [21]. Searle indeed explains:

> As to whether or not machines will be conscious, it is important to remember that we are machines. We are biological machines and we are conscious. I do not see any reason, in principle, why we could not build an artificial machine that was conscious, but we are unable to do that now because we do not know how the brain does it. The question, "Can you build an artificial machine that is conscious?" is just like the question "Can you build an artificial heart that pumps blood?" We know how to build artificial hearts because we know how the biological heart works. We do not know how to build an artificial brain because we do not know how the brain works. But assuming we knew how the brain worked, I see no obstacle in principle to building an artificial conscious machine. The important thing to see is that the human brain is a machine, a biological machine, and it produces consciousness by biological processes. We will not be able to do that artificially until we know how the brain does it and we can then duplicate the causal powers of the brain. Perhaps we can do it in some completely different medium as we build artificial hearts in a completely different medium from muscle tissue, but at present we do not know enough about the brain to build an artificial brain. [22][1]

A similar remark was given even earlier by Shannon & McCarthy.

> A disadvantage of the Turing definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli.... With a suitable dictionary such a machine would surely satisfy Turing's definition but does not reflect our usual intuitive concept of thinking. [24, p. vi]

Where Searle [19] makes a great leap though is when he describes the instructions, or the program in his Chinese room example (similarly the dictionary for Shannon & McCarthy [24] or the tree of sensible replies for Block [25]). They all describe a problem of the TT which is indeed real and offer conceptual examples that would pass a TT. But are these examples doable in practice? Is such a detailed and exhaustive list of instructions possible? It is extremely unlikely [26].

Indeed conversations do not follow any set of rules as strictly as one might assume. Sure, one might start a conversation with hello and end it with goodbye, as politeness would indicate. In fact, philosophers and linguists have attempted to produce a set or rules that would explain how we converse with others, starting the field of conversational pragmatics. Grice is one of them [27]. He came up with the *Cooperation Principle* (the idea that conversational partners try to cooperate during a conversation), and with four maxims that are a direct consequence of this principle: 1) the maxim of quality focusing on the truthfulness and certainty of an information given, 2) the maxim of quantity focusing on the amount of information given (neither too little nor too much), 3) the maxim of relation suggesting that participants in a conversation try to remain relevant and 4) the maxim of manner focusing on how the information is given (briefly, clearly, orderly and without ambiguity). Yet again these are not rules, they are more like expectations that each agent in a conversation has of the production of the other agents. Speakers very often do not follow them strictly: "He's a shark" is obviously not a statement that must be taken literally, but it instead conveys the idea that "he" will take everything he can from you. Grice was well aware of that and considered this practice in conversations

[1]Note that Turing himself did not seem to be against that idea. He indeed claims multiple times that in order to pass the Turing Test the best strategy would be to learn from the way humans think, though he does not dismiss the possibility that other strategies could work as well (see [23, p. 472]). The main difference between Turing and Searle being that Turing suggests this can be done at the software level while Searle considers it to only be possible at the hardware level.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

to be "opting-out" of the maxims: a deviation from the maxims still within the context of a cooperation. And then there are actual violations of the maxims in the cases where participants in a conversation no longer try to cooperate: for example lying in a conversation would be a violation of the maxim of quality regarding to the truthfulness of the information given, which would be done without the knowledge of the other conversational partner: thus voluntarily removing oneself from the act of cooperation in the conversation. Still, violating these maxims does not make it less human, but any violation that is detected by the other partners will give rise to different inferences, and the violation itself will be considered to be a piece of information in its own right. The most important concepts that Grice offers which is of importance for chatbots is the distinction between what is *said* and what is *meant*. Take the following example: "Come in! But I do not have alcohol". At face value, it would be difficult to tell how inviting someone in would be this directly related to alcohol without any other information. Yet this sentence is easily understood and can trigger an offended reply, a disappointed one or an amused one, depending on the relationship between the two participants in the conversation. What is *meant* here is "Come inside, but there is no alcohol inside and I know you might have expected that we would share alcohol". Here the key to understand the mention of alcohol (drinking was expected) is completely implicit. While the mention of the alcohol seem to be coming out of nowhere and thus violating the maxim of relation, it is understood as being perfectly relevant within the given context, because of prior expectations about the situation.

To explain such productions, Sperber and Wilson developed the *Relevance Theory* [28]. The main idea behind it is that participants in a conversation actively search for relevance in the utterances of others. The Relevance Theory describes an utterance with optimal relevance as an utterance which has the greatest contextual effect on the listener's mental representations for the least cognitive cost (least effort in retrieving what is *meant* from what is *said*). Indeed, in the previous example what is the use of including that drinking is something expected when it's an expectation both participants in the conversation already share? This would only be making the interpretation of the sentence harder, would take more time, and would not add anything (it would not change the mental representations of the listener as this is something they would already know). Thus adding it in the utterance is irrelevant, and it remains implicit.

Because the Relevance Theory expects participants in a conversation to have an idea of what is in the mind of the others, participating in an actual conversation (at the human level) requires a Theory of Mind [29]. The Theory of Mind is the concept according to which humans (among others animals) are mindreaders. Not in the metaphysical way of course, but humans understand that other humans also think, that they have mental representations of the world that might be different or similar to their owns. There are things others do not know that we know, and there are things we do not know that they might know... It is the reason why there are questions in conversations: we understand that others might have the answers we are looking for, and we ask them to share the information they have with us. Reciprocally, the only reason why people answer questions is because they assume people who ask them do not already know the answers and will learn (their mental representations will change once the answer is given to them). Evidence indicate that humans acquire this capacity very early in their life [30], [31] and the presence or absence of this ability in other species is still a strong debate in the scientific community [32], [33], which is not entirely surprising given the difficulty in finding ways to explicitly communicate the question without ambiguity to young humans [30].

Because conversations are built around these principles, the replies given during a conversation are not fixed and will depend heavily on what each participant in a conversation believes the other knows. Thus, predefined rules, as mentioned in Searle's Chinese room (the instructions given to the man inside the room on how to match a string of symbols as a reply to another string of symbols) cannot do more than imitate and drastically reduce the range of possibilities that natural conversations have. Not only would the set of instructions be infinitely large, but it would also need constant updating to be tuned for the specific audience and for the changes in time as the natural language evolves. Thus, it is our belief that rule based AI such as ELIZA, A.L.I.C.E, and Kuki.ai will not be able to reliably pass the TT for their inability to learn from their interactions. Similarly, retrieval based systems using machine-learning in order to detect the intent still likely will not be able to reliably pass the TT as they are not able to generate new answers that would fit new situations like a human would. Only a machine learning to infer meaning and to change how it expresses itself should be able to pass the Turing Test reliably, even though currently generative AIs are less useful and more frustrating than retrieval-based AIs.

But would a judge be able to tell the difference in a TT? Is not understanding context and the mind of others enough to significantly prevent a machine from passing the TT? As we will show in the next section, the answer is yes.

V. Understanding humans

Comparing how human-chatbot conversations with human-human conversations has many benefits for the two fields of psychology and computer science. Investigating how humans behave compared to chatbots can help us make better chatbots, and investigating interactions with chatbots can give us valuable information on what humans expect of a conversational partner. And yet, despite the fact that the TT can be used as en experimental paradigm useful to explore human expectations in conversations, it is remarkably absent in international publications in the field of psychology and pragmatics. Indeed, doing a quick search on Google Scholar reveals about 33.000 entries for "Turing Test" while adding keywords from the field of pragmatics makes the search drop to below 300 entries ("Turing Test" Implicatures: 209 results,

"Turing Test" "Relevance Theory": 96 results, "Turing Test" "Cooperative Principle": 104 results), most only mentioning these topics without focusing on them.

Chatbots are still quite far from meeting human expectations of a conversational partners. surveys and studies showing that people get quickly frustrated when using them are not hard to find (see [10], [34], [35] to name only a few). An extensive survey conducted on the literature of chatbots indicates many of the current challenges they still face [36], especially regarding social characteristics of the chatbots. This feeling of frustration can be mitigated when it is made clear to the user what can be expected of the chatbot. For example, Woebot clearly sets its users' expectations beforehand which allows the users to adapt their own behavior [37]. In the case of this chatbot (which acts as a coach to help deal with anxiety and depression), the bot remains in control of the conversation at all times as the user navigates pre-defined decision trees, and in doing so it is able to carry out its task, though in cases that are too severe the user is redirected to a hotline through which they can interact with professionals to seek help. This transparency about the chatbot's abilities (along with its very sparse use of natural language understanding) allows it to be efficient in its task of helping people cope with anxiety and depression, at least for a short time (as the study did not investigate long term effects). Similar effectiveness of this chatbot seems to be observed to help control substance use [38].

Indeed, the closer the chatbot is to feeling like a human, the more users will be expecting human-like abilities in their interactions with them. It is possible to observe this effect even on the same chatbots depending on how they are introduced. For example one study can find the bot entertaining enough for the users to keep conversing with it for extended conversations despite a quality of conversations significantly lower than that with humans [39], while another can observe judges in a TT being quite perplexed when they are not made aware that the author of the messages might be a chatbot, wondering whether the person writing such messages might be "mentally ill" [40].

These situations of violating the user's expectations are common when interacting with chatbots, creating a feeling similar to Mashiro Mori's uncanny valley [41] which is a famous effect observed with robots [42] (The closest an artificial agent, robot or chatbot, gets to human behavior or appearance, the greater the expectations of humans interacting with it will be, and the greater the frustration or uncomfort if they are not meant). Still one might wonder if all other aspects remaining similar, violating such expectations would be enough to prevent a machine from passing the TT. Saygin & Cicekli [40] investigate this issue by trying to assess to what level each of Grice's maxims [27] has (or does not have) an effect on the participants responses in a TT. Their findings indicate that not all the maxims have similar influences on the answers in the TT. Indeed, violations of the maxim of manner (which deals with how information is given to the user) has no detrimental effect on the judges' perception of the humanness of the chatbot. In fact, they even observe that it has a positive effect as long as no other maxims are violated. They explain this finding by the fact that violating this maxim can produce a seemingly more emotionally loaded reply, emotions being a feature more readily (and understandably) associated to humans than to machines. Violations of the maxim of quantity was shown to have either no effect on the TT (when the maxim was violated to give too little information) or to be quite detrimental to perceived humanness (when the maxim was violated to give too much information, giving an encyclopedic feel to the reply). The difficulty with assessing the individual effect of violations of this maxim is that when violated it also has the tendency of violating the maxim of relation, which produced by far the strongest adverse effect to the feeling of humanness: the judge was left feeling like the chatbot simply did not understand the question (or did not want to talk about this topic for no understandable reason when the judges were not aware that a chatbot was present). Finally, the authors were unable to show a specific influence of the maxim of quality on the humanness of the chatbot for it also had a tendency of being violated along with other maxims.

An important difference remains between the above paper and a regular TT. In Saygin & Cicelki's paper judges were reading excerpts of conversations recorded during a Loebner prize competition and did not actually interact with the chatbots. Would users interacting with a chatbot for which the only issue would be a lack of relevance or other violations notice this flaw enough to correctly label the chatbot as a machine? We have tried to answer this question in previous papers [43]–[45] by inviting participants to play the judges in a TT. The main interest in our approach here was to test the influence of these violations only: indeed the judges participated in two conversations in a random order, being informed that one would be with a chatbot and the other with a human. In truth there was no chatbot at all. Indeed using one would have made it more difficult to test whether or not the observed differences would have been caused by the violations or by other factors related to the chatbot. Both conversations were played by the same human experimenter, each time portraying a fictive character (the same fictive character between the two conversations), except that in one conversation the experimenter was tasked to produce violations of one of Grice's maxims. Once again, the violations which had the most effect on the feeling of humanness were violations of the maxim of relation [43], [45] and violations of the maxim of quantity giving rise to an encyclopedic feeling [44]. This effect was also visible in the delay between the experimenter's utterance and the participant's turn (which is longer following a violation than following an expected reply), further indicating that these violations are indeed the cause of the observed difference. In addition, the kind of violations of the maxim of relation in these papers were slightly more subtle than the blatant violations that can often be found in chatbots: the experimenter was not allowed to use previous knowledge of the conversation in their replies, but could still answer relevantly if all the necessary information to do so was contained in the participant's last message. For example:

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Human: Do you like reading?
Experimenter: Not really no. It's not really my thing.
Human: Why not?
Experimenter: It's hard to tell. Do you have any brothers or sisters?

In the experimenter's first reply they are allowed to give a relevant answer, but in their second answer they were not allowed to use the knowledge that the topic was about reading. Thus they used a generic reply instead, producing a violation of the maxim of relation.

This type of violation is very easy to get on any chatbot currently available. Asking generic questions such as "Why?" or "Why not?" requires the chatbot to use the context of the message (the conversation's history) to be able to reply correctly. In the human's second question here, they assume that their reader still has in mind the topic of the conversation (reading is not the experimenter's thing), while the experimenter must infer that what the participant *means* is "Why is reading not really your thing?" when they *say* "Why not?".

More studies need to be carried out to explore just how sensitive the TT is to even more subtle violations, but with the evidence at our disposal today, it seems highly likely that only a chatbot able to converse in a relevant way in every situation would be able to pass the TT (especially in its 3 players version: judge, machine and human, with no limits on the topics of discussion), and this would require the ability to develop an idea of what is relevant to the user, and thus for the chatbot to have a theory of mind [46]. We are not there yet [4].

VI. Conclusion

While we have only scratched the surface of the literature regarding the TT, we explored the existing literature discussing the importance of conversational pragmatics within chatbots, and we have attempted to show how the TT is a very relevant tool in evaluating the ability of chatbots to generate relevant replies in an open conversation which is (so far) not matched by any other evaluation method.

We have also discussed how the TT in its design suggests that only an agent with a theory of mind could reliably pass it, though of course it does not set any requirements on how this theory of mind is implemented.

We also believe that the TT should be more widely used in human sciences like psychology, especially in the case of studying reasoning and conversational pragmatics. It is still a tool that is extremely rarely used despite being a valuable experimental paradigm which enables experimenters to collect direct measures (the response in the TT) and indirect measures (the delay between utterances during the conversations for example). This area of research is still underdevelopped despite its great potential for fundamental and applied research. One example is to test the influence of the use of textisms (SMS language) on the cognitive cost of processing messages in a conversation [47], or to use chatbots to investigate how behaviors are influenced by different pragmatic clues in the ultimatum game [48].

Finally, some readers might object that we did not settle the issue of whether passing the TT proves that one has a mind. After all, do we need a mind to have a theory of mind?

References

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433–460, 1950.

[2] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[3] R. S. Wallace, "The anatomy of alice," in *Parsing the Turing Test*. Springer, 2009, pp. 181–210.

[4] B. Jacquet and J. Baratgin, "Mind-reading chatbots: We are not there yet," in *International Conference on Human Interaction and Emerging Technologies*. Springer, 2020, pp. 266–271.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu, "Alime chat: A sequence to sequence and rerank based chatbot engine," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 498–503.

[7] M. Dhyani and R. Kumar, "An intelligent chatbot using deep learning with bidirectional RNN and attention model," *Materials Today: Proceedings*, vol. 34, pp. 817–824, 2021.

[8] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *Twenty-Fourth international joint conference on artificial intelligence*, 2015.

[9] B. Wei, S. Lu, L. Mou, H. Zhou, P. Poupart, G. Li, and Z. Jin, "Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7290–7294.

[10] M. Jain, P. Kumar, R. Kota, and S. N. Patel, "Evaluating and informing the design of chatbots," in *Proceedings of the 2018 Designing Interactive Systems Conference*, 2018, pp. 895–906.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[12] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[13] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[14] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic turing test: Learning to evaluate dialogue responses," *arXiv preprint arXiv:1708.07149*, 2017.

[15] A. Venkatesh, C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, M. Cheng, B. Hedayatnia, A. Metallinou *et al.*, "On evaluating and comparing open domain dialog systems," *arXiv preprint arXiv:1801.03625*, 2018.

[16] M. de Gennaro, E. G. Krumhuber, and G. Lucas, "Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood," *Frontiers in Psychology*, vol. 10, p. 3061, 2020.

[17] S. Devaram, "Empathic chatbot: Emotional intelligence for empathic chatbot: Emotional intelligence for mental health well-being," *arXiv preprint arXiv:2012.09130*, 2020.

[18] J. Lassègue, "What kind of turing test did turing have in mind?" *Tekhnema: Journal of Philosophy and Technology*, vol. 3, pp. 37–58, 1996.

[19] J. R. Searle *et al.*, "Minds, brains, and programs," *The Turing Test: Verbal Behaviour as the Hallmark of Intelligence*, pp. 201–224, 1980.

[20] S. Harnad, "What's wrong and right about searle's chinese room argument?" 2001. [Online]. Available: http://cogprints.org/4023/

[21] J. Hawkins, M. Lewis, M. Klukas, S. Purdy, and S. Ahmad, "A framework for intelligence and cortical function based on grid cells in the neocortex," *Frontiers in Neural Circuits*, vol. 12, p. 121, 2019.

[22] D. Turello, "Brain, mind, and consciousness: A conversation with philosopher john searle," 2015. [Online]. Available: https://blogs.loc.gov/kluge/2015/03/conversation-with-john-searle/

[23] A. P. Saygin, I. Cicekli, and V. Akman, "Turing test: 50 years later," *Minds and machines*, vol. 10, no. 4, pp. 463–518, 2000.

[24] C. E. Shannon, J. McCarthy *et al.*, *Automata studies*. Princeton University Press Princeton, NJ, 1956, vol. 11.

[25] N. Block, "Psychologism and behaviorism," *The Philosophical Review*, vol. 90, no. 1, pp. 5–43, 1981.

[26] D. McDermott, "On the claim that a table-lookup program could pass the turing test," *Minds and Machines*, vol. 24, no. 2, pp. 143–188, 2014.

[27] H. P. Grice, "Logic and conversation," in *Syntax and Semantics 3: Speech arts*, P. Cole and J. L. Morgan, Eds. New-York: Academic Press, 1975, pp. 41–58.

[28] D. Wilson and D. Sperber, *Relevance Theory*. Oxford: Blackwell, 2004, pp. 607–632.

[29] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[30] J. Baratgin, M. Dubois-Sage, B. Jacquet, J.-L. Stilgenbauer, and F. Jamet, "Pragmatics in the false-belief task: let the robot ask the question!" *Frontiers in Psychology*, vol. 11, p. 3234, 2020.

[31] I. Bretherton, S. McNew, and M. Beeghly-Smith, "Early person knowledge as expressed in gestural and verbal communication: When do infants acquire a "theory of mind"," *Infant social cognition*, vol. 333, p. 73, 1981.

[32] C. Krupenye and J. Call, "Theory of mind in animals: Current and future directions," *WIREs Cognitive Science*, vol. 10, no. 6, p. e1503, 2019.

[33] D. C. Penn and D. J. Povinelli, "On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 731–744, 2007.

[34] P. B. Brandtzaeg and A. Følstad, "Chatbots: changing user needs and motivations," *Interactions*, vol. 25, no. 5, pp. 38–43, 2018.

[35] E. Luger and A. Sellen, "" like having a really bad pa" the gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5286–5297.

[36] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design," *International Journal of Human–Computer Interaction*, pp. 1–30, 2020.

[37] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "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 Ment Health*, vol. 4, no. 2, p. e19, June 2017.

[38] J. J. Prochaska, E. A. Vogel, A. Chieng, M. Kendra, M. Baiocchi, S. Pajarito, and A. Robinson, "A therapeutic relational agent for reducing problematic substance use (woebot): Development and usability study," *J Med Internet Res*, vol. 23, no. 3, p. e24850, March 2021.

[39] J. Hill, W. Randolph Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations," *Computers in Human Behavior*, vol. 49, pp. 245–250, 2015.

[40] A. P. Saygin and I. Cicekli, "Pragmatics in human-computer conversations," *Journal of Pragmatics*, vol. 34, pp. 227–258, 2002.

[41] J. Vallverdú, H. Shah, and D. Casacuberta, "Chatterbox challenge as a test-bed for synthetic emotions," in *Creating Synthetic Emotions through Technological and Robotic Advancements*. IGI Global, 2012, pp. 118–144.

[42] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, "All robots are not created equal: the design and perception of humanoid robot heads," in *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, 2002, pp. 321–326.

[43] B. Jacquet, J. Baratgin, and F. Jamet, "Cooperation in online conversations: the response times as a window into the cognition of language processing," *Frontiers in psychology*, vol. 10, p. 727, 2019.

[44] B. Jacquet, A. Hullin, J. Baratgin, and F. Jamet, "The impact of the gricean maxims of quality, quantity and manner in chatbots," in *2019 international conference on information and digital technologies (idt)*. IEEE, 2019, pp. 180–189.

[45] B. Jacquet, J. Baratgin, and F. Jamet, "The gricean maxims of quantity and of relation in the turing test," in *2018 11th international conference on human system interaction (hsi)*. IEEE, 2018, pp. 332–338.

[46] B. Jacquet and J. Baratgin, "Towards a pragmatic model of an artificial conversational partner: opening the blackbox," in *International Conference on Information Systems Architecture and Technology*. Springer, 2019, pp. 169–178.

[47] B. Jacquet, C. Jaraud, F. Jamet, S. Guéraud, and J. Baratgin, "Contextual information helps understand messages written with textisms," *Applied Sciences*, vol. 11, no. 11, 2021.

[48] B. Beaunay, B. Jacquet, and J. Baratgin, "A selfish chatbot still does not win in the ultimatum game. [accepted]," in *6th International Conference on Human Interaction & Emerging Technologies: Future Systems*. Cham: Springer International Publishing, October 2021.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

To Digital Technologies of Patent Processing for Development of Critical Products

Vasyl Gorbachuk
Department of Intelligent Information Technologies
V.M.Glushkov Institute of Cybernetics of the NASU
Kyiv, Ukraine
GorbachukVasyl@netscape.net

Serge Gavrilenko
Department of Intelligent Information Technologies
V.M.Glushkov Institute of Cybernetics of the NASU
Kyiv, Ukraine
S.A.Gavrilenko@nas.gov.ua

Gennady Golotsukov
Department of Intelligent Information Technologies
V.M.Glushkov Institute of Cybernetics of the NASU
Kyiv, Ukraine
Golotsukov@nas.gov.ua

Dmytro Nikolenko
Department of Intelligent Information Technologies
V.M.Glushkov Institute of Cybernetics of the NASU
Kyiv, Ukraine
Nikolenko@nas.gov.ua

Abstract − The current picture of digitalization of the world's economy is quite accurately conveyed by the processes of anti-epidemic measures, distance learning and work, development and use of appropriate vaccines. While rich countries can successfully provide their residents with anti-epidemic remedies based on modern information or communication and biological technologies, poor countries usually need international assistance in the development and application of such technologies. The development of advanced technologies requires not only entrepreneurial activity, but also the social organization and institutional capacity inherent in rich countries. The work on empirical data highlights the current problem of the relationship between antitrust regulation and regulation of patent funds.

Keywords − pandemic; network; register; distributed technolgy; digital organization.

I. INTRODUCTION

Since 2015, the V.M.Glushkov Institute of Cybernetics (GIC) of the National Academy of Sciences of Ukraine (NASU; Національної академії наук України, НАНУ) has been developing and maintaining the Distributed Information Technology (Розподілену інформаційну технологію, PIT; RIT) to support scientific and organizational activities (науково-організаційної діяльності, НОД; NOD) of the NASU (known as PIT НОД НАНУ, or RIT NOD NASU), which covers tens of thousands of customers who, in turn, produce about a half of the objects of intellectual property rights (IPR) in Ukraine. In the process of development, the subsystem of maintaining the register of IPR of the NASU for the Center for Intellectual Property Research and Technology Transfer of the NASU (CIPRTT NASU) is being developed [1; 2]. The pandemic forces us to review previous developments in order to adapt to new conditions, as well as to better realize the role of this subsystem and understand the directions of its further development [3]. First of all, the RIT NOD NASU proved to be convenient for remote work, the scale of which also makers us aware of novel challenges to contemporary academic and educational activities in Ukraine and around the world [4−7]. During the period of global remote work since the pandemic started in 2020, the capitalizaion of Zoom Video Communications startup (established by Eric Yuan in 2011) exceeded that of known Exxon Mobil (established by John Rockefeller in 1870): data are the oil of the future. As a result, the issues of ownership and property rights protection arise in the field of data, data processing and data application. The modern complex products (goods and services) with high valued added are based on such property rights.

II. PROBLEM IDENTIFICATION AND INPUT DATA

On April 19, 2021, at a traditional conference call chaired by the President of Ukraine, he stressed that Ukraine urgently needs to prepare a state-of-the-art laboratory for the development of vaccines and drugs as soon as possible. For that purpose, the state budget for 2021 included UAH 100 million (about USD 3,6 million). The Ministry of Health is looking for ways to solve the problem stated. The President noted: «We need to gather our brightest heads and provide them with working conditions at the level of Western specialists. Ukraine needs such a laboratory not only during this pandemic. In general, this may be a response to the challenges of the future». It should be noted that some Western specialists in question are former Ukrainian scientists (from the NASU mainly) currently working for leading transnaional corporations in the field of biotechnologies. Such corporations are constantly developing a lot of patents: for instance, the known IBM Corporation is producing more patents than any other business in the world each year since 1994.

According to the Department of Communications of the Secretariat of Cabinet of Ministers of Ukraine, on April 22, 2021, under chairmanship of the Prime Minister of Ukraine and with participation of heads of leading academic institutions and institutes, a meeting was held to establish vaccine development infrastructure in Ukraine. The Prime Minister stressed that the situation with shortage of vaccines against COVID-19 in the world has demonstrated the need to create (or recreate) its own corresponding laboratory in Ukraine: «This is a strategic issue, as the construction of modern infrastructure on world standards

will allow Ukraine to restore its scientific potential for the development of vaccines and develop drugs to protect against various diseses». During the meeting, the Deputy Minister of Health of Ukraine – Chief State Sanitary Doctor noted that one of the options considered by the Ministry of Health is the creation of a laboratory on the basis of the State Institution (SI) «L.V.Gromashevsky Institute of Epidemiology and Infectious Diseases» of the National Academy of Medical Sciences of Ukraine (NAMSU): «In parallel, we have a plan to create an open academic space at the Institute that will unite all scientists for conducting research». It is assumed that such an open cluster will give Ukraine an impetus to resume production of immunobiological drugs, in particular, for the treatment of cancer. Following the meeting, the Prime Minister of Ukraine instructed to work on the realted infrastructure issues for two weeks and identify ways to finance the establishment of a high-level laboratory in Ukraine.

The meeting was also attended by the Minister of Health of Ukraine, Head of the National Research Foundation (established in 2018), Head of the Department of the D.K.Zabolotny Institute of Microbiology and Virology of the NASU (established in 1928), Chairman of the Board of Private Joint Stock Company «Research and Production Company DIAPROPH-MED» (diaproph.com.ua; established in 1999), Director of the Institute of Cell Biology of the NASU (established in 2000 in Lviv), Director of the O.V.Palladin Institute of Biochemistry of the NASU (established in 1925 in Kharkiv), Director of SI «L.V.Gromashevsky Institute of Epidemiology and Infectious Diseases» of the NAMSU (established in 1896), Director of the Institute of Molecular Biology and Genetics of the NASU (established in 1973).

Despite the availability of sufficient financial resources, relevant highly qualified personnel, the necessary industrial capacity, the production of vaccines in Ukraine is constrained by the problems of IPR [8]. Such problems are evidenced by the place 105 of Ukraine among 129 countries in the world ranking of IPR. The neighboring countries have the better IPR ratings: Slovakia − 39, Hungary − 43, Romania − 54, Bulgaria − 55, Poland − 58, Turkey − 66, Georgia − 73, the Russian Federation − 88; Moldova has a rating of 111, and Belarus is not in the list of countries studied [9].

At the conditions of pandemic, the success of health care system depends significantly on the regular implementation of effective research tools and population monitoring [10−13]. Decisions on which the people's lives depend are regularly made not only by individuals, but also by legislative and executive institutions of power that implement the function of state health care system. These decisions take into account the possibility of preserving and prolonging human life by means of scarce resources (for example, financial, human, temporal resources). Such decisions are made by government institutions to implement the functions of defence and security, law and order, macroeconomic management, protection of property rights. The contemporary international data of total vaccination (TV), case fatality rate (CFR), and the derived ant-epidemic reaction index AERI = TV/CFR (Tables 1, 2) confirm the problem statement above.

TABLE 1. THE ANTI-EPIDEMIC REACTION INDEX FOR UKRAINE AND NEIGHBORING COUNTRIES AS OF APRIL 24, 2021 [14]

Country / Indicator	TV (%)	CFR (%)	AERI
Belarus	3.5	0.7	4.9
Bulgaria	10.3	4.0	2.6
Georgia	0.9	1.3	0.7
Hungary	53.1	3.4	15.5
Moldova	3.0	2.3	1.3
Poland	26.6	2.1	12.6
Romania	24.3	3.0	8.1
Russian Federation	12.4	2.6	4.8
Slovakia	15.9	2.4	6.7
Turkey	25.0	0.8	30.1
Ukraine	1.2	2.1	0.6

TABLE 2. THE ANTI-EPIDEMIC REACTION INDEX FOR UKRAINE AND NEIGHBORING COUNTRIES AS OF JUNE 12, 2021 [14]

Country / Indicator	TV (%)	CFR (%)	AERI
Belarus	9.79	0.73	13.41
Bulgaria	22.65	4.26	5.32
Georgia	6.25	1.43	4.37
Hungary	97.56	3.71	26.30
Moldova	13.1	2.41	5.44
Poland	63.53	2.59	24.53
Romania	44.01	2.95	14.92
Russian Federation	22.43	2.42	9.27
Slovakia	53.75	3.18	16.90
Turkey	39.64	0.91	43.56
Ukraine	3.88	2.36	1.64

During the 50-day period from April 24, 2021, to June 12, 2021, the AERI increased the most in Georgia (by 524 %), Moldova (by 318 %), Belarus and Ukraine (by 174 %). Taking into account the visible organizational changes of those countries in that period (for instance, in the course of pandemic, the Parliament of Ukraine on May 20, 2021, approved the fourth Minister of Health Care), one can conclude the proper organization and management are of significant importance for answering modern challenges and saving lives. During the period considered the least change of AERI took place for Turkey (45 %) because of its relatively rapid vaccination and stable low case mortality. The case mortality rate decreased in Romania and the Russian Federation.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

At the beginning of the 21-st century, the number of patent awards is growing rapidly. Against the background of increasing number of patents, the volume of litigation between competitors among a wide variety of technologies, which is not always socially useful: the overlap of IPR complicates the commercialization of innovative products [15] and the orientation in patent thickets for potential inventors [16]. As a result, the price of final products and the duration of entry of new products into market increase, reducing social welfare. To solve the patent-thicket problem, patent funds or pools (formal or informal organizations) have been proposed, where IPR subjects share patent rights with each other and with third parties [16]. Empirical observations and practical examples show that the current regime of patent pool regulation should focus not only on the extreme cass of perfect substitutes and complements, but also on intermediate cases.

In this mode, independent licensing, royalty controls and grant-back policies (in licensing agreements that require the licensee to inform about the improvements made by the licensor's license object, which recognize the right to filing a corresponding patent applicaion) have an important role. Patent pools have become an economically significant institution. For example, in 2001, sales of devices based on patent pools (pool associations) in whole or in part exceeded 100 billion dollars. If the abovementioned proposals were accepted, then at the beginning of the 21-st century the role of patent pools could approach the role of patent association agreements in many essential branches of manufacturing industry at the beginning of the 20-th century.

III. CONTEMPORARY HEALTH CARE ORGANIZATION

The countries, having a national health service or national health insurance, usually allow government agencies to make decisions on orders for new products (pharmaceuticals, therapies and medical devices). As a rule, innovations, that promote therapeutic treatment with a lower probability of early death within a certain risk group, predominate. Because such innovations involve additional costs, cost-cutting innovations are often neglected.

For instance, providing a multimillion-dollar mobile coronary unit can help treat patients with heart attacks quickly, significantly reducing the number of lethal cases on the way to a hospital. The long-term drug therapy for patients with hypertension who use antihypertensive drugs can also prevent heart attacks, significantly supporting the economy of research and development (R&D) in pharmaceuticals. The installation of dialysis equipment for patients with chronic renal failure promotes R&D in manufacturing medical equipment.

The earlier the risks of disease can be identified, the more effective process of preventive measures or treatments can be. Life-saving costs are borne not only in the field of health care: in the field of transport, in locations with a higher number of road accidents, there are issues of improving the quality of roads (not only road surface), which must be met by local communities and government agencies; in the field of transport, there are also issues of proper arrangement of roads within residential areas in order to reduce speed of vehicles and to conduct permanent video surveillance. Of course, the practical realization of responses to those issues involves certain expenditures of the local or state budget.

In the field of environmental protection, there are questions about ensuring the levels of security systems for such dangerous enterprises as a nuclear power plant or a chemical plant; if the level of security system is insufficient, an accident can occur threatening the lives of millions of people. One of the consequences of the 1986 Chornobyl disaster was an increase in cancer cases, especially in Ukraine and Belarus. In thermal power plants burning coal, there are questions about the cost of filters that can contain sulfur dioxide and other harmful emissions into the atmosphere. Such emissions increase the incidence of respiratory diseases among the people.

In all the above issues, government institutions cannot make rational decisions without a comprehensive and accurate assessment of future gains (and losses) caused by the implementation of a particular project, as well as without comparison of such gains with the present value of cost flow associated with the project. It is important for decision makers to measure gains and costs in the same units. Since project costs are usually measured in monetary terms, it makes sense to measure all gains in monetary terms as well. Therefore, the prolongation of life or improvement of human health, caused by the implementation of project should also be measured in monetary units. Since it is difficult to assess the status of health and life for a human being in monetary units, economists have developed alternative methods for assessing the state of health and human life.

Different approaches to economic health assessment compare the benefits of medical intervention with the costs of this intervention. Gains from intervention can be measured by physical units on a one-dimensional scale, monetary units, units of cardinal utility function reflecting the multidimensional concept of health in a scalar index.

Since the 1990-s, several states in the world have taken steps to increase competition for their health care insurers, hoping to improve efficiency in their fields of health insurance and health care. Then the generalized equality of price and marginal cost will mean that competing health insurers will charge a high premium for high risks and at the same time a low premium for low risks: high risks are characterized by a relatively high expected cost of treatment due to the high probability of disease. As the state wants all its citizens to be provided with health insurance, there are issues of risk selection in health insurance markets.

One way to ensure an universal access to health insurance is to provide targeted subsidies to the poorer strata of population to cover insurance premiums. In practice, governments regulate premiums, effectively eliminating the dependence of premium charged by an insurer on risk: in the United States, for example, premium regulation applies so called a community rating. In addition, the German and Swiss regulators typically require insurers to follow an open enrollment policy and accept all the applications. In the United States Medicare gives its beneficiaries a choice between the

Medicare Plan itself and competing health care plans, which receive a capitation payment for every policyholder.

Therefore, in the countries mentioned, there is a natural incentive to risk selection. If each person pays the same insurance premium, the insurer will expect losses with high-risk individuals (of high-risk type) and gains with low-risk individuals (of low-risk type). The economic viability and balance of any health insurer presumes a sufficient number of low-risk persons insured: insurers try to attract as many such persons as possible. Therefore, under the pressure of competition, all the insurers will take part in the collection of cream on market (cream-skimming), attracting favorable risks and avoiding adverse risks.

Risk selection can take many forms. On the one hand, health insurers can implement direct risk selection by influencing who would sign the insurance contract: for example, the insurers may not pay their attention to the draft contract from a high-risk person. Individuals who are likely to need some medical care may be asked to sign a contract that provides additional discount services or outright payments. On the other hand, indirect risk selection is the development of payment packages or contracting with service provides that involve low-risk individuals but do not involve high-risk persons. Direct risk selection concerns the problem of individual access to a service, and indirect one − the quality problem.

The both forms of risk selection will occur only when insurers or their consumers possess information about individual health care costs. Direct risk selection require insurers to be able to observe the characteristics of physical persons that correlate with their expected costs − gender, age, social behaviour, and so on. For instance, if healthy people use the Internet more often, the risk selection strategy is to market insurance contracts online: this way people do not have to know their type of risk. However, people need to know their type of risk in indirect risk selection: for example, people need to know the likelihood that they will use certain services. Such personal data allow insurers to develop payment packages and attract service providers with different types of risk.

Direct and indirect risk selection can take place simultaneously: measures that exclude one selection should not affect another. For instance, if the benefit package is strictly regulated, preventing indirect risk selection, insurers may remain interested in attracting favorable risks and thus turn to another risk selection − direct risk selection. On the contrary, if insurers do not have the ability to select risks directly, they retain the incentive to develop a benefit package that attracts low risks and avoids high risks. Indirect risk selection is closely related to the phenomenon of unfavorable (adverse) selection in insurance markets, which happens when policyholders have more information about their type of risk in comparison with their insurers. This phenomenon takes place regardless of the actions of state. At the same time, indirect risk selection is an implication of state regulations for premiums.

To avoid unwanted behavior by insurers in selecting risks, certain measures can be taken based on the assumption of compulsory health insurance, forcing them to cover high risks by means of low risks.

First, open enrollment guarantees that some insurers will take some high risks. At the same time, legislation, regulation and reporting may prevent obvious opportunities for direct risk selection: for example, the law may limit the insurer's financial and other benefits from taking low risks.

Second, the measure against indirect risk selection is the regulation of benefit package. On the one hand, lower bounds of benefits can be envisaged, forcing insurers to offer benefits that are important for high risks (say, for the treatment of different types of diabetes). On the other hand, upper bounds of payments may prevent insurers from including low-risk services (say, fitness center services) in their contracts. In addition, certain types of payments, that are convenient for risk selection, can be regulated by separate provisions. However, the payment package includes supply of services from specific partners provided for in the contract (say, subcontractors), which may be selected by the insurer in question. Such selection is especially important in Managed Care: for example, by involving many sports medicine professionals, the insurer can count on the attention of healthy lifestyle advocates (low-risk consumers).

Third, the measure of creating incentives via additional payments to insurers, who take high risks, and imposing financial sanctions to insurers, who skim creams (favorable risks), is a risk adjustment scheme (RAS). The payments mentioned depend on such characteristics observed as age and gender. The measure of reimbursing the share of actual costs for medical treatment is a cost reimbursement scheme (CRS). The idea of CRS is to reduce gains from risk selection by decreasing the impact of costs on the profits of insurers. At the same time, the CRS reduces incentives of insurers to control their costs.

The RAS and CRS can be substantiated by modeling risk selection. First of all, due to various reasons insurers may differ in their terms of insurance for population, the RAS and CRS can create a competitive system where the favorable risk structure of an insurer does not give her a starting advantage. Besides, the health insurance market may be destabilized as new insurers enter the market and move from high to low risks. The RAS and CRS can reduce differences of insurers in premiums, thereby reducing incentives to the movement (transition).

The insurers, entering the market, can insure mostly low risks by facilitating more frequent changes of insurers by consumers (policyholders) and mixing the overall health insurance market. Because insurers, that have entered the market earlier, would appear at high risks, they eventually have to increase their premiums or file for bankruptcy. In such circumstances, insurers will have no incentive to invest in proving effective payments.

Indeed, there is evidence of higher low-risk mobility in the German health insurance market, based on a comparison of the health care expenditure (HCE) of those who change insurers and those who do not change their insurer: depending on age categories, people, who changed insurers, had on average 45−85 % less HCE than the HCE of those who did not change insurers. Studies, based on the German socioeconomic panel, have shown that (adult) people, who remained loyal to their

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

insurer, had significantly worse health status than people who changed insurers. In the United States, there is a case of Harvard University's decision to increase employers' contributions to insurance premiums if employers did not choose the cheapest option (Health Maintenance Organization (HMO) plan).

Types of risk began to be identified during the year: those who switched from the most expensive insurance plans to HMOs had a mean age of 46 years and were 9% higher in HCE than the overall average HCE; those who remained on expensive insurance plans had an average age of 50 years and a 16% higher HCE compared to the general average HCE. The rapid loss of low risks by broad insurance plans forced the experiment to stop.

Thus, the RAS and CRS can help ensure a level playing field during the transition to a competitive market and the stabilization of health insurance market. In the absence of schemes such as the RAS and CRS, the market may lose the most efficient insurers. For actuaries and other financial professionals, risk adjustment means the accrual of a premium or per capita payment in proportion to the expected expenses of an individual or group. The RAS is based upon risk adjusters – the observed characteristics of individuals. The development of RAS and the search for appropriate risk adjusters require empirical testing of their ability to predict HCE.

Socio-demographic variables can be risk adjusters. Since age and gender have a relatively small explanatory power, other socio-demographic variables were studied – marital status, retirement status, disability status, educational level, income level. Data from the German health insurance funds showed that elderly pensioners with disabilities have significantly higher HCE. In addition, higher HCEs are revealed by single retirees and low-income individuals.

HCE in previous periods is an obvious indicator of morbidity: an increase in HCE leads to an increase in HCE in the next period by 20−30%. At the same time, the explanatory capacity of HCE should be weighed against the weakening of person's incentives to reduce her costs, because higher current HCE will to some extent be compensated to the person later. It is through HCE that insurers try to identify favorable risks, and there may not be better risk adjusters. Prescription medications in previous periods have predicted the value of HCE.

The morbidity can be measured by gathering available diagnostic information to identify chronically ill patients and to classify individuals according to their expected HCE. This classification can be done by various methods. The empirical studies show that diagnostic information gives an accurate prediction of HCE values. In turn, the corresponding gathering of information can be expensive. Because insurers have an interest in beneficial diagnoses for their policyholders, they are also interested in the ability to interpret relevant information – upcoding: insurers can encourage their policyholders to consult with doctors more often to select as many diagnoses as possible. Many countries and health care systems use diagnostic information to determine the reimbursement to a service provider, revealing the necessary data. For processing and analysis of these data, software implementations of construction for classifiers, allocation of informative features, processing of heterogeneous medical and biological variables for carrying out scientific research in the field of clinical medicine are developed.

IV. TO DIGITAL ORGANIZATION OF PATENT POOLS

At the beginning of the 21-st century, biomedical research communities have shown great interest in developing patent funds for biomarkers of cancer, patents for HIV/AIDS-related diseases (human human immunodeficiency viruses (HIV) and acquired immunodeficiency syndrome (AIDS)) and severe acute respiratory syndrome (SARS) or acute respiratory viral infection, as well as biotechnologies used in crop and livestock agriculture (animal cloning).

It should be noted that in the 1990-s, the GIC of the NASU began developing the first HIV/AIDS registers in Ukraine.

Although patent pools have been successfully developed in the basic manufacturing and electronics industries for decades, such pools are increasingly seen as potential solution to common patent licesing issues in biotechnology-related industriess. The Organization for Economic Cooperation and Development (OECD) has outlined the development of biomedical patent funds as an area for future research. State policy on patent pools gradually shifted from the extreme approach of non-interference (laissez-faire) in the early 20-th century to the approach of strict control in the middle of century.

Creation and use of inventions and other objects of IPR (OIPR), patent protection of scientific and technical results and the efficiency increase of their commercialization is one of the tasks of scientific and organizational activities (hereinafter − NOD) of the NASU to ensure modern academic research (AR) in the NASU. The subsystem of maintaining the register of OIPR (SMROIPR) of the RIT NOD NASU is directed on information and communication support of the decision making process for the task. Within this subsystem, inventions and other results of AR implementation are considered as OIPR, officially registered with certain government agencies. The information about such OIPR, in particular, about their conditions, is stored in the register of OIPR (ROIPR). The SMROIPR of the RIT NOD NASU automates the work with the ROIPR as an important data source for the NASU in general. The SMROIPR significantly simplifies the accumulation of information on OIPR, as well as the search, analysis and use of the SMROIPR data.

The SMROIPR provides the implementation of technological procedures on electronic documents (EDs) containing the SMROIPR data. Those procedures are carried out by the following SMROIPR user groups:

employees of divisions of academic institutions (AIs) of the NASU, responsible for technology transfer, innovation activity and OIPR;

Scientific Secretaries of the NASU or specialists of their departments, authorized to support the conduct of ROIPR;

employees of the Presidium of NASU, authorized to support the conduct of ROIPR;

employees of the CIPRTT NASU.

The abovementioned users are the RIT entities. Each of them is assigned a certain role or certain roles (according to job responsibilities) in the SMROIPR as a part of the RIT NOD NASU. The SMROIPR provides support to users with the OIPR cards from AIs of the NASU, containing (input or primary) information from documents on current laws, administrative instructions and regulations, in particular, with:

notifications on submission of applications for the creation of OIPR;

applications for registration of OIPR;

information on the protecting document (patent, certificate), copyright, and others;

information on accounting for intangible assets;

information on contractual relations with the creators (inventors, authors) of OIPR;

other information (applications and decisions on the issue of protecting documents, the creation and use of OIPR).

The incoming (primary) documents of SMROIPR include the Forms 1, 2, 3 on intellectual property (IP), according to the Order № 469 of State Statistics Committee of Ukraine (SSCU) of July 8, 2004 amended by the Order № 342 of SSCU of November 1, 2005:

Form IP-1 «Journal of registration of applications for inventions, utility models, industrial designs, layouts (topographies) of integrated circuits submitted in Ukraine»;

Form IP-2 «Journal of registration of applications for inventions, utility models, industrial designs, layouts (topographies) of integrated circuits submitted to the competent authorities of foreign states»;

Form IP-3 «Journal of registration of used inventions, utility models, industria; designs, layouts (topographies) of integrated circuits».

Reporting forms of the documents above are source EDs of SMROIPR, which are automatically created by means of the subsystem.

Besides, the reporting EDs supported in SMROIP include the Forms VII-1, VII-2, VII-3, VII-4 (Annex to the Order № 654 «On preparation of reports on the activities of academic institutions and the Report on the activities of NASU in 2018» of Presidium of NASU of November 20, 2018):

Form VII-1 «Results of inventive work, creation and use of objects of intellectual property rights in 2018»;

Form VII-2 «Agreements for the use of objects of intellectual property rights»;

Form VII-3 «Applications for the issue of protecting documents»;

Form VII-4 «Decision on the issue of protecting documents».

In addition, the SMROIPR supports receiving of reporting documents (tables) for the CIPRTT NASU:

OIPR created in the AIs during the reporting year and in previous years if used during the reporting year;

the contest on inventive activity during the reporting year;

the main indicators of AIs of the NASU for development, protection and use of IP during the reporting year.

The automated workplace (hereinafter − workstation) of an employee of division of AI of the NASU, responsible for technology transfer, innovation activity and OIPR, within the SMROIP is designated for processing input (primary, secondary or derivative) documents on the OIPR of AI of the NASU and obtaining output (reporting) EDs. The guideline [1] describes the appropriate actions of such an employee as the following technology procedures:

creating of an OIPR card;

entering in that card information on 1) notification on the creation of OIPR, 2) application for registration of OIPR, 3) protecting document (patent, certificate), copyright, and others, 4) accounting for intangible assets, 5) contractual relations with the creators (inventors, authors) of OIPR.

The procedures above are sufficient for the initial filling of information on the OIPR of AI of the NASU.

V. THE DIGITAL ECONOMY OF INTELLECTUAL PROPERTY

The modern economy of intellectual property, in particular the patent economy, has become an important part of the global economic activity of the world's leading nations [16]. The patent ecosystem includes a wide range of participants and interests: inventors put forward an idea, after which patent agents and attorneys help them create, defend and improve a patent application, interacting with patent experts in patent offices, reviewing it for novelty and comparing with the previous techniques, as well as evaluate its usefulness or industrial applicability. After issuing a patent, or at the stage of filing a patent application, intermediary companies can buy and sell a patent or patent application. The company that owns the patent may then require other companies to license its use.

In the telecommunications industry, the 3G (3-rd Generation Partnership Project, 3GPP) standard was a strategic initiative of Nortel Networks (1895–2012) and AT&T Wireless (Cingular Wireless LLC in 2000–2006; AT&T Mobility LLC, a subsidiary of AT&T with more than 176 million subscribers by 2020). In 1998, AT&T Wireless operated the IS-136 Interim Standard (IS) wireless network in the United States. The IS-136 and IS-54 are the 2G generation mobile communication systems that developed the Advanced Mobile Phone System (AMPS), the 1G mobile system, towards the digital D-AMPS. The first commercial IS-136 and IS-54 networks in America were launched in 1993 and became widespread. D-AMPS is now considered an end-of-life (EOL) product, or an EOL product whose supplier will not provide updates to users. These networks have been replaced by the Global System for Mobile Communications (GSM), General Packet Radio Service (GPRS), and CDMA 2000 technologies.

The IS-136 is often referred to as the time division multiple access (TDMA) system, which is used in most common 2G

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

standards, including most 2G standards, including GSM. The D-AMPS system competed with GSM and code division multiple access (CDMA) systems. GPRS is a packet-oriented mobile data standard in GSM on 2G and 3G cellular networks.

CDMA 2000 (C2K, IMT Multi-Carrier (IMT MC)) is a family of 3G mobile technology standards for the transmission of voice, data and signals between mobile phones and cell sites. The CDMA 2000, developed during the 3GPP2 collaboration, is a backward compatible successor to the IS-95 (the first CDMA-based digital cellular technology developed by Qualcomm (established in 1985; QCOM listed on the NASDAQ)), adopted by the Telecommunications Industry Association. Association, TIA) and the Electronic Industries Alliance (EIA) in 1995) 2G and distributed in North America and the Republic of Korea.

IS-95 has a patented name cdmaOne. TIA was established in 1988 and accredited by the American National Standards Institute (ANSI) to develop voluntary, consensus-based industry standards for a wide range of information and communication technology (ICT) products; in 2021, TIA represented about 400 companies. Established in 1918, ANSI is a private nonprofit organization that oversees the development of voluntary consensus standards for products, services, processes, systems, and personnel in the United States; ANSI also coordinates the U.S. standards with international ones so that American products can be used worldwide.

3GPP2 (3-rd Generation Partnership Project 2) is a collaboration launched in 1998 between telecommunications associations to create a globally applicable specification of the 3G generation mobile telephone system within the framework of the International Mobile Telecommunications-2000 (IMT) project, or IMT-2000, of the International Telecommunication Union (ITU). The ITU (established in 1865 as the International Telegraph Union) is a specialized UN agency responsible for all matters related to ICT.

GPRS was introduced by the European Telecommunications Standards Institute (ETSI) in response to previous cellular packet switching technologies [7], Cellular Digital Packet Data (CDPD) and Mobile Internet (i-mode). CDPD is a territorial mobile data service that used for data transmission (at speeds up to 19.2 kbit / s) bandwidth in the range from 800 to 990 MHz, which was not used by AMPS mobile phones; AMPS have been replaced by faster services. Mobile Internet (other than wireless Internet) is offered by Japan's dominant mobile operator NTT DoCoMo (established in 1991 as a subsidiary of NTT in Japan) with the slogan «DO COmmunications over the MObile network» (docomo means «everywhere» in Japanese). Unlike the Wireless Application Protocol (WAP), a technical standard for accessing information over a mobile wireless network, i-mode covers a wider variety of Internet standards, including web access, e-mail, and a packet-switched network that delivers data.

At the proposal of the European Commission, ETSI was established in 1988 by the European Conference of Postal and Telecommunications Adaministrations (CEPT), founded in 1959 in Montreux, Switzerland, also known as the 1936 Montreux Convention on control of Turkey over the Bosphorus and Dardanelles, as well as on the regulation of the passage of warships. ETSI is an officially recognized body responsible for ICT standardization.

Founded in 1961, the European Committee for Standardization (French Comité Européen de Normalisation, CEN) is an association of European standards bodies, whose mission is to promote the economy of the European Single Market (ESM) and the entire European continent in global trade, the well-being of European citizens and the environment by providing an effective infrastructure for stakeholders to develop, maintain and disseminate agreed sets of standards and specifications.

In 1973, the European Committee for Electrotechnical Standardization (French Comité Européen de Normalization Électrotechnique, CENELEC) was established on the site of CENELCOM and CENEL, responsible for European standardization in electrical engineering. Together ETSI, CENELEC, CEN form a European system of technical standardization in telecommunications, electrical engineering and other technical fields, respectively. Harmonized ETSI, CENELEC, CEN standards are regularly adopted in many countries outside Europe that follow European technical standards.

Although CENELEC works closely with the European Union (EU), CENELEC is not an EU institution but a non-profit organization registered in Brussels under the Belgian law. CENELEC members are the national bodies of electrical standardization of most European countries. According to EU Regulation 1025/2012 of 25 October 2012, CENELEC standards are considered European Standards (ENs) for the EU and the European Economic Area (EEA), including the European Free Trade Agreement (EFTA)), which was joined by Ukraine on June 24, 2010 (at the suggestion of one of the authors of this work).

ENs are technical standards ratified by one of the three European standards organizations − ETSI, CENELEC, CEN. ENs are a key component of ESM. All ENs are developed and created by all stakeholders through a transparent, open and consensual process. The role of ETSI, CENELEC, CEN is to support EU regulations and strategies by developing Harmonized European Standards and other deliverables.

In 1997, the Wireless Research and Development Center Nortel Networks, a division of Bell-Northern Research (established in 1971), developed a vision of the entire Internet Protocol (IP) wireless network under the internal name «cellular web» (cell web). Over time, this vision has become an industry vision called «wireless Internet».

These visions drew the interest from AT&T Wireless, which about a year later launched the global third-generation 3GIP IP initiative. The initiative was first joined by British Telecom (established in 1845; BT.A listed on the LSE), France Telecom (established in 1988; ORAN on the NYSE), Telecom Italia (established in 1925; TIT listed on the BIT), Nortel Networks, followed by NTT DoCoMo, BellSouth (established in 1983 and acquired by AT&T in 2006), Telenor (established in 1855; TEL in the OSE listing), Lucent (established in 1996 and acquired by Alcatel in 2006), Ericsson (established in 1876; ERIC on the NASDAQ listing), Motorola (1928−2011),

Nokia (established in 1865; NOK on the NYSE listing), and others. Standards organizations (ETSI, 3GPP, IEEE, etc.) are also important participants in the life cycle of many essential patents, as the implementation of standards often requires certain patents and therefore the structure of licensing agreements: for example, fairly acceptable and non-discriminatory (fair reasonable and non-discriminatory, FRAND) agreements are concluded to coordinate the licensing of patents for the purpose of their application. When a company decides that another company is likely to infringe its patent rights, it can initiate a court case.

Today, each of these problems is solved by a person, almost without using the appropriate specialized software or algorithmic tools. Solving such problems is part of a broader agenda aimed at developing the newest products of the modern economy. One can identify two current problems: i) the selection of deep meaning (deep meaning) from the claims (patent claims); ii) the use of in-depth content in recommender systems that reflect patents and standards. To update these problems, one can create and maintain appropriate datasets, as well as develop and evaluate prototypes of basic systems to solve problems.

VI. STOCHASTIC GRADIENTS FOR DATA PROCESSING

Multilayer neural networks trained in back-propagation are the best example of a successful gradient-based learning method to solve the problems i), ii). For a given suitable network architecture, gradient learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional images such as handwritten letters with minimal pre-processing. A comparison of the different methods used to recognize handwritten characters on the standard handwritten number recognition task shows the advantage of convolutional neural networks specifically designed to analyze the variability of two-dimensional (2-D) shapes.

Real document recognition systems consist of multiple modules, including field extraction, segmentation, recognition, language modeling. Deep learning involves a transformer model with mechanisms of attention and weighing the impact of different parts of the input data, which is used mainly for natural language processing and is used to understand visual data. The graph transformer networks (GTNs) approach allows training of such multi-modular systems globally, using gradient methods to optimize the overall measure of system performance. A comparison of online handwriting recognition systems shows the benefits of global learning and the flexibility of GTNs. The GTN example uses convolutional neural network (CNN) character recognizers to read bank checks, combined with global learning techniques to ensure the accuracy of checks from legal entities and physical persons. Such a GTN is capable of reading several million checks a day, and is therefore suitable for commercial use.

Since the 1990s, machine learning (ML) methods, especially applied to neural networks (NNs), have played an important role in the development of pattern recognition (PR) systems: the availability of training methods has been a decisive factor in the successful use of PR for recognition continuous speech and handwriting. In 1997, the Deep Blue chess program defeated world champion Garry Kasparov.

Thanks to advances in ML and computer technology, it is possible to create better PR systems with automatic learning instead of intuitively built-in heuristics. In the case of letter recognition, one may show that the manual feature extraction can be successfully replaced with carefully designed learning machines that work directly on pixel images. In the case of understanding of documents, one may show that the traditional way of building PR systems by manually integrating separately designed modules can be replaced by a unified and principled design paradigm of GTN, which allows training of all modules to optimize the global performance criterion.

From the very beginning of the development of PR, it was known that the variability and diversity of natural data of language, (hiero)glyphs, and other types of images make it almost impossible to build a system of accurate recognition only manually. Therefore, most PR systems are built by combining automatic learning methods and heuristic algorithms. The usual (traditional) method of recognizing individual images is to divide the system into two main modules − the feature extraction module (input − primary data, output − vector of features) and the capable of learning classifier module (input − vector of features, output − numerical characteristics of the respective classes). The feature extraction module is called an extractor.

Feature vectors must meet the following requirements: 1) low dimensionality of vectors or strings of symbols; 2) high ability to compare with each other; 3) relative invariance with respect to transformations and distortions of input images (data) without changing their essence. The feature extraction process is quite specific, as it takes into account a lot of prior knowledge and design efforts, which are often done by hand. On the other hand, the classifier is a general-purpose module capable of learning. One of the main problems of the traditional method of recognition is that the accuracy of recognition is largely determined by the ability of the designer to offer a suitable set of features. This is a difficult task that has to be performed for each recognition task separately: a large amount of literature on PR is devoted to describing and comparing the relative advantages of different sets of features for individual tasks.Historically, the need for suitable extractors has been related to the fact that the learning methods used by classifiers have been limited to low-dimensional spaces with easily separated classes. This view was changed by a combination of several factors at the end of the last century. First, the availability of inexpensive machines with fast arithmetic and logic devices allows any user to rely more on more powerful numerical procedures than on algorithmic improvements. Second, the availability of large databases with broad markets and wide applications, including handwriting recognition, allows designers of recognition systems to rely more on real data than to manual feature selection. Third, the availability of powerful ML methods that can process high-dimensional input data and generate complex decision functions on this data allows the development of universal classifiers.

It can be argued that progress in the accuracy of speech and handwriting recognition systems is largely due to the greater

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

use of modern ML and large training data sets: most modern commercial optical character recognition (OCR) systems use some form of multilayer NNs, trained using the back propagation algorithm.

Among the approaches to automatic ML, one of the most successful in the NN- community is considered to be numerical or gradient learning on data: the learning machine computes a function $Y^p = F(Z^p, W)$, where Z^p is the p -th input image, W is a set of customizable parameters in the system. In the PR setting, the output Y^p is interpreted as a recognized image Z^p label, a numerical characteristic, or a probability associated with each class.

The loss function $E^p = D(D^p, F(Z^p, W))$ measures the discrepancy between the correct or desired result D^p for an image and the output Y^p produced by the system, $p = 1,..., P$, where P is the number of training samples (images).

The average loss function $E_{train}(W)$ is the average value of errors E^p on a set of labeled examples $\{(Z^1, D^1), (Z^2, D^2), ..., (Z^P, D^P)\}$, which is called a training set. In the simplest formulation, the task of learning is to find a value W that minimizes $E_{train}(W)$.

In practice, the performance of the system on the training set is not decisive: a more acceptable indicator is the level of error of the system in the area that will be used in practice.

This indicator is evaluated by measuring the accuracy of a set of samples separated from the training set − the test set.

Many theoretical and experimental works have shown that the gap between the expected level $E_{test}(W)$ of generalization error on the test set and the level $E_{train}(W)$ of error on the training set is equal to $E_{test}(W) - E_{train}(W) = k(P^{-1}h)^{\alpha}$, where h is the measure of effective capacity or complexity of the machine [17], k − some constant, $\alpha \in (0.5,1)$ [18].

This gap always narrows with growth of P.

As $E_{train}(W)$ decreases with growth of h, the gap increases and the question of the optimal value h that minimizes $E_{test}(W)$ arises. In most algorithms of ML, $E_{train}(W)$ and some gap estimation are minimized.

Formally, this is called minimizing structural risks [18], based on determining the sequence of training machines of increasing capacity, which corresponds to the sequence of subsets of the parameter space so that each subset is an extension (superset) of the previous subset.

In practice, structural risk $E_{train}(W) + \beta H(W)$ is minimized, where $H(W)$ is the regularization function, β is a constant. $H(W)$ is chosen to take large values on parameters W that belong to subsets of high capacity parameter space: minimization of $H(W)$ actually limits the capacity of the available subset of parameter space, thereby controlling the trade-off between minimizing $E_{train}(W)$ and minimizing the gap mentioned.

The general problem of minimizing a function by a set of parameters underlies many issues of computer science.

The gradient-based ML is based on the fact that it is usually much easier to minimize a fairly smooth continuous function than a discrete (combinatorial) function.

The loss function can be minimized by estimating the effect of small variations in parameter values on the loss function, which is measured by the gradient of this function by these parameters.

When the gradient vector can be computed analytically (as opposed to the numerical method that uses perturbations at a given point), efficient learning algorithms can be developed.

Analytical expressions of gradients underlie many learning algorithms with continuous parameters.

Suppose a set of parameters is a vector W whose elements are real numbers, and the function $E_{test}(W)$ (or $E_{train}(W)$) is continuous by W and also differentiated almost everywhere by W. Then the simplest optimization procedure is an iterative gradient descent algorithm

$$W_k = W_{k-1} - \varepsilon \frac{\partial E_{test}(W)}{\partial W},$$

where the step ε is a scalar constant in the simplest case. More complex procedures use a variable step, a diagonal matrix, and an inverse Hessian estimate (in Newtonian or quasi-Newtonian methods) on the place of ε. One can also use the conjugate gradient method for optimization.

Gradient learning algorithms can use two approaches to update parameters: a) the classic batch approach, where gradients are accumulated throughout the training set, and the parameters are updated after computing the exact gradient; b) the stochastic (partial or noisy) gradient approach, where the gradient is estimated on the basis of a single training sample or a small number of samples, and the parameters are updated using such an approximate gradient. Training samples can be selected randomly or according to a properly randomized sequence. In the stochastic version of gradient, the estimates are noisy, but the parameters are updated much more often than in the batch version. The empirical result of significant practical importance is that, in problems with large redundant data, approach b) is much faster than approach a) [17].

This advantage of approach b) over approach a) is easier to explain by the example of a training database consisting of two copies of the same subset. Then the accumulation of the gradient on the whole training set will lead to redundant computations. On the other hand, computing one stochastic quasi-gradient on this set would mean performing two complete iterations of learning on a small subset. This example of the advantages of approach b) can be generalized to learning

sets, where the elements are not exact copies of each other and where approach a) leads to somewhat redundant computations.

Many authors have argued that the classical Gauss – Newton or Levenberg – Marquardt methods, the quasi-Newtonian method of Broyden – Fletcher – Goldfarb – Shanno or its variant with limited-storage, different versions of the method of conjugate gradients should be used to teach NNs, instead of the gradient descent.

However, these methods proved unsuitable for training large NNs on large data sets: Gauss – Newton and Levenberg – Marquard methods require $O(N^3)$ update operations (iterations), where N is the number of parameters, and quasi-Newtonian methods – $O(N^2)$ operations; the Broyden – Fletcher – Goldfarb – Shanno method with limited storage and conjugate gradient methods require $O(N)$ update operations, but their rate of convergence relies on an accurate estimate of successive conjugate descent directions, which only makes sense in approach a).

For large data sets, the acceleration of these methods compared to the method of regular batch gradient descent is significantly inferior to the acceleration of the stochastic gradient method. Acceleration of the stochastic gradient method was not achieved by attempting to use a conjugate gradient in small batches or batches of increasing size.

When applying approach b), the axes of the parameters were scaled so as to minimize the eccentricity (the degree of difference of the ellipse from the circle) of the error surface. A common optimization procedure is a stochastic gradient or online update algorithm, which consists of updating parameters of a vector based on a single sample $p(k)$ on the iteration k using a noisy or approximate version of the averaged gradient:

$$W_k = W_{k-1} - \varepsilon \frac{\partial E^{p(k)}(W)}{\partial W}.$$

This procedure fluctuates the vector W_k around some middle trajectory, but usually on large training sets with excessive samples (those found in the recognition of language or letters), at $k \to \infty$ this vector converges faster than the usual gradient descent and second-order methods [19].

VII. The infrastructures and institutions available

The State Enterprise «Scientific and Telecommunication Centre UARNet» of the Institute for Condensed Matter Physics of the NASU (established in 1990 in Lviv on the basis of department of the Bogolyubov Institute for Theoretical Physics of the NASU (established in 1966)) is one the major backbone operators in Ukraine. The UARNet is subordinated to the Institute for Condensed Matter Physics of the NASU and is the only 100% state-owned telecommunications company in Ukraine.

The UARNet provides Internet-access to about 160 corporate clients in all regions of Ukraine, including Internet-providers, higher state authorities and higher education institutions of Ukraine, AIs of the NASU, AIs of other National Academies (National Academy of Medical Sciences, National Academy of Agrarian Sciences, National Academy of Educational Sciences, National Academy of Arts, National Academy of Legal Sciences) of Ukraine, banking and other structures. Today the UARNet has tens of thousands of users. The UARNet is a member of the Internet Traffic Exchange Network (UA-IX).

The UA-IX (ix.net.ua) is a subsidiary of the Internet Association of Ukraine (INAU). The UA-IX is operating as a Ukrainian Internet traffic exchange point since August 10, 2000, while the INAU was officially registered in November 2000. Then eight Ukrainian companies signed the agreement establishing the UA-IX. Exchange of Internet Protocol traffic is carried out in accordance with the All-All rule. The UA-IX provides connectivity of networks of Ukrainian and foreign providers and exchange of traffic between them on the shortest routes. Today the UA-IX is an example of technical and financial accessibility for all Internet operators and providers of Ukraine. The INAU (inau.ua) is a trade and industrial association that combines the efforts of its member companies in the development of Internet in Ukraine. The INAU implements projects promoting the development of Ukrainian segment in global Internet, giving its members consulting and legal support, providing a dialogue with government agencies. The INAU cooperates with corresponding committees of the Ukrainian Parliament (today the previous Chairman of INAU Board is the Member of Parliament, Deputy Chairman of the Parliamentarian Committee on Digital Transformation and Chairman of the Parliamentarian Subcommittee on Digital and Smart Infrastructure, Electronic Communications, Cybersecurity and Cyberdefense) and international public organizations, analyzes, develops and submits proposals for adjusting the legislative and regulatoty framework of Ukraine.

The INAU organized annual meetings on the following topics:

the tasks of interaction between the state, business and the public in the development of information society in Ukraine;

the ways to solve the priority problems of market development for information and communication (ICT) technologies;

creating conditions for sustainable development of the Internet domestic segment.

The INAU members are the largest Ukrainian Internet service providers, content providers, electronic and common mass media, equipment suppliers, public organizations. As of April 25, 2021, the INAU included 222 full members and 14 associate members.

The availability of UA-IX allowed Ukraine to have rank 6 in the terms of domestic (total aggregate) traffic among 124 traffic exchange points in 33 European countries as of 2010.

The UA-IX was based on the equipment of Extreme Networks and Cisco Systems, switches connected by 10 Gigabit Ethernet communication channels and two Cisco 2811 routers. Each UA-IX participant connected to a port of nearest Ethernet switch at speeds from 10 Mbps to 10 Gbps. For

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

connections 1 Gbps and above in the optical port of switch, the extension module for installation in switch was supplied by the participant. The routers provided exchange by routing tables within the network using the route information exchange protocol BGP (Border Gateway Protocol) v4. BGP belongs to the class of external gateway routing protocols (EGPs) and today is the main dynamic routing protocol on the Internet.

Each UA-IX participant inserts a BGP session with both routing servers and transmitted to them its network information. This information is then spreading among other participants, thus forming a single routing domain within the traffic exchange network. Both routing servers have an identical configuration. Failure of one of the routing servers does not cause the network malfunction. The network has a server that provides:

information about the current state of Network (Looking Glass(LG));

information on the amount of traffic passing through network in total and for each participant.

LG is a class of servers on the Internet for checking routing from a remote autonomous system. LG presumes a limited access (read-only) to the routers of organizations where LG is running. Those organizations are usually Internet service providers. Ukraine has several dozens of various LGs (lookinglass.org). The UA-IX is constantly increasing its speed capacity due to the installation of new switches, for instance, 24-port Extreme Networks Summit X650 with a bandwidth of 10 Gbps each port.

VIII. Conclusions

Developing, serial manufacturing, and mass supplying of some specific critical products within the rigid time limits appeared to be a serious challenge for many nations and international organizations during the pandemic started since 2020. Nevertheless, modern information and communication technologies seem to be capable of making a difference.

Nowadays the SMROIP system developed should satisfy some technology-driven and data-driven criteria promoting necessary data processing and modeling, first of all, the IPR-related data in question. The SMROIP function is different from that of Ukrpatent subordinated to the Ministry of Economy of Ukraine. The Ukrpatent would be an essential node in the general national IPR-related network as well as distributed AIs with their divisions and employees producing and registering OIPRs. The digital tools for management patent pools and other IPR-related pools are becoming efficient ways for developing high technology products and their marketing.

References

[1] Автоматизоване робоче місце співробітника підрозділу з питань трансферу технологій, інноваційної діяльності та інтелектуальної власності наукової установи НАН України. Технологічна інструкція користувача. Київ: Інститут кібернетики імені В.М.Глушкова НАН України, 2019. 43 с.

[2] Капіца Ю.М. *Право інтелектуальної власності Європейського Союзу: формування, інститути, напрями розвитку*. Київ: Центр досліджень інтелектуальної власності та трансферу технологій НАН України; Академперіодика, 2017. 664 с.

[3] Горбачук В., Гавриленко С., Голоцуков Г., Ніколенко Д. Економіка Internet-застосунків і цифрового контенту. *The role of technology in the socio-economic development of the post-quarantine world*. M.Gavron-Lapuszek, A.Karpenko (eds.) Katowice: Katowice School of Technology, 2020. P. 81–88.

[4] Горбачук В.М., Ткачев И.И. Стратегическая роль экономической информатики в успешном развития Евразии. *Вестник Таджикского национального университета*. 2012. 2/9. С. 68–80.

[5] Горбачук В.М. Організація функціонування мережних галузей. *Інноваційна економіка*. 2014. № 4 (53). С. 320-328.

[6] Горбачук В.М. На порозі Четвертої промислової революції. *Причорноморські економічні студії*. 2016. Вип. 8. С. 216–220.

[7] Горбачук В.М. Постіндустріальна організація державних замовлень у розвитку AUTODIN, ARPANET, PRNET, NSFNET та Інтернету. *Вісник Одеського національного університету. Економіка*. 2016. Т. 21. Вип. 8. С. 116–122.

[8] Горбачук В.М., Лещинська Л.В. Міжнародні інтеграційні процеси та вимірювання рівня піратства. *Актуальні питання міжнародних відносин*. 2012. Вип. 109 (І). С. 40–42.

[9] Levy-Carciente S. *International Property Rights Index 2020*. Property Rights Alliance, 2020. 89 p.

[10] Bardadym T.O., Gorbachuk V.M., Novoselova N.A., Osypenko S.P., Skobtsov V.Yu. Intelligent analytical system as a tool to ensure the reproducibility of biomedical calculations. *Штучний інтелект.* 2020. № 3. P. 67−81.

[11] Горбачук В.М., Гавриленко С.О., Голоцуков Г.В., Дунаєвський М.С., Ніколенко Д.І. До інтегрованого менеджменту і фінансового забезпечення інфраструктури охорони здоров'я районів Запоріжчини. *Кібернетика та комп'ютерні технології*. 2020. № 4. С. 87−99.

[12] Горбачук В., Скобцов Ю., Том І. Економічні аспекти забезпечення здоров'я в інформаційну еру. *National health as determinant of sustainable development of society*. N.Dubrovina, S.Filip (eds.) Bratislava, Slovakia: School of Economics and Management in Public Administration, 2021. P. 697−720..

[13] Кнопов П.С., Норкін В.І., Атоєв К.Л., Горбачук В.М., Кирилюк В.С., Біла Г.Д., Самосьонок О.С., Богданов О.В. Деякі підходи використання стохастичних моделей епідеміології до проблеми COVID-19. Київ: Інститут кібернетики імені В.М.Глушкова НАН України, 2020.
http://incyb.kiev.ua/archives/3988/dejaki-pidhodi-vikoristannja-stohastichnih-modelej-epidemiologii-do-problemi-covid-19/

[14] H. Ritchie, E. Ortiz-Ospina, D. Beltekian, E. Mathieu, J. Hasell, B. Macdonald, C. Giattino, C. Appel, M. Roser, E. van Woerden, D. Gavrilov, M. Bergel, J. Crawford, M. Gerber. Coronavirus Pandemic (COVID-19). https://ourworldindata.org/covid-vaccinations. 2021.

[15] Горбачук В., Гавриленко С., Голоцуков Г., Ніколенко Д. Цифрова економіка й освіта. *Vzdelávanie a spoločnosť. Prešovská univerzita v Prešove*. R.Bernatova, T.Nestorenko (eds.) 2021. VI. P. 253–258.

[16] Горбачук В.М., Гавриленко С.О., Голоцуков Г.В. Розвиток інтелектуальної власності, індустріалізації та цифровізації. *Цифровізація економіки як фактор економічного зростання*. О.Л.Гальцова (ред.) Запоріжжя: Класичний приватний університет; Херсон: Гельветика, 2021. С. 24–43.

[17] Le Cun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998. 86 (11). P. 2278–2324.

[18] Вапник В.Н. Восстановление зависимостей по эмпирическим данным. М: Наука, 1979. 448 с.

[19] Гайворонский А.А., Горбачук В.М. Субградиентный метод решения детерминированных и стохастических задач оптимального управления. *Вычислительные аспекты в пакетах программ и опыт решения оптимизационных задач*. К.: Ин-т кибернетики АН УССР, 1981. С. 41–46.

Thermal-Mask – A Dataset for Facial Mask Detection and Breathing Rate Measurement

Leonardo Queiroz, Helder Oliveira and Svetlana Yanushkevich
Biometric Technologies Laboratory
Department of Electrical and Software Engineering
University of Calgary, Canada
{leonardo.queiroz, helder.rodriguesdeol, syanshk}@ucalgary.ca

***Abstract*—This paper demonstrates the usability of thermal video for facial mask detection and the breathing rate measurement. Due to the lack of available thermal masked face images, we developed a dataset based on the SpeakingFaces set, by generating masks for the unmasked thermal images of faces. We utilize the Cascade R-CNN as the thermal facial mask detector, identifying masked and unmasked faces, and whether the mask colour indicates a inhale or exhale state. The latter is used to calculate the breathing rate. The proposed Cascade R-CNN is a multi-stage object detection architecture composed of detectors trained with increasing Intersection-of-Unions thresholds. In our experiments on the Thermal-Mask dataset, the Cascade R-CNN achieves 99.7% in precision, on average, for the masked face detection, and 91.1% for recall. To validate our approach, we also recorded a small set of videos with masked faces to measure the breathing rate. The accuracy result of 91.95% showed a promising advance in identifying possible breath abnormalities using thermal videos, which may be useful in screening subject for COVID-19 symptoms.**

***Index Terms*—Biometrics, Breathing Rate, Bounding Box Regression, Thermal Facial Mask, Decision Support, Covid-19, Object Detection.**

I. Introduction

According to the World Health Organization, people with COVID-19 experience a wide variety of symptoms, which include fever or chills, shortness of breath or difficulty breathing [1]. Despite numerous efforts, conventional temperature screening (such as non-contact infrared thermometers or oral thermometers) can result in long lines in places such as airports and hospitals, which compromises social distancing and increase the screeners' risk to get infected. Thermal imaging systems have been used to measure human skin surface temperature accurately without being physically close to the person being evaluated. The US Food and Drug Administration (FDA) has issued guidance for initial temperature assessment during a triage process using thermographic systems (thermal cameras) to determine skin temperature from a distance [2].

In the COVID-19 context, thermography can be used not only for the temperature estimation at a distance, but also to evaluate breathing rate. This is because thermal video reflects the variations in the thermal image intensity during inhale and exhale. This, in turn, can help identify the difficulty of breathing which is one of the symptoms of the disease.

This paper proposes to apply deep learning techniques to thermal videos in order to assess whether the subjects are wearing masks, as well as estimate the respiration rate by analyzing the thermal image intensity rate over time.

This work's contributions are as follows:

1) We propose a synthetic dataset based on the SpeakingFaces dataset [3]. To the best of our knowledge, this is the first effort to develop a dataset that have thermal images of subjects wearing protective masks. We developed an algorithm that automatically applies protective surgical masks to the faces, thus creating a Thermal-Mask dataset with both masked and unmasked faces.
2) We apply the Cascade R-CNN deep learning model to detect whether individuals are wearing masks. For those with masks, we also identify the state of breathing as either inhaling or exhaling.
3) Based on the output of the previous step that detects the breathing state, we evaluate the breathing rate.

This paper is organized as follows: Section II presents the related works; Section III describes the approach to synthesize our dataset; Section IV explains the facial mask detection applying the Cascade R-CNN as the model and the breathing rate measurement based on the mask colour variation. In Section V, we present the experimental results; Section VI contains conclusions and future directions.

II. Related Works

The requirement to use masks in public places such as in hospitals, schools, and airports during the pandemic has made facial biometrics such as automated face detection and feature recognition, much more difficult. Several studies that apply deep learning models for visual mask detection have been published lately, and also many datasets with masked faces were developed for machine learning applications. However, most studies focus on images in the visual spectrum. Mask detection in infrared (thermal) spectrum have not been yet studied as of present date, to the best of our knowledge.

Listed below are the most used datasets in the visual spectrum applied in the deep learning approach for face mask detection:

- MAsked FAces dataset (MAFA) [4]: It contains 30,811 visual images from the internet and 35,806 annotated masked faces, with different levels of occlusions.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

- Paper [5] referred to the three types of masked face datasets:
  (1) Masked Face Detection Dataset (MFDD) of 24,771 masked face images from the internet.
  (2) Real-world Masked Face Recognition Dataset (RMFRD) created using a Python crawler tool using front-face images of celebrities and their corresponding masked face images from the internet. It includes 5,000 pictures of 525 people wearing masks and 90,000 images of the same subjects without masks.
  (3) Simulated Masked Face Recognition Dataset (SMFRD) developed using an algorithm to put masks on faces of the LFW [6] and the Webface [7] datasets. In total, it has 500,000 face images of 10,000 subjects.
- Paper [8] reports a new MaskedFace-Net dataset based on the Flickr-Faces-HQ3 (FFHQ) dataset [9]. The masks were placed on the unmasked faces; the set is divided into the Correctly Masked Face Dataset (CMFD) and the Incorrectly Masked Face Dataset (IMFD). The dataset has a total of 137,016 images.
- "Prajna Bhandary" dataset [1] which is a synthetic dataset with 1,376 web-scraped images of 690 masked faces and 686 unmasked faces.
- FaceMaskDetection dataset [2]which is a combination of the MAFA [4] and the WIDER-Face [10] datasets respectively for masked and unmasked faces; the incorrect annotations were corrected resulting in 7,971 images.

Few publications exist that apply those datasets on the deep learning approach for facial mask detection. In the study, [11] the authors applied a hybrid deep learning for facial mask detection and used the RMFRD and SMFRD datasets for masked faces and the Labeled Faces in the Wild (LFW) dataset [6] for unmasked faces.

In [12], among other datasets, the authors also used the 'Prajna Bhandary' dataset and developed a real-time facial mask detection using a modified SSD (Single Shot Multibox Detector) with the MobileNetV2 backbone to deploy on embedded devices (like NVIDIA Jetson Nano, Raspberry Pi).

In addition to detecting masked faces, some studies focused on detecting a proper mask wearing. Using the MAFA dataset [4], the study [13] proposed an approach based on the Mask-RCNN to perform detection and segmentation of faces and masks. An additional CNN with a Soft Attention unit was then used to detect the correctness of the mask usage.

Many current works only use images and do not consider the applied physics and variables of the environment and cameras, with most studies on machine learning or image processing. The authors of [14] analyzed physical properties to identify breaches in a mask's integrity, which can lead to a high risk of exposure and subsequent infection. They have applied thermal image analysis to detect leaks in the face mask, using the heat transfer and thermodynamics principles.

[1]https://github.com/prajnasb/observations
[2]https://github.com/AIZOOTech/FaceMaskDetection

A study reported in [12] involved visual and infrared (thermal) spectra of masked faces. The visual images were applied to detect the masked faces, and the infrared (thermal) images were used to measure the facial temperature.

In [15], thermal images of the masked faces were used to measure the facial temperature and the discomfort when wearing the protective masks. The purpose was to analyze the hands moving the mask, the facial skin temperature and the heat flow when wearing medical surgical masks and N95 respirators.

After the outbreak of Severe Acute Respiratory Syndrome (SARS) in 2003, some studies were carried out to analyze the feasibility of the thermography system for fever screening. In [16], an Infrared Fever Screening System (IFSS) for screening a large group of people was reported. A thermal camera was pre-adjusted such that it focuses on the face and neck, detecting febrile subjects and confirming the temperature with a conventional thermometer. More recently, Wang et al. [17] validated the diagnostic effectiveness of standardized thermography-based fever screening. In their work, they conducted a clinical study with 596 individuals. They confirmed that measurements based on maximum temperatures in both the inner canthi and the entire face are the best regions for detecting febrile individuals.

In addition to fever screening, thermal cameras were applied to breathing rate measurement. [18] recorded 11 thermal videos from 5 healthy subjects at rest. The subjects were mostly stationary with minor head movements. The Region of Interest (ROI) was just below the tip of the nose and up to the mouth level, where the exhalation airflow could be seen and measured the respiration rate.

Paper [19] proposed to track breath pattern in the infrared spectrum. Face physiology was used to select salient thermal features, and to identify the periorbital regions, with high temperature, and the nose, with low temperature. After selecting those regions, a Monte Carlo method for Bayesian tracking was applied. The series of mean temperature variations were utilized to compute the breathing rate.

In [20], a pose estimation algorithm was applied to detect the body parts and track the thoracic region in the near-infrared spectrum for breathing rate measurement.

Study [21] proposed a portable system applied to a smartphone, combining visual and thermal cameras. It collected images from healthy and unhealthy people with chronic respiratory diseases, some of them with fever, and applied a deep learning model for classification. The device detects the forehead temperature and the breathing rate acquired from the nostril region of the masked face. Note that the images of frontal face view were collected in a well-controlled environment at a close distance of 50 cm away from the camera.

Summarizing, most approaches for face mask detection were developed for visual and not thermal domain. Our project, however, aims to work with images in the infrared spectrum, covering a wider range of application. In the context of COVID-19, thermal cameras shall assist in the screening in

airports, hospitals and schools to detect febrile individuals. Our approach expands the usability of thermography further, and aims at detecting whether the individuals wear protective masks, as well as estimation of their respiratory rate.

III. Dataset Description

Given the lack of available datasets with masked face images in the infrared spectrum, or thermal images, we consider an enhancement of the existing dataset called SpeakingFaces [3], by applying an algorithm to add protective masks taking into account the thermal pattern they create.

A. SpeakingFaces Dataset

The SpeakingFaces dataset is a large-scale multimodal dataset that combines thermal, visual, and audio data streams. It include data from 142 subjects, yielding over 13,000 synchronized instances with video, image and audio, total 3.8TB of data. The video and still images were collected in both visual and infrared spectra, and aligned. For each of the 142 subjects, 900 frames of 9 different positions of the head were acquired at each of 2 trials. Figure 1 exemplifies one subject in the 9 different positions in visual and thermal spectra.

Fig. 1: Visual and thermal facial images of one subject taken from the SpeakingFaces dataset [3].

B. Thermal Mask Dataset

Out of over 4.5 million images in the SpeakingFaces dataset, we have selected 153,360 images in both spectra (visual + thermal). Following the steps shown in Fig. 2, we created the thermal patterns of surgical masks on the unmasked face thermal images, and formed a new data set called the Thermal-Mask Dataset.

The creation of the new data set included (Fig. 2):

1) On the visual images, we applied the RetinaFace [22], a state-of-the-art face detector, pre-trained on the WIDER Face dataset. We visually confirmed the good localization of the face based on the bounding box created. Next, the High-Resolution Network (HRNet) [23] was applied to detect the facial landmarks and localize the facial parts such as nose, eyes and mouth.
2) Considering that the visual and thermal images are all aligned, we used the facial landmarks acquired in the previous step and applied them to the thermal images.
3) To create the "templates" of the mask patterns that were not present in the original data set, we took images of one subject in a surgical mask using the thermal camera Flir A700. The subject simulated 9 head positions, in both the inhalation and exhalation respiratory mode. It created 18 variations of face masks patterns on the thermal images. Note that the intensity variation of the pixels (illustrated in color map as variance in color) on the face mask image region is different for the different breathing state. It is more intense on the exhale, indicating the higher temperature of the exhaled air. Next, we segmented and cropped out the masks to overlay them onto the original data set images.
4) Each cropped mask region has four annotated landmarks as seen in Fig.3. The Algorithm 1 was developed to impose those masks on the unmasked face images. It should be noted that we cannot "reuse" the same individuals with and without masks in our training set, since the model will become heavily biased and fail to generalize.

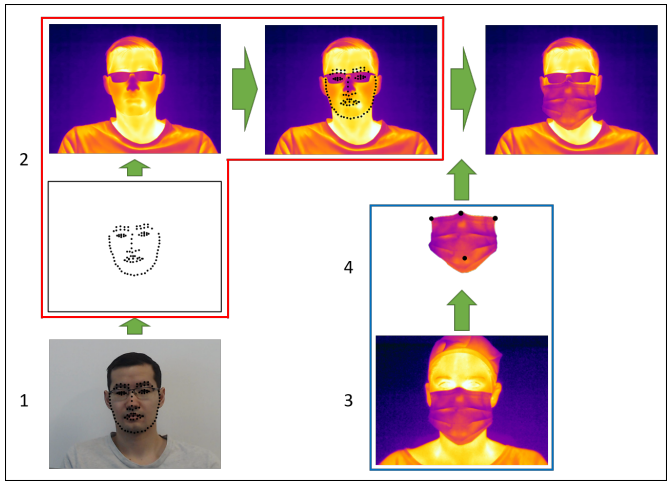


Fig. 2: Overview of the process of creating the Thermal-Mask Dataset

Figure 3 presents the landmarks used to match the facial masks and the thermal faces. To adjust the facial masks to each face, we performed the following image transformations:

Translation: The translation is done by moving the mask on the referential point 0 to point 52 on the face.

Resizing: For each face, we calculated the vertical distance between points 52 and 16 ($face_{vert\_dist}$) and horizontal distance between points 4 and 28 ($face_{hori\_dist}$). The same is done for the facial mask; we calculate the vertical distance between points 0 and 2 ($mask_{vert\_dist}$) and horizontal distance between points 1 and 3 ($mask_{hori\_dist}$). We calculate two proportions based on those vertical and horizontal distances, and then the average proportion. This final value is scaled up using the facial mask pixels, thus, resizing the image to fit with the corresponding face.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

$$
\begin{aligned}
scale_1 &= \frac{face_{vert\_dist}}{mask_{vert\_dist}} \\
scale_2 &= \frac{face_{hori\_dist}}{mask_{hori\_dist}} \\
scale_{total} &= \frac{scale_1 + scale_2}{2} \\
mask_{resized} &= mask * scale_{total}
\end{aligned} \tag{1}
$$

Rotation: Using the line connecting points 52 and 16 on the face, and the line connecting points 0 and 2 on the mask, we rotate the mask so that its corresponding line is aligned with the face line.

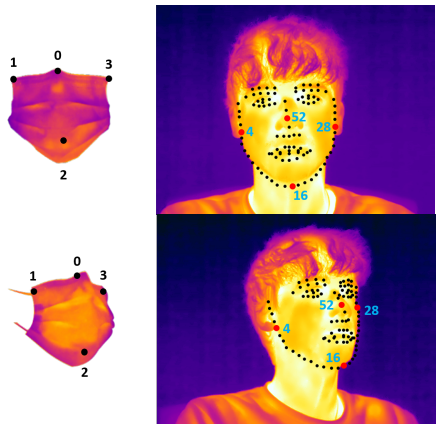


Fig. 3: Example of two Mask and Facial Landmarks

Algorithm 1 Generation of Thermal-Mask Dataset

```
function MASKING FACE(visualFace, thermalFace, mask)
    for visualFace ∈ SpeakingFaces do
        faceBbox ← FacialBoundingBoxDetection(visualFace)
        landmarks ← FacialLandmarksDetection(faceBbox)
        Match(landmarks, thermalFace)
        Map(mask, thermalFace)
    end for
    return Thermal-MaskedFace
end function
```

For our dataset, we initially started with 153,360 images in total (visual + thermal). However, there were some inconsistencies in the original dataset, such as blurred faces, images without faces and incorrect alignment between visual and thermal spectra. In such cases, the landmark algorithm did not work, so we removed those images during the step 1 and the final step after we check the incorrect mask locations. Figure 4 exemplifies some of those failed images.

Among the 142 subjects, we selected 80 to be masked and 62 to be unmasked, and created a total of 42,460 thermal masked faces and 33,448 thermal unmasked faces.

IV. Proposed Approach

In this section, we discuss how the dataset created was organized. We also describe the model used for mask detection with the respective breathing status. Finally, we cover the breathing rate measurement based on the mask colour variation.

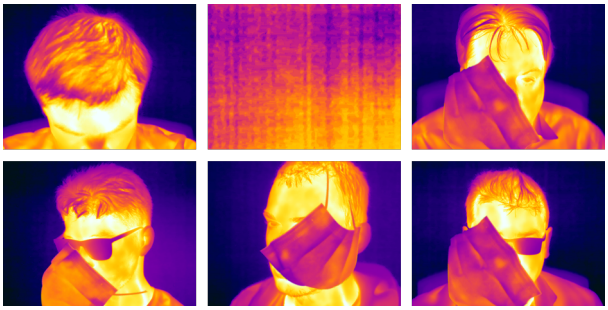


Fig. 4: Example of failed images

A. Dataset Preparation

With 42,460 faces with masks (80 subjects) and 33,448 faces without masks (62 subjects), we joined the samples, and among the 142 subjects, we randomly selected the samples as 70% for training (100 subjects), 20% (28 subjects) for validation and 10% (14 subjects) for testing. Those percentages will not be precise among the total number of images because the subjects do not have the same number of images since we have deleted the bad ones. The distribution ended up been 69.86% for training, 19.55% for validation and 10.59% for testing. Those samples will be applied to the Cascade R-CNN model for object detection. Table I summarizes the total number of samples in the final subset of the Thermal-Mask dataset.

TABLE I: Number of samples in each of the train, validation, and test data splits for the Thermal-Mask Dataset.

Set	Unmasked Faces	Masked Faces	Subjects
Train (69.86%)	23,188	29,842	100
Validation (19.55%)	5,940	8,905	28
Test (10.59%)	4,320	3,713	14
Total (100%)	33,448 (44.06%)	42,460 (55.94%)	142

The dataset has been labelled to the three classes:

1) unmasked face,
2) masked exhaling face,
3) masked inhaling face.

Face-based detection can be roughly divided into two application scenarios: uncontrolled and controlled application environments. We initially considered the Thermal-Mask dataset as being in a controlled scenario since the images were taken at a similar distance and had no background.

B. Mask Detection

To create our masked/unmasked face detection architecture, we used the MMDetection, a PyTorch-based [24] object detection toolbox [25] that, besides training and inference codes, also provides weights for more than 200 network models for object detection and instance segmentation.

Following the model representation of the MMDetection [25], we divided our model architecture into the following parts:

1) **Backbone,** a general feature extractor made up of convolutional neural networks to extract information in images to feature maps.

2) **Neck,** an intermediate component between a backbone and the heads; it performs enhancements or refinements on the feature maps.
3) **Head,** a detector and can achieve the network's final objective of detecting the masked/unmasked faces.

1) Backbones - ResNet and ResNext: The backbone refers to a convolutional neural network (CNN), which takes as input the image and extracts the feature map. In our approach, we applied the ResNet backbone and its variant ResNext.

a) ResNet: We used the Residual Network (ResNet), which is based on residuals or shortcut connections that skip one or more layers [26]. The shortcut connections perform identity mapping, and their outputs are concatenated to the outputs of the stacked layers as seen in Fig. 5. In our architecture, we applied the ResNet-50 and Resnet-101 with 50 and 101 layers, respectively.

b) ResNeXt: Another CNN we applied as backbone was a variant of ResNet that is called ResNeXt [27]. It follows the split-transform-merge paradigm, and the outputs of different paths are merged by adding them together with all paths sharing the same topology. The cardinality, the number of independent parallel pathways provides a new way to adjust the model capacity. Experiments show that the performance can be gained more efficiently by increasing the cardinality rather than increasing the number of deep layers. With the ResNext backbone, we applied two versions, the ResNeXt-101-32x4d, which stands for the architecture with 101 layers, 32 parallel pathways and a bottleneck width of 4 dimensions. The second version is the ResNeXt-101-64x4d that differentiates only due to its higher cardinality being equals to 64. Adding groups in parallel improve performance with the same computational complexity.

All the backbones were previously trained on the ImageNet dataset. The building block of both types of backbones in a computational equivalence is shown in Fig. 5:

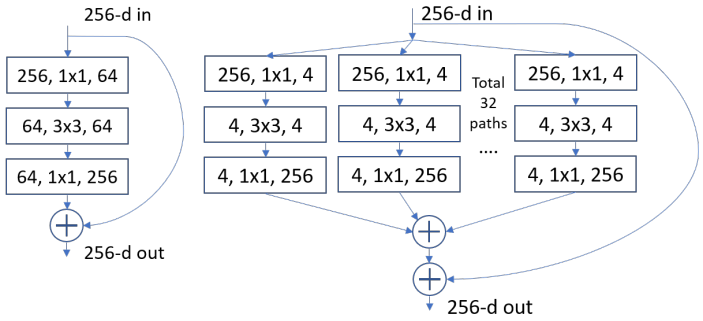


Fig. 5: Left: a building block of ResNet, Right: a building block of ResNeXt with cardinality = 32

2) Neck - Feature Pyramid Network (FPN): The long sequence of the CNN layers leads to increase of the semantic value of feature maps, while the spatial dimension (resolution) decreases. Following the blocks in Fig. 6, the lower layers of ResNet CNN are not selected for object detection due to its low semantic value that significantly slow-down the training. Therefore, the detector uses only upper layers, thus performing much worse for small objects.

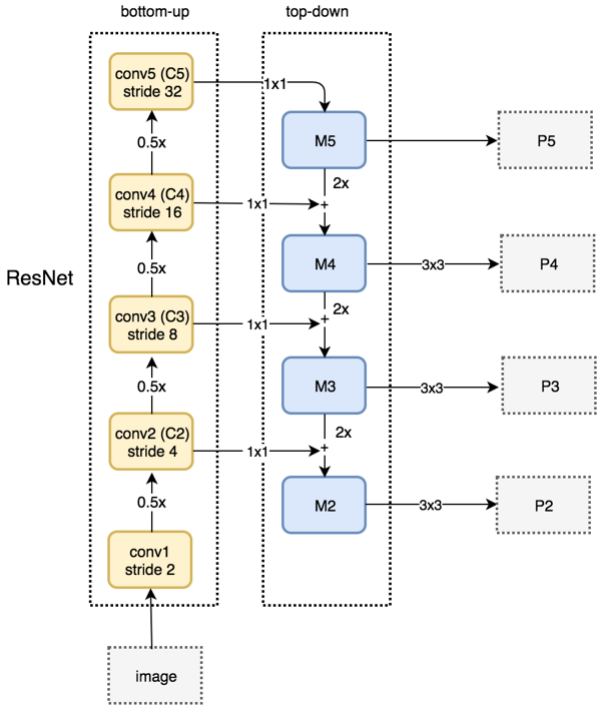


Fig. 6: Architecture of the FPN applied to the ResNet Backbone

To overcome the low resolution of the feature maps in upper layers and create a more generalized model, we applied the Feature Pyramid Network (FPN) [28]. The FPN takes an image as an input and outputs feature maps at multiple levels of size in a fully convolutional fashion. The construction of the pyramid is made up of bottom-up and top-down pathways.

The bottom-up pathway is composed of the backbone, which computes the feature map at various scales. For the ResNet seen in Figure 6, we use each stage's last residual block (C2, C3, C4, C5). As we go up, the spatial dimension is reduced by half, and the strides are doubled. The top-down pathway takes the feature maps from the bottom-up pathway and applies a 1x1 convolution filter to reduce the channel depth. As we go down, it is applied the nearest neighbour and upsample the previous layer by 2, adding them element-wise and creating the merged feature maps (M2, M3, M4, M5). Finally, a 3×3 convolution is applied on each merged map to reduce the aliasing effect of upsampling and generate the final pyramid feature map (P5, P4, P3 and P2). Since we share the same classifier and box regressor of every output feature map, all pyramid feature maps have 256-dimentional output channels.

In short, the top-down pathway creates higher resolution features by upsampling spatially coarser although semantically robust feature maps from higher pyramid levels. These features are then enhanced with features from the bottom-up pathway via lateral connections.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

3) Head - Cascade R-CNN model: After we process the image by a backbone (CNN) with the Feature Pyramid Network (FPN), we apply the Cascade R-CNN object detector [29]. Commonly the detector models are trained with a single threshold for Intersection over Union (IoU), being at least 50% for the object to define positives and negatives examples. It is quite a low threshold that creates many bad proposals from the Region Proposal Network (RPN), and it is difficult to reject close false positives to detect various quality and size objects in an image accurately. Detection performance also degrades for larger thresholds.

To overcome this problem, the Cascade R-CNN, a multi-stage object detection architecture, creates detectors trained with increasing IoU thresholds. For our application we applied IoU = (0.5, 0.6 and 0.7). It uses the same architecture as Faster R-CNN but with more branches in a sequence. The detectors are trained sequentially, using the output of a detector as the training set for the next. Fig. 7 shows the Cascade R-CNN architecture with the ResNet-50 Backbone. It is going to simultaneously boost the quality of hypotheses and detectors to improve object detection. This is accomplished with a combined cascaded bounding box regression and detection.

Figure 7 B) shows a few results of Cascade R-CNN applied to the test set of the Thermal-Mask dataset. Besides identifying whether the individual is wearing a mask, it also classifies the breathing status as inhale or exhale based on the mask image intensity distribution, displayed using colour.

C. Breathing Rate Measurement

The resulted images shown in Fig. 8 also highlight that masks with a lighter colour indicate that the individual is exhaling, therefore, heating the region between face and mask. Masks with a lower intensity, or a darker colour, on the contrary, indicate an inhalation state.

Our ultimate goal is to measure the breathing rate of masked people in the COVID-19 context. We created a small dataset to validate the breathing rate measurement application based on the intensity, or mask's colour variation. Using the thermal camera Flir A700, at a distance of 1.5m, we collected thermal videos from 11 masked subjects. Each subject is recorded three times for 1 minute each. In the first video, we ask them to breathe slowly; in the second video, we ask for faster breathing, and in the last recording, we ask for normal breathing. In total, we collected 33 videos from all 11 subjects. With that, we were able to apply the model previously described and measure each one's breathing rate.

Figure 8 presents one example of a subject "slowly" breathing with 12 breaths per minute (BPM). Considering that each individual has a different breathing pattern and the videos were taken in various locations with different thermal distributions, we had to standardize each frame's pixel intensity distribution. For this, we applied histogram matching, where we chose an image from the Thermal-Mask dataset as the reference and matched each frame histogram using the histogram matching function available on the scikit-image library. In sequence, we applied the frames to the Cascade R-CNN mask detection model to detect faces and classify them between unmasked, masked exhaling (level 1) and masked inhaling (level 0) subjects.

V. Experimental Results

We applied the Cascade R-CNN model for the object detection in thermal images of subjects with and without a mask. The model detects whether the subjects are wearing a mask, and, consequently, evaluates the respiratory state.

To evaluate how the Cascade R-CNN performs on the mask detection task, we used the Average Precision (AP) and the Average Recall (AR) metrics.

This measure is related to the Intersection over Union (IoU) that measures the overlap between the ground truth box and the predicted box over their union. The IoU score ranges from 0 to 1, where 1 indicates that the boxes are precisely in the same position. Predicted bounding boxes with IoU above a chosen threshold are classified as True Positive (TP). Bounding boxes with an IoU below the threshold are classified as False Positive (FP), and when there is an object but not a predicted box, we classify it as False Negative (FN). True Negative (TN) is not applied since the model would identify multiple empty boxes as non-object.

Precision represents a ratio of the true positive cases of the masked/unmasked face detection and the total number of positives, both true and false ones:

$$Precision = \frac{TP}{TP + FP} = \frac{true\ object\ detection}{all\ detected\ boxes} \tag{2}$$

Recall stands for a ratio of the true positive cases of the masked/unmasked face detection and the number of ground truth cases, including both true positives and the missed true cases called false negatives:

$$Recall = \frac{TP}{TP + FN} = \frac{true\ object\ detection}{all\ ground\ truth\ boxes} \tag{3}$$

The metric that summarizes both recall and precision and provides an assessment of the entire model is the Average Precision (AP). It is calculated for each category as the area under the curve (AUC) of precision-recall curve. The AP computes the average value of precision over the interval from $Recall = 0$ to $Recall = 1$ of $Precision = p(r)$:

$$AP = \int_0^1 p(r)dr \tag{4}$$

The AP tells us how well we are doing overall in terms of each category (class). To compute it over all three classes (unmasked, masked exhaling and masked inhaling), we calculate the average of AP with each class, calling it as mean Average Precision (mAP).

We applied the MS-COCO averaging approach, that calculates the mAP over different IoU thresholds (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95). This metric is denoted by mAP@[.5,.95] or just AP as we see in the second column

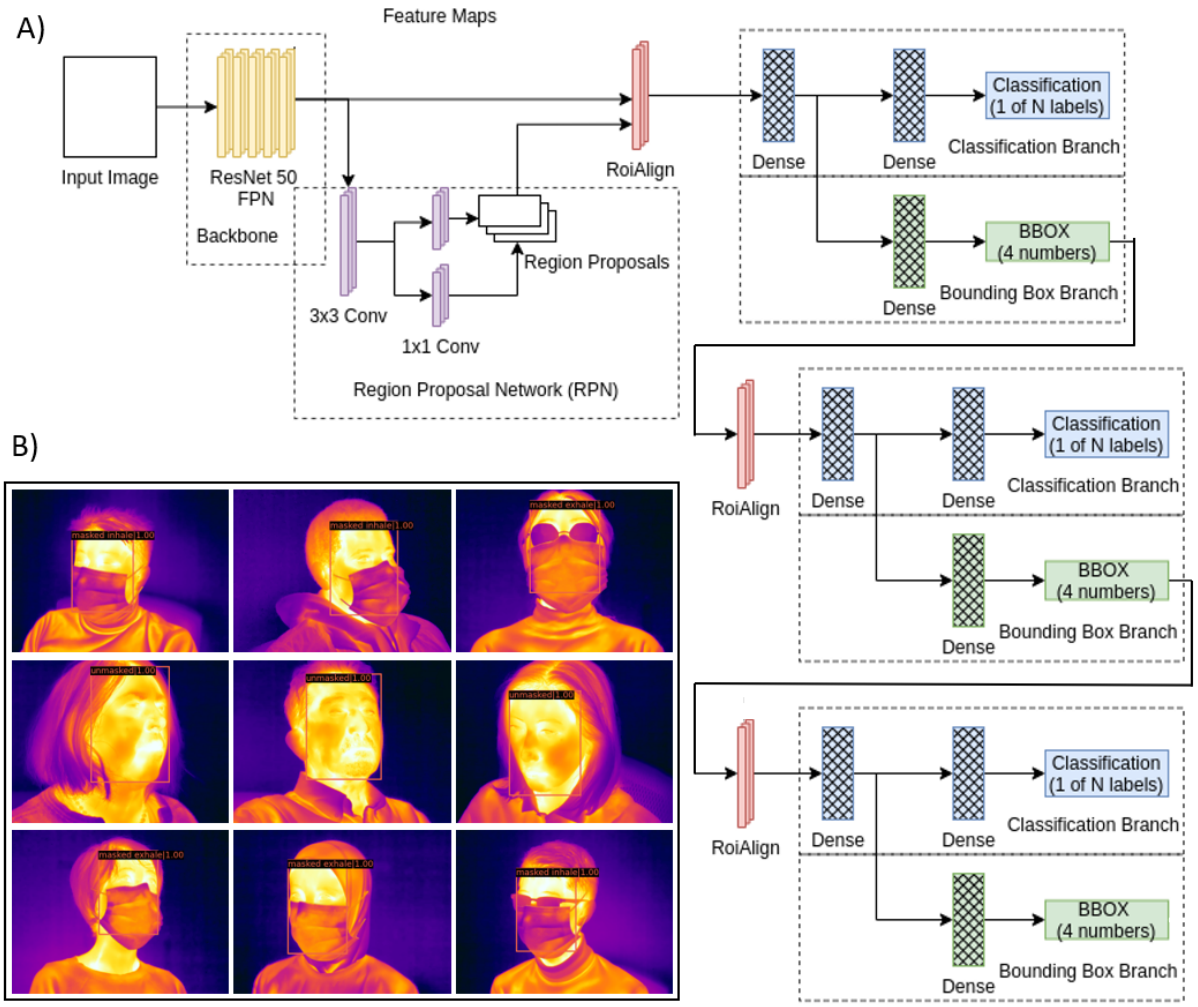


Fig. 7: A) Architecture for Cascade R-CNN Model and B) Results of Cascade R-CNN applied to the Thermal-Mask Dataset test images

of Table II. We calculate the averages AP not only over all classes but also on the defined IoU thresholds. The AP (interchangeably used for mAP) is also computed at the fixed value of IoU equal to 0.5 (AP_{50}) and at the IoU of 0.75 (AP_{75}).

Rather than computing recall at a fixed IoU, we also calculated the AR at IoU thresholds varying from 0.5 to 1. Average recall describes the area under the $Recall$ x IoU curve. The $Recall$ x IoU curve plots the Recall for each IoU threshold value within the interval $\in [0.5, 1.0]$, and is calculated as follows:

$$AR = 2 \int_{0.5}^{1} recall(IoU) \mathrm{d}IoU \tag{5}$$

Similarly to the average precision (AP) metric for the class-specific object detection, AR summarizes the performances across IoU thresholds (for a given number of proposals) and also correlates well with the detection performance. Likewise, mAR is the average of ARs over each of the three classes, and we apply AR interchangeably for mAR, as shown in the fifth column of Table II.

TABLE II: Cascade R-CNN.

Backbone	AP	AP_{50}	AP_{75}	AR
ResNet-50-FPN	0.873	0.997	0.986	0.904
ResNet-101-FPN	0.873	0.997	0.990	0.905
ResNeXt-101-32x4d-FPN	0.877	0.997	0.989	0.911
ResNeXt-101-64x4d-FPN	0.879	0.997	0.990	0.910

Table II presents the results of AP, AP_{50}, AP_{75} and AR for each backbone applied to the Cascade R-CNN. We reached the best result with the ResNeXt backbone, applying 101 layers, 64 parallel pathways and a bottleneck width of 4 dimensions with the Feature Pyramid Network (FPN) as the neck of our architecture. This result indicates that the model chosen for detecting the masks in the infrared spectrum, using the dataset developed, has a high level of reliability for applications with

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

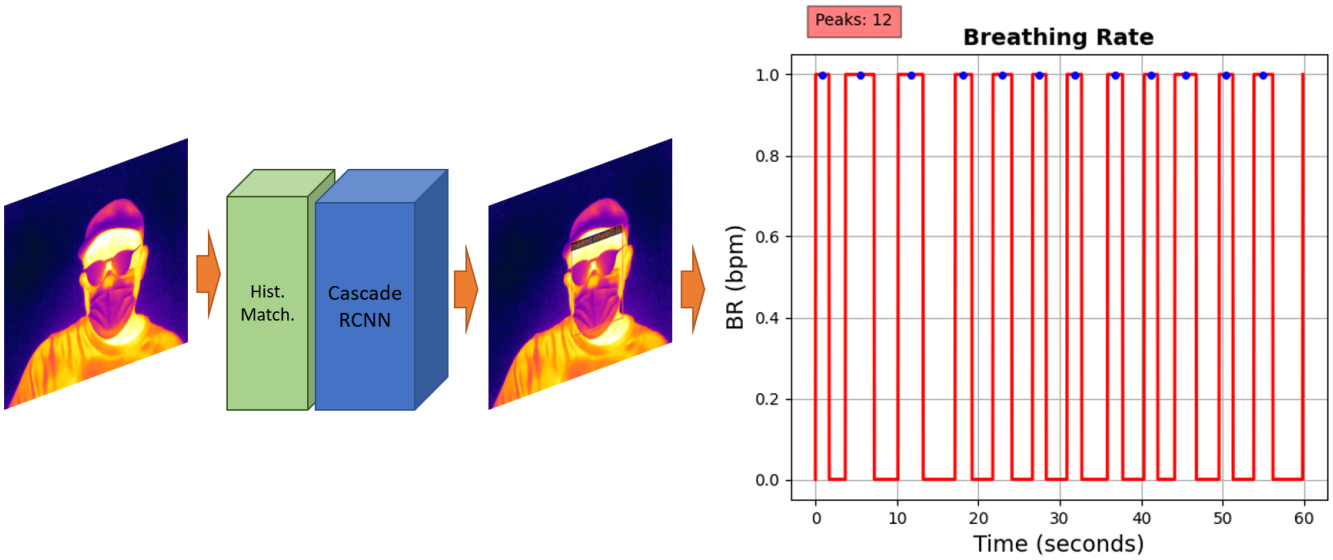


Fig. 8: Breathing rate measurement based on image intensity variation

thermal images acquired in controlled environments.

Figure 9 illustrates the scatter plot of the breathing rate visually measured, and also estimated using the proposed approach. We highlight the breathing rate correctly evaluated using green circles, and the misdetected ones using red circles. Although abnormal, the green circle on the far right differs only by having a very high rate of respiration, simulated by one of the participants. The perfect match line (blue) indicates the perfect linear correlation where the breathing rate estimated is equal to the breathing rate measured. The closer the coefficient of determination (R^2) to 1, the better the linear correlation results. We reach an R^2 of 0.779, which indicates that the estimated values' distribution has just a few outliers (red circles). After a careful analysis of the images and results, we found that the subjects, whose respiratory rate is outside the confidence interval range, are wearing masks of inadequate size. This might cause the airflow from the respiration to escape through the sides of the mask and not accumulate. Therefore, it will not heat the region between the mask and the face, making it difficult to detect the exhaling state. For the proper use of the created model, the participants must wear masks suitable for their respective face shape and size. This information is also important to affirm that the use of masks will not always provide the necessary protection if the individual does not wear the proper mask size and shape.

We also calculate the Accuracy (Acc), the Mean Absolute Error (MAE) of breaths per minute, and the Standard Error (SE), as described below.

With the average values of the breathing rate we calculated the percentage accuracy as:

$$Acc = 100 - 100 * \frac{\mid MBR_{avg} - EBR_{avg} \mid}{MBR_{avg}}, \qquad (6)$$

where MBR_{avg} stands for measured breathing rate and EBR_{avg} the estimated breathing rate.

To compare the breathing rate measured and the breathing rate estimated we also applied the *MAE* to measure an error between the paired observations:

$$MAE = \frac{1}{n} * \sum_{i=1}^{n} |EBRi - MBRi|, \qquad (7)$$

where n stands for the number of samples, $EBRi$ and $MBRi$ are an individual estimated breathing rate per minute (BPM) and an individual measured breathing rate per minute.

The *Standard Error* (SE) is calculated as:

$$SE = \frac{SD}{\sqrt{n}}, \qquad (8)$$

where SD is the standard deviation of the absolute difference between estimation and the reference data $(EBR - MBR)$.

Table III summarizes all the results. After using the thermal videos recorded with the masked subjects to detect the breathing rate based on the intensity or colour variation of the mask, we compared them to the actual breathing rate. It reached an average accuracy of 91.95% and an MAE of 3.76 bpm. To identify uncertainty around the mean measurement estimate, we use SE that provides a confidence interval. We calculated the 95% confidence interval as 1.96×SE, resulting in 1.28 bpm.

TABLE III: Accuracy, MAE and 95% confidence interval of the measured and estimated breathing rate for the recorded videos.

Source	Acc(%)	MAE	1.96 x SE (bpm)
Recorded videos	91.95	3.76	1.28

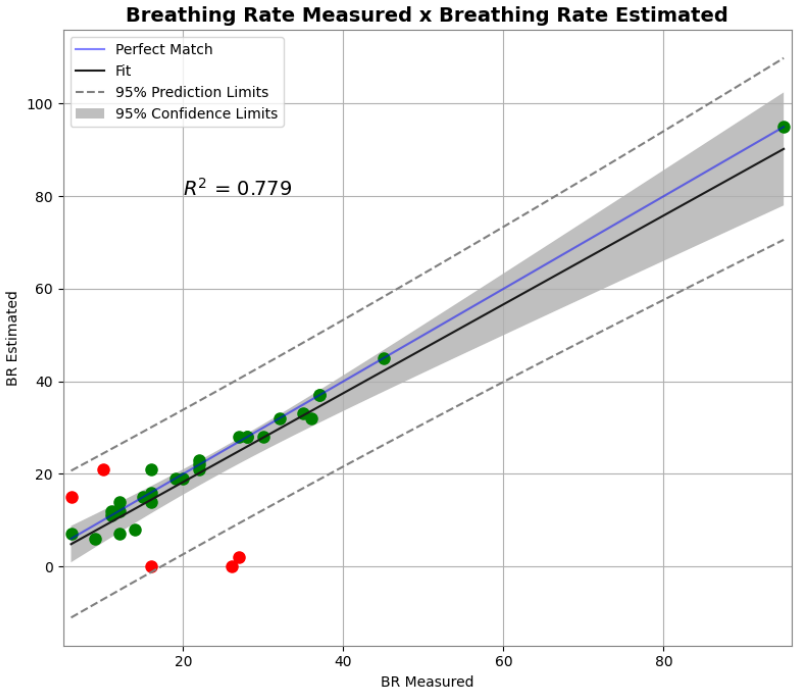


Fig. 9: Linear Correlation Analysis

VI. Conclusion and Future works

In the wake of COVID-19 pandemic, multiple control measures have been applied to mitigate the spread of COVID-19. It includes proactive methods such as screening and temperature measurement in public environments such as airports, hospitals, and shelters. The use of thermal cameras allows to remotely estimate the temperature of multiple people simultaneously. Our application aims to expand the usability of thermal cameras beyond measuring body temperature. We show that thermography can be used to detect whether the subjects are wearing masks. As well, in video thermography, the image intensity or colour variation helps determine the respiration rate. This is useful because the subject infected with COVID-19 may also have difficulties breathing or shortness of breath.

Our approach is to apply a deep learning model to thermal images to localize the face and classify it as a masked/unmasked face. As well, analysis of the image intensity in the mask region also indicates the state of breathing as inhaling or exhaling. Due to the lack of a dataset with thermal images of people wearing masks, our project includes creation of a dataset to be used in a machine learning model. Using the existing SpeakingFaces Dataset, we have developed a dataset with 42,460 thermal masked faces and 33,448 thermal unmasked faces called Thermal-Mask Dataset. To validate the dataset created, we applied an object detection model based on Cascade R-CNN. We tested with four different backbones: ResNet-50, ResNet-101, ResNeXt-101-32x4d and ResNeXt-101-64x4d, and also implemented the Feature Pyramid Network (FPN).

The model will detect whether the subject is wearing a mask or not and, based on the colour of the mask, will indicate the inhale or exhale breathing status. Once we have the breathing status, we can calculate the respiration rate and assess the difficulty in breathing, this being one of the symptoms of COVID-19. To validate this last goal, we also created a small dataset with 33 videos of 11 subjects. We recorded 1 minute where each one simulates a slow, quick and normal breath. The subjects were recorded in different locations, with varying temperature patterns and with various head positions during the measurement.

Table II presents the implementation results of the Cascade R-CNN on our Thermal-Mask dataset, reaching 99.7% for mean average precision and 91.1% for mean average recall. The best performance was reached using the backbone ResNeXt-101-64x4d-FPN.

Creation of a small thermal video dataset also revealed a promising way to measure the breathing rate on thermal images based on image intensity, or colour variation. We reached a breathing rate measurement accuracy of 91.95% on 33 videos of 11 subjects.

The overall results indicate that:

1) The developed Thermal-Mask dataset contributes to detecting masked/unmasked faces in the infrared (thermal) spectrum. It can also be the first step towards a more in-depth analysis of the proper use of the masks, assessing the position and the extent to which the region between the mask and the face contains the airflow exhaled by individuals.
2) The Cascade R-CNN model was proved to be a great

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

choice. It has highlighted the fact that the models with multiple branches associated with different IoU can be used in the detection of the protective masks in the infrared spectrum. Increase in the number of these branches will generate increasingly better results.

3) The ultimate goal of our work is to calculate the respiration rate. We were able to verify that for the subjects wearing the appropriate mask, even with continuous head movements in video, we can estimate the respiratory rate with high accuracy. More research is needed to assess whether the breathing difficulties, other than the symptoms of COVID-19, affect this measurement.

As future work, we plan to further expand the Thermal-Mask dataset with many more facial mask options, exploring variations in the protective mask material, types, and subject biometrics and/or demographics. In addition, we plan to create the sets with subjects incorrectly wearing the mask and other occlusions to identify its correct use, and distinguish between occlusions that are not masks.

Considering that the original dataset (SpeakingFaces) and therefore the created one (Thermal-Mask) were obtained in a controlled environment with specific face positions, distance from the camera and having no backgrounds, we found it difficult to generalize our model to the real-world applications. In this direction, the next step is to do experiment using some complex thermal background, as well as different distances between the subjects and the thermal camera.

Acknowledgment

This Project was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), via SPG "MOST: Biometric-enabled Identity Management and Risk Assessment for Smart Cities", and CREATE "WeTrac".

References

[1] World Health Organization (WHO). (2021) Weekly epidemiological update on COVID-19 – 16 March 2021. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210316_weekly_epi_update_31.pdf?sfvrsn=c94717c2_17&download=true

[2] Food and Drug Admistration (FDA). (2020) Enforcement policy for telethermographic systems during the coronavirus disease 2019 (COVID-19) public health emergency: guidance for industry and food and drug administration staff. [Online]. Available: https://www.fda.gov/media/137079/download

[3] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, and *et al.*, "Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams," *Sensors*, vol. 21, no. 10, 2021.

[4] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 426–434.

[5] Z. Wang, G. Wang, B. Huang, and *et al.*, "Masked face recognition dataset and application," *arXiv preprint arXiv:2003.09093*, 2020.

[6] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in 'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[7] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[8] A. Cabani, K. Hammoudi, H. Benhabiles, and M. Melkemi, "MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19," *Smart Health*, vol. 19, 2021.

[9] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 4396–4405.

[10] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 5525–5533.

[11] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, 2021.

[12] I. Farady, C.-Y. Lin, A. Rojanasarit, K. Prompol, and F. Akhyar, "Mask classification and head temperature detection combined with deep learning networks," in *2020 IEEE 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)*. IEEE, 2020, pp. 74–78.

[13] T. Truong, D. Lalseta, R. Ittyipe, and S. Yanushkevich, "Detecting proper mask usage with soft attention," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 1745–1750.

[14] J. B. Dowdall, I. T. Pavlidis, and J. Levine, "Thermal image analysis for detecting facemask leakage," in *Thermosense XXVII*, vol. 5782, International Society for Optics and Photonics. SPIE, 2005, pp. 46–53.

[15] A. Scarano, F. Inchingolo, and F. Lorusso, "Facial skin temperature and discomfort when wearing protective face masks: thermal infrared imaging evaluation and hands moving the mask," *International Journal of Environmental Research and Public Health*, vol. 17, no. 13, 2020.

[16] Y. H. Tan, C. W. Teo, E. Ong, L. B. Tan, and M. J. Soo, "Development and deployment of infrared fever screening systems," in *Thermosense XXVI*, vol. 5405, International Society for Optics and Photonics. SPIE, 2004, pp. 68–78.

[17] Y. Zhou, P. Ghassemi, M. Chen, D. McBride, J. P. Casamento, T. J. Pfefer, and Q. Wang, "Clinical evaluation of fever-screening thermography: impact of consensus guidelines and facial measurement location," *Journal of Biomedical Optics*, vol. 25, no. 9, 2020.

[18] J. Fei and I. Pavlidis, "Analysis of breathing air flow patterns in thermal imaging," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 946–952.

[19] Z. Zhu, J. Fei, and I. Pavlidis, "Tracking human breath in infrared imaging," in *Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*. IEEE, 2005, pp. 227–231.

[20] L. Queiroz, H. Oliveira, S. Yanushkevich, and R. Ferber, "Video-based breathing rate monitoring in sleeping subjects," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 2458–2464.

[21] Z. Jiang, M. Hu, Z. Gao, L. Fan, R. Dai, Y. Pan, W. Tang, G. Zhai, and Y. Lu, "Detection of respiratory infections using RGB-infrared sensors on portable device," *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13 674–13 681, 2020.

[22] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, 2019.

[23] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv preprint arXiv:1904.04514*, 2019.

[24] A. Paszke, S. Gross, F. Massa, and *et al.*, "PyTorch: an imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[25] K. Chen, J. Wang, J. Pang, and *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.

[27] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5987–5995.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2117–2125.

[29] Z. Cai and N. Vasconcelos, "Cascade R-CNN: high quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2021.

Exploratory analysis of the fire statistics using automatic time series decomposition

Dr. (Tech.), prof. M.M. Tatur
The Educational Establishment "Belarusian State University of Informatics and Radioelectronics"
Minsk, Belarus
tatur@bsuir.by

Ph.D. (Tech.), assoc. prof. A.G. Ivanitskiy
State Educational Establishment "University of Civil Protection of the Ministry for Emergency Situations of the Republic of Belarus", Minsk, Belarus
a.ivanitski@gmail.com

V.M. Prorovsky
The Educational Establishment "Belarusian State University of Informatics and Radioelectronics"
Minsk, Belarus
slawapro@gmail.com

Miroslav Kvassay
Faculty of Management Science and Informatics, University of Zilina
Zilina, Slovakia

***Abstract*— The aim of the work was to determine the possibility of using software based on machine learning technologies at the stage of exploratory analysis of data on the situation with fires in settlements, to assess the possibility of using the results obtained earlier by other researchers. Exploratory analysis of data and comparative assessment with results from other sources was carried out.**

***Keywords*— emergency situation, data mining, exploratory analysis, time series, forecasting, fire statistics**

I. INTRODUCTION

Exploratory data analysis is an important step in data mining. The researcher uses various methods of data transformation and visualization, which can be rather time-consuming. Currently, software products based on machine learning technologies are available for these purposes.

The main algorithm of this study is close to the SEMMA methodology (Sample, Explore, Modify, Model, and Assess) [1], which defines the stages of data mining.

The Ministry for Emergency Situations of the Republic of Belarus performs accounting of fires and their consequences, which is the basis for state statistical accounting of fires. A man-made fire is an emergency, defined as an uncontrolled combustion outside a special fireplace, resulting in damage and losses [2]. The above analysis does not cover minor fires (ignitions), which do not entail material damages. The studies in this article mainly contain the result of the application of stages corresponding to stages 1-4 of SEMMA.

II. DATASETS

The initial set of data was obtained from the database of the software package "Accounting for Emergency Situations" [3] and respectively processed (grouped by calendar days for the period from 2011 to 2020 and divided into additional data sets, such as: the total number of fires ("a"), fires in cities and urban areas ("b"), fires in rural settlements ("c"), fires in apartment buildings ("d"), the number of fires in single-family houses, summer cottages, outbuildings, yards and adjacent territories ("e").

III. ANOMALIES DETECTION

Modification stage included detection of anomalies - identification of rare data, events or observations that are unusual due to their significant difference from the rest of the data. As a rule, such anomalous data indicate an existing problem [4]. Despite the large number of studies on the detection of anomalies, not all methods are applicable to fire data due to the peculiarities of the seasonal and trend components. Therefore, the S-ESD (Seasonal Extreme Studentized Deviate) methods were used for automatic detection of data anomalies [5]. Statistical learning methods are used to detect anomalies. Decomposition by seasons is applied to filter trends and seasonal components of a time series and then robust median and median absolute deviation (MAD) statistics is considered to accurately detect anomalies and distinguish them from seasonal outliers. To exclude anomalous values, the outlier is removed or replaced with the nearest neighbors.

Reviewing anomalies in the course of data mining can lead to beneficial and previously unknown findings. For example, on July 17, 2016, an abnormal number of fires untypical for this time of the year was recorded - 41 cases (usually about 20). A detailed study proved that 23 of them were due to lightning strikes, and strong winds made the conditions even more complicated (strong wind is a registered meteorological emergency, the wind passes over the territory of three regions). A few days before the weather has changed to hot and dry, the air temperature reached 30 °C. Therefore, a rare simultaneous impact of several dangerous meteorological phenomena has led to a sharp increase in the number of fires.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Data visualization during preliminary study (Fig. 1) enabled researches to put forward a hypothesis about the signs of probable time series (trend and seasonal fluctuations).

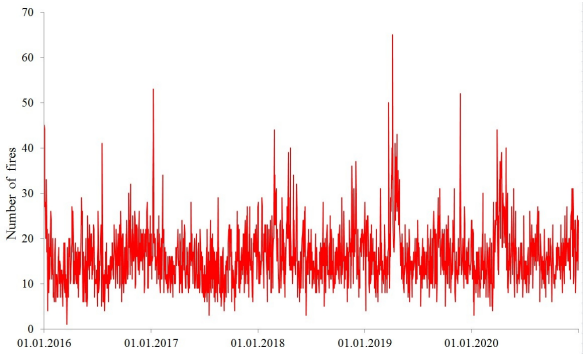


Figure 1. Five year fires chart

A time series is a sequence of observations, usually arranged by time. The main feature that distinguishes time series analysis from other types of statistical analysis is the importance of observations order. Time series analysis allows revealing the hidden patterns of a dataset, and forecasting methods provide information about the possible value of the studied indicator in the future. The main task of time series analysis, as a rule, is to define a model that describes the structure of the time series and can be further used for forecasting.

IV. PROPHET FEATURES

Classical methods of time series forecasting based on statistical models require additional costs to engage the experts in forecasting, who should configure the model and adjust the parameters of the applied methods depending on the specific problem area. Customizing these methods requires a deep understanding of time series models and their operation. Many organizations can't afford employing data scientists, and in most cases they lack resources for building complex forecasting platforms.

The choice of Prophet software framework as a forecasting tool is based on the studies of its estimated forecasting accuracy compared to other models [6, 7] (the current review does not cover the comparison with other models).

Prophet is designed to forecast the data using time series methods based on an additive model, in which non-linear trends correspond to yearly, monthly, weekly or daily seasonal fluctuations. It allows tracing of the holidays effect and other special days. The best forecasting results are achieved with time series that have strong seasonal effects and several seasons of historical data. Prophet is an open source software [8] developed by Facebook Core Data Science. The optimal parameters for the models are selected using machine learning methods expressed in the Stan probabilistic programming language, so the forecast can be obtained within a few seconds and in a fully automatic mode for disorganized data. Prophet is resistant to outliers, missing data and abrupt changes in time series, and includes many options for customizing and adjusting the forecasts based on available to user parameters for a specific area. The methodology is described in detail in [6]. It is based on the procedure for fitting additive (Generalized Additive Models, GAM) regression models of the following type:

$$y_t = g_t + s_t + h_t + e_t , \qquad (1)$$

where g_t and s_t are functions that approximate the trend of the series and seasonality (for example, yearly, monthly, weekly, etc.), h_t is a function reflecting the effects of holidays and other influencing events, e_t is normally distributed random disturbances.

The following methods are used to approximate the listed functions:

trend: piecewise linear regression or piecewise logistic growth curve;

yearly seasonality: partial sums of the Fourier series, ; the number of terms (order) determines the smoothness of the function;

weekly seasonality: presented as an indicator variable;

" holidays " : (public holidays and weekends - New Years, Christmas, etc., as well as other days when the properties of the time series can change significantly - sports or cultural events, natural phenomena, etc.) presented as indicator variables.

Estimation of the model's parameters is performed using the principles of Bayesian statistics (either by the maximum a posteriori (MAP), or by Bayesian inference) [9].

V. ANALYSIS OF DECOMPOSED DATA

The performance of the model was evaluated using experimental software in the Python programming language, which provides loading and processing of initial data, setting up and training the model, building a forecast, visualizing the initial and resulting data, cross-validation and calculation of metrics, saving the results.

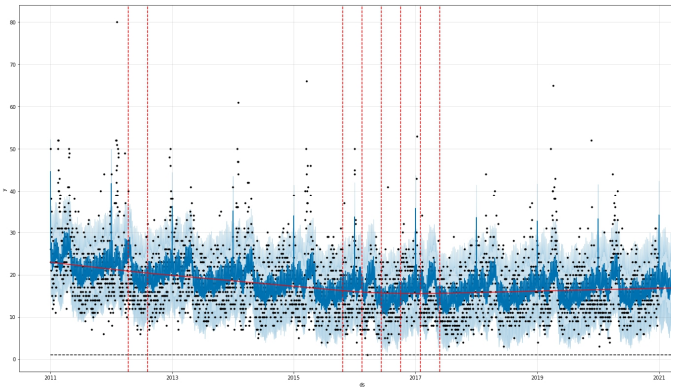


Figure 2. Extracting trend and chaingepoints

Let's train the model on the set “a” and draw the conclusion of the main components of the expanded time series.

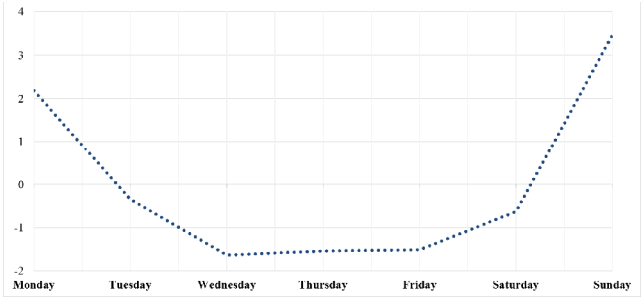


Figure 3. Weekly component

As is follows from the diagram (Fig. 2), the trend line (continuous red line) of the fire situation has several changepoints (vertical dashed lines). A slight fluctuation was noted in 2012. A total change in direction took place in 2016-2017. The similar section of the growing trend over the last five-year period in Russia is presented in [7], and suggests the presence of a common influencing factor for both states. In this case, it is possible to consider following hypotheses:

the consequences of the currency crisis in Russia in 2014-2015, which was caused by a rapid decline in world oil prices and led to a sharp weakening of the Russian ruble against foreign currencies, rise of inflation, a decrease in consumer demand, an economic recession, an increase in poverty and a decrease in real income of the population [10]. For example, in [11], it was suggested that the fire situation grew worse in connection with an economic recession and decline in the living standard of people;

changes due to climatic conditions. Studies of the influence of climatic conditions on the fire situation in Russia are given in [12].

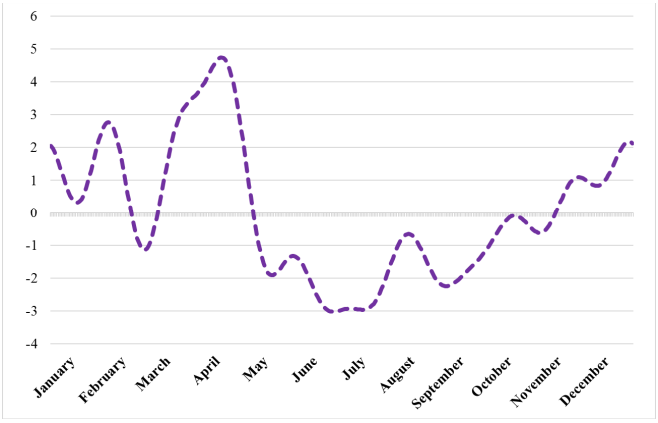


Figure 4. Yearly component

Assessment of the influence of days of the week on the frequency of fires (Fig. 3) is fully consistent with previously published data [12], [13]. There are more fires on weekends than on weekdays. Considering the fact that approximately 80% of fires occur in the residential sector [12], it seems reasonable to assume that there is a positive correlation between the time of stay at home and the frequency of fires. The model has correctly calculated and reflected this dependence.

The yearly component review (Fig. 4) provides a lot of additional information. The presence of a burst in April-May is a priori associated with a large number of dry vegetation fires within that period and is confirmed by other sources [7, 12]. At the same time, it is interesting to identify the features and conditions that have led to this burst. For this purpose we can use two additional data sets of time series “b” and “c”, which we obtained by grouping the initial data according to the type of settlement in which the fire occurred.

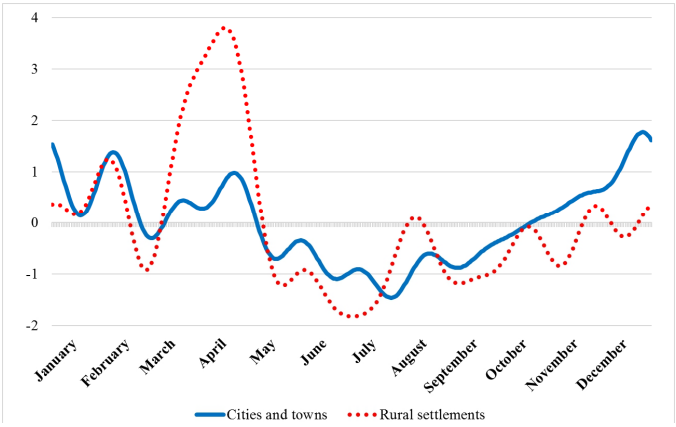


Figure 5. Yearly component by type of settlement

The spring avalanche-like increase in the number of fires (Fig. 5) was due to fires in rural settlements, enables us to continue research on the level of objects on where the fire occurred. We use the data sets “d” and “e”.

The yearly component reflecting the data for multi-apartment (including multi-storey) buildings has a shape close to a straight line, and minor fluctuations are presumably associated with fires in one- and two-story buildings (Fig. 6). For an in-depth analysis, we could consider such parameters as the number of residential buildings of each type, as well and their residents.

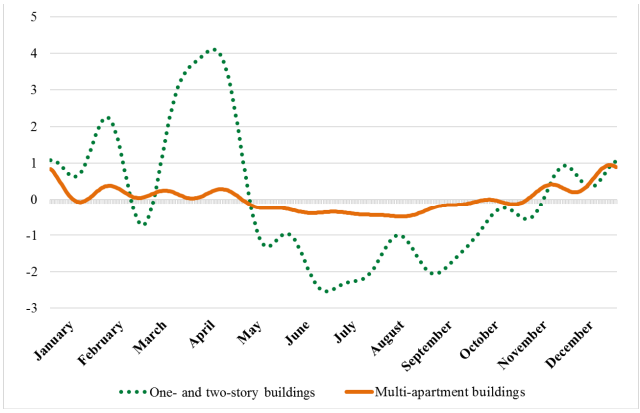


Figure 6. Yearly component by buildings type

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Let us review the diagram of influence (Fig. 7) of national holidays in Belarus, “standardly” used by the framework and the assessment of days based on statistical data. The highest score falls on January 1, the rest of the values are distributed almost evenly. Obviously, a more detailed study of this specificity is required.

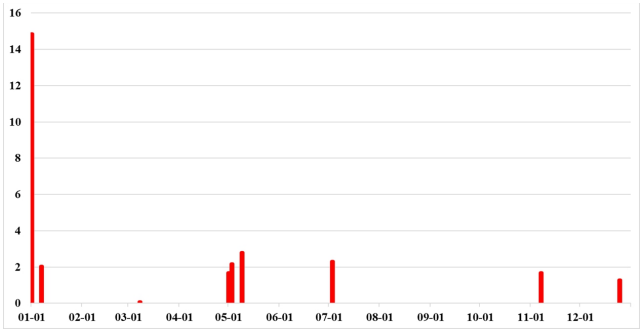


Figure 7. Holiday influence component

Taking into account the fact that the model supports a more accurate procedure for entering the information about the days on which social activity changes, it is obvious that the forecasting accuracy can be improved by setting the length of periods for consecutive holidays, as well as adding information about other anomalous dates, which include the shift of working.

Conclusions

1. The use of Prophet framework with additive models has enabled us perform exploratory data analysis, visualize the trend and its changepoints, weekly and yearly seasonal components, and determine the effect of holidays preset in the model by default on the fire situation.

2. Using the decomposition of the time series into independent separate series that reflect the aggregation of individual section of indicators, we reviewed the effect of avalanche-like increase in the number of fires in April and May.

3. Hypotheses of the reasons for the change in the direction of the trend line in 2016-2017 have been identified for further investigation.

4. A hypothesis was proposed that the model can be strengthened by improving the quality of the initial data on holidays, the total duration of weekend periods and adding other anomalous dates to the set.

5. In the process of detecting anomalies, previously undefined dependencies were revealed - an avalanche-like burst of fires under certain hazardous meteorological conditions.

Acknowledgment

This work was partially supported by the Slovak Research and Development Agency under the contract No. SK-FR-2019-0003 "Mathematical Models based on Boolean and Multiple-Valued Logics in Risk and Safety Analysis".

References

[1] Azevedo, A. KDD, SEMMA and CRISP-DM: A parallel overview [Electronic resource] / A. Azevedo, M. Santos // IADIS Multi Conference on Computer Science and Information Systems, Amsterdam, 22-27 July 2008 / Intern. Assoc. for Development if the Inform. Soc.; Associate Ed.: Luís Rodrigues and Patrícia Barbosa.– Amsterdam, 2008. – P. 182-185.

[2] On the Approval of the State Statistical Reporting Form 1-os (fires) «Report on Fires (Except Forest Fires) and the Consequences of them» and Instructions for Filling it out : Regulation of the Nat. Statistic Comm. of the Republic of Belarus, 27th June, 2017 № 49 // Nat. juridical Internet-portal of the Republic of Belarus [Electronic resource]. – Minsk, 2021 – Mode of access: https://pravo.by/upload/docs/op/T21703807p_1501016400.pdf. – Date of access: 20.04.2021.

[3] Develop a Software Package for Collecting and Analyzing Information about Emergency Situations and their Consequences : sci. rep. (final) / Sci.-Res. Inst. of Fire Safety and Emergencies of the MES of the Republic of Belarus; V.M. Prorovsky, M.V. Hodin, N.D. Chistyakov, T.A. Kornacheva, O.E. Kozlova. – Minsk, 2017. – 54 p. – № ГР 20163551. – Deposited in BelISA 04.07.2018, № Д201828.

[4] Zimek, A. Outlier Detection / A. Zimek, E. Schubert // Encyclopedia of Database Systems. Living Edition / Springer. – New York, 2018. – P. 40. – DOI:10.1007/978-1-4899-7993-3_80719-1

[5] Hochenbaum, J. Automatic Anomaly Detection in the Cloud Via Statistical Learning [Electronic resource] / J. Hochenbaum, O.S. Vallis, A. Kejariwal. – 2017. – Mode of access : https://arxiv.org/pdf/1704.07706.pdf. – Date of access : 20.04.2021.

[6] Taylor, S.J. Forecasting at Scale [Electronic resource] / S.J. Taylor, B. Letham // The American Statistician. – 2018. – Vol. 72, № 1. – P. 37-45.

[7] Babenyshev, S.V. Time Series Forecasting Based on Machine Learning Methods for Ensuring Natural and Technosphere Safety / S.V. Babenyshev, O.S. Malyutin, E.N. Materov // Siberian Fire-Rescue Bull.– 2021. – Vol. 1 (20). – P. 75-83. DOI: 10.34987/vestnik.sibpsa.2021.20.1.013.

[8] Prophet: Automatic Forecasting Procedure [Electronic resource]. – GitHub, Inc., 2021. – Mode of access : https://github.com/facebook/prophet. – Date of access : 20.04.2021 (MIT License Copyright (c) Facebook, Inc. and its affiliates).

[9] Mastitsky, S.E. Time series analysis with R [Electronic resource] / S.E. Mastitsky. – 2020. – Mode of access: https://ranalytics.github.io/tsa-with-r/. – Date of access: 20.04.2021.

[10] Vinokurov, M.A. Economic Crisis in Russia 2014-2015: Reasons, Consequences and Ways Out of the Crisis / M.A. Vinokurov // Socio-economic and Legal Region Safety Problem: Materials of the Intern. Sci.-Res. Conf., Irkutsk, 19-21 Febr. 2015 / Baikal State Univ. of Econ. and Law. – Irkutsk, 2015. – P. 65-71.

[11] Fire Safety and Up-to-Date Directions for its Improvement / E.A. Serebrennikov, A.P. Chupriyan, N.P. Kopylov; Ed. U.L. Vorob’ev. – Moscow : VNIIPO, 2004. – 187 p.

[12] Tatur, M.M. On Forecasting Fire Situation Related to Technogenic Emergencies in the Republic of Belarus: Approaches and Problems / M.M. Tatur, A.G. Ivanitskiy, V.M. Prorovskiy // Bull. of the Univ. of Civil Protection of the MES of the Republic of Belarus. – 2020. – Vol. 4. – № 3. – P. 237-250. – DOI 10.33408/2519-237X.2020.4-3.237.

[13] Prorovskiy, V.M. Regularity of the Spatio-Temporal Allocation of the Fires on the settlements in the Republic of Belarus / V. M. Prorovskiy, M.V. Hodin // Actual Problems of Fire Safety : Collected thesis of the XXX Intern. Sci.-Res. Conf., Noginsk, 06-08th June 2018. – Noginsk: VNIIPO, 2018. – P. 108-110.

A Concept of a Multi-robotic System for Warehouse Automation

Alexei Belotserkovsky
Department of Intelligent Information Systems
United Institute of Informatics Problems, National Academy of Sciences
Minsk, Belarus
orcid.org/0000-0002-8544-8554

Pavel Lukashevich
Department of Intelligent Information Systems
United Institute of Informatics Problems, National Academy of Sciences
Minsk, Belarus
orcid.org/0000-0001-9138-545X

Mert Doganli
Novosim Müh. Hiz. San. ve Tic. Ltd. Sti.
Istanbul, Turkey
mdoganli@novosim.com

Jan Rabcan
Department of Informatics
University of Žilina
Žilina, Slovakia
orcid.org/0000-0003-2835-9114

***Abstract* — The paper describes the concept of a warehouse multi-robotic system based on visual navigation methods. In contrast to the existing approaches for device navigation and planning of their movement, it is proposed to use external stationary cameras in operation. This approach will simplify and reduce the cost of the design of warehouse robots by using the existing warehouse video surveillance infrastructure (if available), and will also allow more flexible and reliable implementation of robotic automation tools for small warehouses of different configurations.**

Keywords — warehouse robotics, multi-robotic system, visual navigation, digital twin

I. INTRODUCTION

With the rapid growth of e-commerce, warehouses must meet the ever-increasing demands for processing speed while maintaining high precision in the picking of the cart. With the growth of sales, the management of basket picking and the product delivery is becoming a big challenge for companies. As a result, e-commerce companies started looking for solutions that would help them scale. For example, Amazon Kiva and Ali Baba Cainiao use tiered automated storage and retrieval systems and autonomous transport picker robots. However, many enterprises cannot afford to organize a multi-level automated system for storing and searching for goods, since it requires very high initial investments and is also difficult to implement technically.

This fact stimulated the emergence of complex commercial solutions for the automation of warehouse facilities based on standardized assembly robots (Magazino, InVia, etc.). Such solutions are suitable for warehouses of a relatively large area and low height, which is typical for suburban logistics centers. They can be adapted for warehouses where robotic automation was not originally planned.

Nevertheless, in addition to the issues of the cost of production and operation of robots, there are many different technical challenges associated with the development and operation of such systems:

- navigation;
- identification;
- manipulation with pallets (commodity items);
- prevention of collisions of robots;
- planning optimal routes;
- hardware.

All this has led to the fact that companies are beginning to experiment with their means of navigation, movement, control of robots, the implementation of a picker, etc. Thus, for example, expensive LIDAR systems, radio, and optical tags were often used for localization and navigation until recently.

These problems are quite universal and ubiquitous since it is obvious that the issues of competent and cheap placement of products in warehouses should affect any country that has an internal and external market. As it was shown above, the tasks reach the level requiring scientific elaboration. As part of a joint initiative of the Turkish Scientific and Technological Research Council of Turkey (TUBITAK) and the State Committee on Science and Technology of the Republic of Belarus (SCST), we initiated applied research on an alternative navigation method based on stationary cameras installed inside the warehouse and to apply it to a group of robots.

We assume that this approach is likely to simplify and reduce the cost of the design of warehouse robots by using the existing infrastructure for video surveillance of the warehouse, which will allow to:

- implement more flexibly and efficiently ready-made robotic automation tools for small warehouses of different configurations;
- solve more accurately and universally the problems of global route planning and collision avoidance for a group of robots by receiving up-to-date video information on the entire territory of the warehouse.

To speed up the development of visual navigation algorithms and expand the scenarios for their use, testing, and debugging of

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

alternative navigation methods, it was decided to develop a digital twin of an automated warehouse based on a model simulation of a warehouse and robots moving around it. As a result, this will allow us, without significant time and material costs, to flexibly change the configuration and dimensions of the warehouse, plan the location of cameras, change the design of virtual robots, their operating conditions, conduct virtual tests of algorithms, develop technical recommendations for setting up, adapting and optimizing the implemented warehouse infrastructure.

II. State of the Issue

The issues of robotics and technology of digital twins in the world are being developed by large IT companies. Under the already mentioned reference to the joint Belarusian-Turkish initiative, we were primarily interested in the world-class solutions available in these countries. For several years now, one of the private Belarusian laboratories of the IBA Group has been dealing with this problem in Belarus. However, the well-known corporation SAP is the leader in this area. As a result, due to commercial interest, all scientific and practical achievements in this direction are not published for the scientific community.

The attention to robotics is constantly growing, but there are few innovative solutions related to the control and navigation of robots, especially when it comes to multi-systems, “peaceful” logistics tasks, and the use of robots as a service. Among the Belarusian companies operating here, one should mention "Rozum Robotics". Their activities are related to the search for innovations in the field of robotic manipulators.

Regarding the topic of autonomous navigation using images and video sequences from onboard cameras, we discovered that R&D is mainly carried out concerning UAVs [1]. There are relatively few publications on this topic, and there are no complete innovative solutions yet. Some practical tasks on the creation of unmanned ground transport were announced, in particular, by the company "Mapbox", but the results of their activities have not yet been highlighted.

Our partner, Novosim company, in cooperation with Turkish scientists is developing means of automating the processes of collecting and storing goods in warehouses. The company is a leader in the country in this area and has experience in creating warehouse robotics in conjunction with its subsidiary Createchnic. For a commercial company, the most important aspect of successful existence in the market for partners is constant contact with local and international companies to monitor modern requests and requirements.

The solution for organizing robotic systems proposed for development within the framework of this project will be necessary for both robot development companies and logistics companies. It is planned to use the system as a test site in the Turkish side Createchnic autonomous assemblers. The implementation of such a test site will allow us to demonstrate visually in action a working prototype to potential customers and form their understanding of possible scenarios for using the product. Various companies are interested in using an autonomous warehouse maintenance product.

Currently, most logistics companies operate with warehouses 24/7 with the condition that they can meet seasonal changes in demand and move goods accurately. Such operating conditions create both hardware and software problems, especially with a wide range, therefore, automated controlled transporters are still widely used. There will be an increase in the number of fully autonomous vehicles and assemblers shortly. In this regard, a joint team decided to deal with the problem of navigation and control of robots.

The planned result of the cooperation is hardware and software that will be used for navigation (including collision avoidance), route planning, and localization of multi-robotic systems. The main areas of application are FMCG (Fast Moving Consumer Goods) and e-commerce. Localization refers to the search for the location of several robots in a known warehouse. The warehouse map may or may not be generated in advance. The calculated deviation from the correct location should be less than 1 m, which is comparable to expensive world analogs.

Since there will be several robots working in the warehouse (along with humans and other machines), it is important to develop algorithms for route planning and image processing to localize objects correctly in the room and avoid collisions between robots or other obstacles. To maximize the synergy of the parties and the interconnection of results, it is planned to develop the most autonomous solutions that can be used for any other robot manufacturer without the need to develop these algorithms anew. This SaaS solution is innovative from all sides: new algorithms, the model environment for the robot, and possible use by the end-user on a commercial basis.

One of the main problems when selling such products is to ensure independence from a specific target customer, which seems wrong from the point of view of product scalability. Another issue is that customers who want to conduct the first tests of automated solutions require a minimum cost of production. Thus, existing solutions are usually offered as a service. Once the customer doesn't need to make a significant initial investment, it becomes a little easier to interact. Such a service allows you to scale the use of the developed products. That is why the main task of the study is to produce the most versatile and safest solution that does not require special modifications to robots and a warehouse.

Preliminary analysis has shown that visual navigation is the most versatile and cheapest method. However, the navigation of robots only with the help of their onboard cameras is possible, but at the same time difficulties may arise in unforeseen situations, such as the breakdown of one of the robots, their collision, an obstacle on the route, etc. In addition, when several small robots are used, it is possible to simplify significantly the size, cost, and reliability of each robot separately by simplifying its indoor navigation system and use its external positioning. These considerations led us to an investigation of the applicability of visual navigation based on fixed external cameras.

In the future, it is planned to adopt several developed robotic solutions specifically for this task, as well as to use the ready-made platform along with the newly created one for carrying out field experiments. To minimize costs and increase the

socio-economic effect, to test innovative solutions, we will conduct our testing of navigation algorithms and simulation modeling of a multicomponent multi-robotic system.

III. The Scope of the Robots to Be Developed

Warehouse premises can differ significantly in many parameters, depending on the purpose. In the international classification, it is customary to divide warehouses into the following 6 classes: A +, A, B +, B, C, and D. The conformity of a warehouse to a particular class is advisory in nature and depends on many factors:

- location;
- a number of storeys;
- height and span;
- availability of engineering equipment (ventilation, heating and other equipment that allows you to create certain micro-climate);
- the presence of security systems and fire extinguishing systems, video surveillance;
- the height of the floors and the presence of an anti-dust coating;
- the presence of a certain number of dock-type gates, loading and unloading platforms, adjustable in height;
- availability of areas for maneuvering and parking for trucks and cars;
- the presence of office, auxiliary premises and buildings;
- and much more like the presence of fencing of the territory with round-the-clock security and others.

In this classification, class A and A + warehouses are the most technologically advanced and large. They are built from lightweight high-quality steel structures and are specially designed for storage needs. The height of the ceilings should provide storage in six to seven tiers. An example of such a warehouse is shown in Fig. 1.

Figure 1. Example of A class warehouse

The development and application of the above-mentioned automation systems based on mobile robots cannot be universal and cost-effective (especially if we are talking about automation as a service). Our target group will most likely be warehouses of class C and D. This is most often a capital production facility, with a ceiling height of at least four meters, with a concrete or other hard floor covering, a heating and ventilation system that maintains temperatures from +8ºC to +14 ºC.

In addition to these warehouses, there will be requirements for a unified storage scheme for goods - boxes, containers, pallets. This requirement will allow the use of unified manipulators for gripping and moving goods. The most common methods for robotic systems are now pallet-based storage methods for bulky goods and containers for small and light items.

Fig. 1 shows an example of a pallet warehouse, and Fig. 2 shows the configuration of shelves for pallet storage of goods. Fig. 3 shows the shelf configuration for a container-based storage system, and Fig. 4 shows a visualization of such a warehouse.

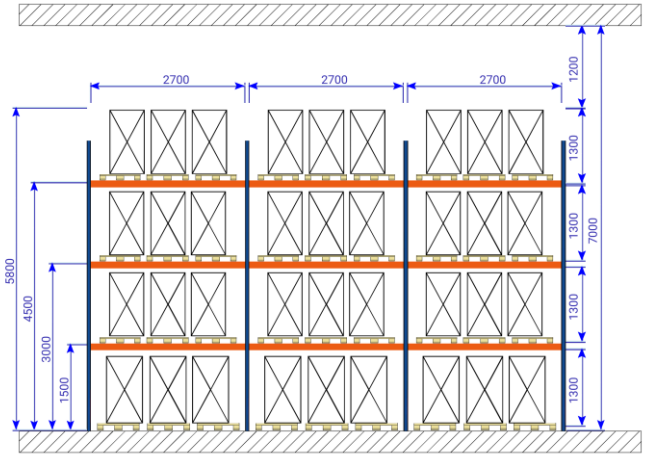


Figure 2. An example of a configuration of shelves for a C-class warehouse with an organized storage system based on unified pallets

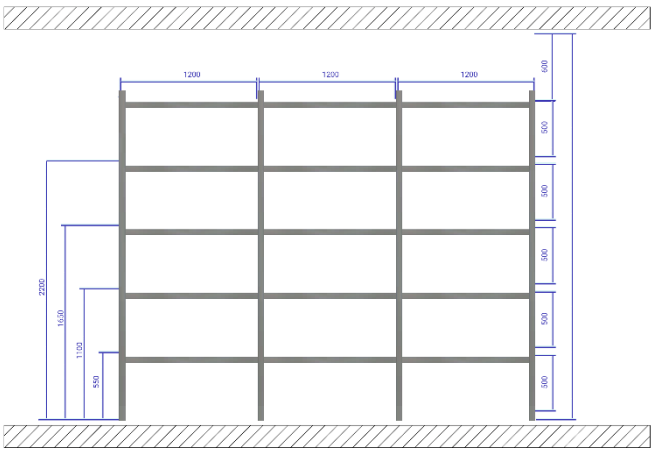


Figure 3. An example of a configuration of shelves for a D-class warehouse with an organized storage system based on unified containers

As already mentioned, based on the formulation of the task (development of a universal robotic system that can be distributed according to the model of providing warehouse automation as a service), the most universal and budgetary solution will be a container-based solution. In this case, the robots being developed will operate with fairly light goods, will be quite compact, safe for humans and cheap.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Figure 4. An example of a D-class warehouse with an organized storage system based on unified containers

IV. A Warehouse Test Model and Expected Robot Configuration

Picker robots have gained a lot interest both in academia and industry during the last couple of years. There are several challenges with autonomous picker robots which are being researched such as:

- having the ability to pick various items – manipulators, flex grabbers, etc.;
- autonomous obstacle free navigation – localization, path planning etc.;
- occupational health and safety – collaborative robots etc.;
- making use of height of the depot – telescopic systems etc.;
- task sharing - distribution of collection tasks.

Considering the pressure from the fast moving producers and e-commerce managers, all these problems are being dealt with.

On the industrial side, Germany Based Magazino [2] developed a picker and a transporter unit. This product was developed for box picking. Interestingly, there are several shelves on the body of the vessel. When the boxes are picked, the mechanical structure carries and places the box on an available shelf. This is a very smart innovation considering the operation because the robot will move to the shipping area only when the shelves are full. Therefore, the total time between items to be picked will be minimized. Also, it uses the height of the warehouse. One drawback of the product is that it occupies a lot of space due to the shelves and placing the structure.

InVia Robotics from the US [3] developed an order fulfillment robot that uses a dynamic platform system to reach higher levels. Both mechanical and pneumatic grabbers were offered and a simple navigation approach was used with landmarks. Compared to Magazino, this product only holds one box at once and delivers it to the target location.

Whereas Fetch Robotics followed a different path and came up with 2 different options; the first one follows a human picker and all picking work is done manually and the items are scanned and placed on the robot. Once the order picking is completed, the robot moves to the shipping area while an empty one approaches the human picker, second option is 2 robots collaborating for picking and delivering. One of the robots picks the product and places it on the carrier and the carrier moves all items to the shipping location. Fetch robotics holds several patents but recently got one for a method for localization of robots [4] using particle filter algorithm.

The first part gives an idea about the industrial applications. However, on the academia side, a lot of effort is being put into developing methods for topics covered by this project. Motion planning is the general name for path planning and is a process of finding a collision-free path for a robot from its initial to goal point while avoiding collisions with any static obstacles or other agents present in its environment.

A diverse variety of algorithms have been proposed for producing a feasible trajectory. Initially, the path planning efforts will focus on Rapidly-exploring Randomized Tree (RRT) [5] and D* Lite algorithms [6]. Sampling-based planners (SBP) such as the RRT [5] and the Probabilistic Road Map (PRM) [7] are probabilistically complete, and computationally efficient for many motion planning problems. Algorithm efficiency is a measure of the average execution time necessary for an algorithm to complete work on a set of data. Algorithm efficiency is characterized by its order. If two algorithms for the same problem are of the same order, then they are approximately as efficient in terms of computation. Algorithm efficiency is useful for quantifying the implementation difficulties of certain problems. Probabilistic completeness is the property that as more "work" is performed, the probability that the planner fails to find a path, if one exists, asymptotically approaches zero. Several SBPs are probabilistically complete. SBP algorithms are known to provide quick solutions for complex and high-dimensional problems using randomized sampling in search space [8], [9]. This algorithm will need to define an optimal collision-free path for multiple robots.

The bidirectional (two-tree) [10] variants of the RRT algorithm have been successfully applied to complex instances of the motion planning where the platform is high-dimensional and must search for paths through narrow corridors, usually referred to as "bug traps", while leveraging the full capabilities of the robot [11]. The name "bug trap" is inspired by actual devices for catching bugs, where bugs can enter easily, but it is hard to escape. Since it is not known whether the initial or target point might lie in a bug trap (or another challenging region), a bidirectional approach is often preferable. This follows from an intuition that two propagating wavefronts, one centered on initial and the other on target, will meet after covering less area in comparison to a single wavefront centered at the initial point that must arrive at the target point.

The D* Lite algorithm [6] is an adaptation of Koenig's Lifelong Planning A* (LPA) [12] which in turn is a derivation of A* [13] search incorporating incremental search. Incremental search methods reuse information from previous searches to find solutions to similar problems much faster than is possible by solving each search task from scratch. A* search is an efficient way to calculate the optimum path along with the graph with heuristic cost estimation and current cost.

Within this project, using the existing methods and algorithms, drawbacks resulting from these algorithms (in

literature) will be repeated and novel algorithms improving the drawbacks will be proposed. Also, from the customer's expectations, an accurate working model (PoC, proof of concept) will be demonstrated. One of the questions is to replace costly LIDAR type of sensors with cheaper cameras so that the penetration of such systems to warehouses would be easier. Therefore, accurate estimates and predictions

As a basis for the development of PoC of such test warehouse robot, we will take a small D-class warehouse with an area of about 100 m^2 with an organized storage system in shelves in unified containers.

Fig. 5 shows our prototype of a warehouse robotic arm for picking and moving standardized containers. As you can see from the image, the robot has a mobile electric platform with batteries and a simple manipulator that allows you to grip, hold and leave the container.

Figure 5. A prototype of a warehouse robotic arm for gripping and moving standardized containers

The figure demonstrates that taking into account the large base of the robot and its height, the first tier of shelves will be at the level of the minimum height of the manipulator.

Fig. 6 gives an example of a warehouse plan for the aforementioned container storage system and a given prototype warehouse robot. The dimensions of the warehouse, the distances in the aisles, the height and number of storeys of shelves take into account the weight and size characteristics of the robot and its other limitations.

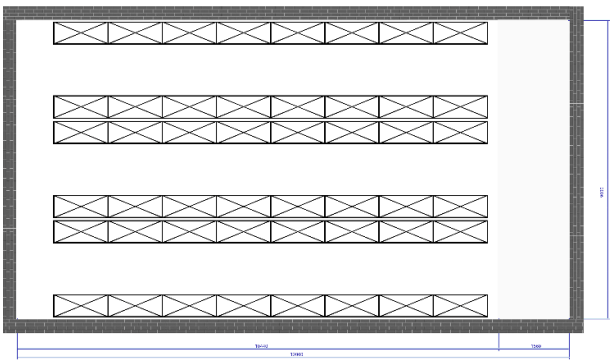


Figure 6. An example of a D-class warehouse plan with an organized storage system based on unified containers

The main navigation task will be solved by video from external cameras fixed in the warehouse. If it is impossible to provide the required accuracy during the simulation using only external video cameras, decisions can be made regarding the additional use of navigation cameras onboard the device.

Below you can see preliminary estimation of the requirements to the warehouse system and to the location of the cameras (can be corrected after experiments or specified by the customer's request):

- the average density of cameras: 1 camera per 10 m of the length of the aisle of the warehouse corridor;
- ceiling height of the warehouse: 3 - 7 m;
- location of cameras: provides information for the movement of robots;
- the number of simultaneously supported cameras per centralized computing device: 10 pcs (position update rate 15 Hz; Server hardware based on 2x Intel Xeon E5530);
- the robot performs the capture of goods by airborne systems within the specified accuracy characteristics; if necessary, a more accurate capture is necessary, the final refinement of the position of the manipulator is made onboard the robot.

Expected technical characteristics of such warehouse robot navigation system with mentioned requirements are provided below in the table 1.

TABLE I. TECHNICAL CHARACTERISTICS OF MULTI-ROBOTIC SYSTEM NAVIGATION SYSTEM

Characteristic	Value
Accuracy of determining the position of the robot	0.05 – 0.15 m
Accuracy of the static determination of the orientation of the robot	5° – 10°
The refresh rate of the position of the robot	15 Hz
The ability to identify and prevent collisions of robots	Yes
The ability to identify outlier objects in the area of movement of robots (falling goods, determining people, other obstacles)	Yes
The ability to detect incorrect capture or loss by the robot when moving the storage unit	Yes
The possibility of online route planning taking into account the blocking of a part of the permitted movement zone and its congestion	Yes

V. CONCLUSION

The paper uncovers the need for an additional or alternative way of visual navigation of warehouse robots and proposes to use stationary cameras located indoors. The use of this type of navigation as the main or additional way of positioning a warehouse robot allows us to build a more versatile and reliable system that can more quickly and adequately respond to difficulties in unforeseen situations (a breakdown of one of the robots, their collision, the formation of an obstacle on the route, etc).

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

To verify and test the algorithms for managing the warehouse, as well as navigating robots without the need for hardware implementation of prototypes, the project announced the use of a model of a digital twin of the warehouse. It is shown that the presence of a digital twin will allow in the future to remove promptly and to eliminate problems arising during operation, as well as to refine and improve the software and hardware solutions of the automated warehouse. Identified problems and recommendations for their elimination will be developed based on the simulation performed on the technical characteristics and location of the stationary cameras of the warehouse and the on-board cameras of mobile robots; assessing the reliability and accuracy of navigation for the implemented model, other recommendations for improving and optimizing the implemented warehouse infrastructure.

ACKNOWLEDGMENT

The study of the problems of robot navigation and the modernization of warehouses was carried out under the initiative of the joint Belarusian-Turkish project TUBITAK-SCST “Development of a warehouse autonomous multi-robotic system”.

This work was also partly supported by grant of the Slovak Research and Development Agency SK-SRB-18-0002.

REFERENCES

[1] Lukashevich, P. The new approach for reliable UAV navigation based on onboard camera image processing / P. Lukashevich, A. Belotserkovsky, A. Nedzved // International Conference on In- formation and Digital Technologies (IDT), – 7-9 July, 2015. – P 230-234.

[2] Magazino GmbH, EP 3 192 616 A1 – European Patent Application, 19.07.2017

[3] in Via Robotics, LLC, US 9,120,622 B1 – US Patent , 01.09.2015

[4] Fetch Robotics , Inc, US 9 ,927,814 B2 - US Patent , 27.03.2018

[5] S. M. LaValle, “Rapidly-exploring random trees: A new tool for path planning,” 1998.

[6] Koenig S, Likhachev M. D* Lite. Aaai/iaai. 2002 Jul 28;15.

[7] Kavraki, L.E., Švestka, P., Latombe, J.-C., and Overmars, M.H. (Aug. 1996) “Probabilistic roadmaps for path planning in high-dimensional configuration spaces”, IEEE Trans. on Robotics and Automation, vol. 12, pp. 566–580.

[8] S. Karaman, and E. Frazzoli, "Sampling-based Algorithms for Optimal Motion Planning", The International Journal of Robotics Research, vol. 30, pp. 846-894, 2011.

[9] M. Elbanhawi, and M. Simic, "Sampling-Based Robot Motion Planning: A Review survey", IEEE Access, vol. 2, pp. 56-77, 2014.

[10] S. M. LaValle and J. J. Kuffner, “Randomized kinodynamic planning,” International Journal of Robotics Research, vol. 20, pp. 378–400, May 2001.

[11] J. J. K. Jr. and S. M. Lavalle, “Rrt-connect: An efficient approach to single-query path planning,” in Proc. IEEE Intl Conf. on Robotics and Automation, pp. 995–1001, 2000.

[12] Koenig, S., Likhachev, M. and Furcy, D., 2004. Lifelong planning A∗. Artificial Intelligence, 155(1-2), pp.93-146.

[13] Nilsson, N. 1971. Problem-Solving Methods in Artificial Intelligence. McGraw-Hill.

Universal Biological Motions for Educational Robot Theatre and Games

Rajesh Venkatachalapathy, Martin Zwick
System Science, Portland State University
Portland, Oregon, USA

Adam Slowik
Department of Electronics and Computer Science
Koszalin University of Technology
Koszalin, Poland

Kai Brooks, Mikhail Mayers, Roman Minko, Tyler Hull, Bliss Brass, Marek Perkowski
Department of Electrical and Computer Engineering
Portland State University
Portland, Oregon, USA

***Abstract* — Paper presents a concept that is new to robotics education and social robotics. It is based on theatrical games, in which students create "biological", "characteristic" and natural motions for social robots and animatronic robots. Presented here motion model is based on Drift Differential Model from biology and Fokker-Planck equations. This model is used in various areas of science to describe many types of motion. The model was successfully verified on various simulated mobile robots and a motion game of three robots called "Mouse and Cheese".**

I. INTRODUCTION. PROBLEM FORMULATION

Games are already an important aspect of modern educational efforts on the levels of high school and undergraduate education because teenagers and young people are enthusiastic about playing and programming games. As an example, paper [41] presents how games can be used to teach logic minimization, circuit design and quantum mechanics. Paper [1] discusses video games in this context, while paper [3] and a comprehensive book [2] present various approaches and ideas of designing video games for education. Different types of games are also used in school theatres to train acting skills [4, 5, 9]. The robot social games are used to help autistic children gain interpersonal skills [7]. Many ideas about didactic and realization aspects of robot theatre are in [6], while [8] discusses the idea of combining games and robots in computer science education. An ambitious task of using robot theatre to promote STEM education for underserved students is discussed in [10].

In addition, robots are increasingly used to teach children, teenagers and university students about mechanical and electrical hardware design, programming, mathematics, gaming and human-interaction skills. Moreover, humanoid and animatronic robots are already used in several areas of human-robot interaction: (1) to interact with elderly and autistic children, (2) as robot museum guides and receptionists, (3) as home robots and popular toys, (4) as robot entertainers. Since 2001 the last author, Marek Perkowski, teaches teenagers about robotics, logic, mathematics and systems science approaches. One of the ways to achieve these educational goals is through creating an improvisational robot theatre that combines the ideas of teaching through robotics, through theatre and through games.

One of important problems in educational robotics is designing the motions of the robot that would be natural: human-like, or specific animal-like. This is also related to the art of animation in puppet theatre, movies and computer games. In robot theatre the motions can be not necessarily human-like but must be "*characteristic and believable*" or "*biological*". How to program an inexpensive robot built from an internet kit to "run like a dog"? To "eat like a nutria"?

As a more advanced example, let us discuss a hand-waving greeting gesture "hello" by a humanoid robot that uses one or two arms for his emotional gestures. If the greeting action would be always the same, the robot would be perceived as boring and unnatural by people who interact with this robot or watch the robot theatre performance. Simple solution to this problem is to design several greeting motions and select one randomly – unfortunately we observed in our many robot theatrical plays that this solution also becomes unnatural after few minutes of watching these repeated behaviors. Another solution may be to design a greeting motion as a sequence of completely random poses. These motions are quite unrealistic and also become not interesting soon.

Thus, several research directions have been tried to create interesting human-like and varying gestures for animatronic and humanoid robots. They include: (1) robot programming languages based on elementary motions and algebraic operators to combine them [12], (2) low-level robot motion editors to design motions (gestures) as sequences of postures (similar to graphic animations), (3) methods based on feedback from position sensors, (4) human motion acquisition and transformation to robot kinematics and dynamics, (5) heuristic systems for converting music to motions, (6) random number generators, (7) Machine Learning approaches, (8) methods based on spectral analysis of individual motion waves from sensors, (9) Perlin noise, and several others [20, 12, 29, 45 44, 38, 37, 14].

These methods and their extensions are used to create complex robot behaviors, which involve also simulated emotions and conversations in natural language. Each of these methods creates better or worse typical examples and variants of motions and behaviors, but as evaluated by psychologists, puppeteers, animators and theatre experts [12], the methods still do not satisfy the criteria of robot's sufficient theatrical realism, which level has been already achieved in the technical areas of graphic animations and games [12], and of course in the art of puppetry. No objective methods exist for the evaluation of behavior quality of such robots.

The problems that we address in this paper are the following:

(1) We outline a new model: realistic ("biological") motion generation problem, based on general principles of motions in Nature.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

(2) We present an interesting issue for robot theatre: "designing new motions for animatronic, humanoid and other theatrical and social robots". Should they have motions similar to humans? Similar to machines? To animals? To Fairy Tale Characters? To electrons in an atom? Attempt to design motions that are new and unexpected, interesting, symbolic, information-carrying and characteristic are important for improvisational robot theatre.

(3) We create a model of motions for robots playing motion-behavior games, like policemen chasing a thief or an airplane battle. The motions may be influenced by noise, have controllable errors or some other parameters so that the motions are not predictable by opponents but are interesting to watch by the game audience (our games are a kind of interactive, improvisational theatres, like a cheese escaping from hungry mice and asking verbally for mercy).

New methods to solve these three highly related problems should be thus developed, implemented, evaluated, and compared with the existing methods. This paper presents our attempt at creating such a new method for game-base improvisational educational robot theatres, distinct from all previous approaches known from the literature.

II. PREVIOUS RESEARCH

When we create improvisational robot theatre for education we want to base our thinking on ***system approaches*** that can solve wide categories of problems and have also deep educational aspects. These issues are discussed with teen teams while solving sequences of design examples, math problems, and programming exercises. Such system approach can be justified also on philosophical, conceptual and aesthetic grounds; it can also be a part of conventional scientific discovery processes. Different perspectives on this process have been offered [30, 37, 25, 22]. Hofstadter emphasizes the role of analogical reasoning in scientific discovery. Analogy is used in SAT tests and plays also a fundamental role in children theatre. Simon [37] offers a more computational approach and works from a discovery of knowledge as a foraging process (foraging is a common topic in plays about animals). Holland and Thagard build on this and the concept of coherence to develop a model of induction. Modern enterprises like Machine Learning find transfer learning to be a key concept underpinning many useful artificial systems.

Taken together, these models of scientific discovery require search and recognition heuristics. The models are meant to capture consistency criteria among scientific explanations, theory induction and other steps during scientific knowledge accumulation [25]. Such heuristics allow a transfer of concepts and models from one domain to another. In this paper, with the goal of creating educational robots and robot team behaviors, we identify these isomorphisms in certain classes of stochastic dynamical systems**,** we construct a notion of reliability as satisfaction of constraints, and we express them as ***first-passage time to a boundary*** (***FPT***). The probability distribution of the first exit out of a region in state space has a natural bio-semiotic interpretation, which in our case is applied to escaping and reaching robot motions in games and educational plays. Running, following, escaping and reaching actions are very common in all puppet plays and computer games. These proposed by us system-based methods have been however never used to generate robot behaviors. This seems to be the most important contribution of this paper from the theoretical point of view, while the practical aspect of our work is designing concrete realistic motions for our theatrical robots.

The biggest issue with the pre-programmed robot motion libraries for entertainment robots is that the robots can quickly become *unnatural* and *boring* to their audiences. This is because often there are only a few ready motions in the memorized motion library, and there are only few circumstances that the robot can respond to. Thus, the results are seemingly repetitive robot responses and the perception of limited interaction. Researchers in the animation field try to find ways to enable real-time response generation for virtual agents (e.g. video game characters). Bruderlin and Williams showed that by representing motion data as signals, common signal processing techniques such as multiresolution filtering, Fourier Analysis, wave shaping, time warping, and interpolation can be applied to the motion data [23]. As signals, motion data can be manipulated in real-time to be exaggerated, subdued, or blended with other motions while maintaining the characteristics of the original motion. Unuma et al. used Fourier Analysis to create transitions between two periodic motions using normalized coefficients [0,1] between the Fourier coefficients of the two motions [24]. Also, Unuma showed by using Fourier Analysis, *characteristic functions* (e.g. 'tiredness') can be extracted by calculating the difference between the coefficients of a 'neutral' motion (e.g. walk), and its variation (e.g. tired walk). The extracted characteristic function can then be applied to other motion (e.g. run) to create a similar characteristic on the other motion (e.g. tired run).

Ken Perlin developed a method to generate pseudo-random noises that can be used for creating noise in animation (one-dimension) to animated solid textures (four-dimensions) [20]. Perlin's noise-generating method (popularly known as *Perlin's Noise*) have been used to create noise in the movement of a virtual character or robot, to simulate those little movements such as breathing, blinking, fidgeting, or sways. Originally, Perlin Noise is often used to create and animate movements of textures in nature such as water, clouds, fire, and other elements. Perlin Noise is generated by creating a sequence of random noises. The sequence of noise usually starts with a smoothed (i.e. interpolated) low frequency noise to *n* number of higher frequency noise, where each subsequent noise frequency is an octave higher than the last (*fn = 2*fn-1*). The sequence of noises is then added together; with the contribution of each random noise decreasing as the frequency of the noise signal increases (noises with higher octave have less contribution than the lower octave noises to the final Perlin's Noise). The Perlin's Noise is especially useful to alleviate the 'static look' when a virtual character or robot is idle; instead of being still without any movement, the little movements (i.e. noise) give an impression of breathing or heartbeats, thus giving the illusion that the agent is 'alive.' Our approach is based on a similar philosophy of creating a special type of "noise", but we base the motions on Fokker Planck differential equations. Thus in addition to Perlin's noise the "Fokker-Planck Noise" is used to create motions for "fairy tale robot actors".

III. New Methodology Based on Universal Motions, Stochastic Dynamical Systems and Fine-Tuning

A. Motions and Emotions of theatrical robots

The motion of a social/theatrical robot should be realistic, interesting, human-like (animal-like, fairy-tale-like or robot-like), and demonstrating some specific "robot personality" and "robot emotion". The motions/behaviors creating method should produce stochastic variants to create very many unpredictable but similar motions, because this makes the robot more interesting and increases its repertoire of behaviors in games. On the other hand, the motion must be robust; which means that with high probability "strange and out of place" motions will not be generated by the stochastic motion generator. We can thus say that the motion-generating mechanism must be reliable. For instance, the motion of a small robot escaping from a light beam should be not programmed. This robot should behave always differently, but in principle it should be able to escape to a safe dark space at least in some cases. Similarly, a robot arm reaching for an apple located above should be ultimately able to grasp the apple. The generated motion should be also related to the environment: a robot control mechanism for "dancing waltz" should have always the same basic pattern, but the actual motion is different for a robot in an open area and a robot dancing in a narrow corridor.

Environment is responsible for some additional constraints in time, space or robot behavior type. As an example, a robot with 22 degrees of freedom (such as rotations in head, arms and knees) can be treated as a dynamic system that solves in real time the set of differential equations with at least 22 variables. These equations can be separate; in the simplest case we have a system of synchronized 22 one-variable differential equations. Parameters of these equations come from robot's memory (like emotional state of the robot, robot being happy or sad – the internal state of a hybrid state machine) or from the environment (like a distance of robot to a wall). In general, the robot-related variables in these differential equations can represent: spatial coordinates, velocities, radial velocities, accelerations, forces, rotation angles, emotions or sensor readings.

It is known from puppet theatre studies that puppet's motion is the most important artistic and educational aspect for very young audiences. Trying to develop a fundamental philosophy for robot theatres, we may ask, what are the most general motions that exist in the world? The linear motion and the circular motion come immediately to mind, as they are solutions of simple differential equations and are common: the motion of planets around the Sun, the motion of electrons in the atom. **Diffusion processes** are another motions that occur in biology, sociology and psychology [11,17,18,19,21,26]. Thus Fokker-Planck and Schrödinger equations come immediately to our mind as universal motion generators. What else? We assume that for a robot theatre one should dispose a system that could generate all basic motions that exist in the world, including cosmic and atomic motions. These types of motions are not only entertaining but have also important educational aspects.

B. Universal motions from differential equations.

One can say that mathematically a robot is a dynamic system of differential equations. Because of the above-mentioned requirements of realism, universality and interestingness, our preferred model is that of stochastic differential equations.

The questions arise:

(1) What equations and on what variables?
(2) How these equations are solved?
(3) How to model noise, slippage and hardware imperfections or even errors?

Differential equations are (together with Cellular Automata) two of the most general mathematical descriptions of system dynamics and problem-solving. They are used in physics, chemistry, biology, psychology and sociology, to list just a few. They describe dynamics of linear and non-linear systems, including systems based on agents. Agents can be robots or their subsystems. Although ***stochastic differential systems (SDS)*** are common in robotics, they have been not yet applied to any of the three groups of problems given in section I. The main idea of our research is to apply general differential equation types from various research areas, such as physics, chemistry, biology, psychology and sociology, to create interesting but reliable motions for theatrical robots. While [26] concentrates on models from biology, sociology and psychology, only a small subset of model equations from [26] will be illustrated in this paper.

C. Biological, psychological and social agents.

One can observe a surprisingly deterministic behavior of biological or social agents in the presence of stochasticity. Notable is the reliability of this behavior. The general nature of this behavior requires a study centered on **general principles**. This makes it an ideal candidate for a systems approach to various processes abstracted as motions. After constructing a specific notion of reliability, we find isomorphisms between biological, social and psychological processes [26] that demonstrate how the same notion of reliability developed in [26] can be used in biology for animal motions, in sociology to understand escape from poverty traps, and in psychology to understand decision making.

Along the way, we discover and discuss connections within related sub-fields in biology, sociology and psychology, illustrating the utility of system approach in scientific inquiry and mathematical exploration, with direct application to universal robot motion generation. This approach is useful mostly for humanoid assistive, theatrical, social and game-playing robots, but for simplification in this paper we present three-wheeled small mobile robots called Braitenberg Vehicles [32]. This is because of: **(1)** they can be built in short time even by young teenagers from inexpensive kits that many are advertised on Amazon, eBay and internet, (**2**) they have the simplest possible kinematics which simplifies programming of motions, **(2)** they allow for easy creation of robot swarms with interesting and non-trivial behaviors.

We formulate a hypothesis that some basic dynamic systems with their differential equations are common in many various areas – we can call them ***system-level motion isomorphisms***. Such systems are treated as describing general motions – a holistic idea that *"motions in various areas represent some characteristic patterns".* These patterns are then used by us as a new methodology for robot motion generation. These model similarities are the system isomorphisms: a trajectory of robot trying to avoid obstacles on his way to the closest door is similar

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

to a status/education/wealth trajectory of a human rising from poverty traps. Can we apply the same differential equations to describe the dynamics of both systems?

In the highly biology-centered perspective [48], the conceptual aspect is based on an assortment of ideas emphasizing biological semantics and the biological level of motions. It carries with it a general biology-like semantics. The mathematical level that is next used for mobile robot programming is more abstract than the conceptual aspect with no semantics associated with it. By lifting a scientific concept and treating it as a system concept, models associated with the original scientific concept are available to be used in disciplinary scientific inquiry in fields unrelated to the original scientific concept, robot motion theory and game applications in our case. The micro- and nano- scale physiological processes are modeled by a certain class of SDS. First-exits or first-passage times of these SDS are adequate descriptors of transient characteristics of these biological processes [23]. This is then used in [26] for scientific inquiry in sociology and psychology. For instance: (1) in sociology, the concept of reliability is associated with dynamics in and out of poverty traps (2) in psychology, the concept of reliability is associated with dynamic judgment and decision tasks. These systems dynamics are already used in theatres and can be used in robot theatre for symbolic motions.

The isomorphism between these two processes extends to the notion of **fine-tuning** and **constraint satisfaction**, both lifted from physiology [26]. Constraint satisfaction methods are already used in robotics and Fokker-Planck equations are already used in biology, but we believe that the system analogies of biological, sociological, psychological, financial and other "motions" can find even deeper applications in designing behaviors and teaching of several types of futuristic robots in future improvisational interactive theatres. The difference of our approach is that while other authors want to analyze Fokker–Planck equations to avoid effects of noise in robotics, we want to add several special types of "noise" for theatrical and game-playing effects.

D. Studying Animal Behaviors

Tinbergen's vision for ethology [40] is a good starting point for researching principles of animal behavior. His "Four Questions" are:

(1) what is its physiological causation?
(2) what is its function or survival value?
(3) how has it evolved over time?
(4) how has it developed in the individual?

Shettleworth [11] summarizes these four questions to be about cause, function, evolution and development respectively and added fifth question on intra-level and inter-level integration of different disciplines answering these *"Four Questions*". Webb [13] expands the discussion to include behavior of artificial agents and is related to an early historical work by Brooks [15], a paper that caused a break-through and initiated a new wave of research in robotics. And finally, focusing on higher organisms, Bullock's work on comparative neuro-ethology [16] emphasizes diversity of neural mechanisms that generate behavior. This can in turn find applications in *"group robotics", social robotics* and "*improvisational robot theatre"* [12, 35, 36]. Animal behavior is typically documented using ethograms [27], recordings of behavioral repertoires decomposed into behavioral primitives. An understanding of these primitives is of general interest to ethologists. Animal behaviors can be captured the same way as human motion is acquired and used to control "biological" and humanoid robot behaviors [38]. These behavioral primitives can be mathematically characterized, modeled, transformed and generated [12, 13]. Among the various observed characteristics, a behavior's reliability is the core focus of this paper. And this reliability is often defined by time-to-event constraints that need to be satisfied in the presence of stochasticity. Here we restrict ourselves to two such primitives: Escape behavior is evasion in response to typically harmful environmental signal. Startle response is an abrupt and fast behavior and may become a building block for more complex startles. From a modeling point of view, we can ignore the subtle differences between escapes and startles and in what follows focus on startles, treating escapes to be conceptually synonymous. Fixed action patterns, on the other hand, are also triggered by environmental cues, but are spatiotemporally complex and longer timescale behaviors. It should be pointed out that our goal is not only to build an interesting robot theatre with teens, but we also want to teach them mathematics (Fokker-Planck), biology (FP in animal motions), programming, game theory and understanding the relation between abstract differential equations and other models and actual motion of the robot related to physical, mechanical design and control aspects such as noise and slipping.

E. Fixed Action Patterns, Gaits and characteristic behaviors

Consider an organism with or without a nervous system whose body size and weight are small enough so that environmental fluctuations induce instability in the organism's movement. Suppose the organism is exposed to sudden harmful stimulus. In response to this environmental signal, the organism must reliably escape or respond to avoid consequences. Adaptation to exposure of harmful signals may occur evolutionarily, developmentally or over its life history. Similarly, consider a situation where the organism, upon presentation of a cue, is expected to locomote in space. This locomotion, ***a fixed action pattern (FAP),*** requires the system repeat its body configurations over an extended but finite period of time, in a stochastic environment. The disturbance may wobble and destabilize locomotion but not by much. FAP is a fundamental method in simple walking robots such as hexapods. In the case of escape and startles, ideal behavior is one that completes the act by a definite time. In the case of FAP, ideal behavior is one that maintains the pattern for a definite time. Both primitives involve time constants; in case of startles and escape, it is time to viability; in the case of FAP, it is persistence of viability. Typically, both these kinds of behaviors are modeled either using optimality principles or are built ad-hoc to reproduce behavioral templates. A key question is the source of reliable behavior under varying conditions of internal and external stochasticity. Reliable startle behavior involves movement of the body of the organism in an environment; it involves avoiding a subset of space of possible configurations while reaching a subset of final configurations within a fixed time. Reliable fixed action pattern involves repetitive movement of the body in an environment, while being subjected to uncertainty in conditions both internal and external to the organism. The notion of reliability is usually given an average case interpretation. According to this, behavior is reliable if it performs well most of the time. Behavior is reliable if we can guarantee that the agent's state is viable, even in the worst-case. Also, reverse engineering

explanations make use of control-theory-based modules to explain reliability. Instead, we find constraint-satisfaction-based reliability guarantees more appealing. Our focus on dynamic systems reliability can be justified biologically [16], and it can also be justified by an argument similar to that of Brooks [15], who denied the need for representation in robot control. One can instead think of behavior as being generated by making use of internal representations. One potential remedial strategy might be to think of behavior as being generated by merely a biological substrate. One can call such a lower form of internal process ***physiological***, as opposed to ***cognitive***; and can correspondingly consider the dynamic systems type of reliability discussed in this paper as also lower than the type of reliability that is achieved via control modules. Our paper does not seek to replace the cognitive; it merely suggests that some aspects of control can be replaced with a lower form of behavior. An interesting example is the crawling locomotion of soft-bodied animals [33]. Such repetitive behavior is usually generated by ***central pattern generators (CPG).*** A neuro-mechanical model that enables rhythmic movement by coupling the neuro-muscular dynamics with the environment, via friction generated by the body, is able to generate full organism coordination of patterned proprioceptive crawling. In this example, even if nervous system is involved, the locomotion is not due to precisely controlled neural circuits. It is because of fine tuning of parameters of the coupled dynamics. This example illustrates one of the main points of the paper: fine tuning with or without constraint satisfaction - under some circumstances - replaces the need for advanced control modules.

IV. Motion Generation for Actors of the "Mice and Cheese Game"

A. Control of Actors – the Braitenberg Vehicles

Although our methods are predominantly for robots or swarms with many degrees of freedom (system variables) here for simplicity we describe a simple mobile robot with two sensors, two wheels and two motors. Such robots are used much in our research and theatres [41] and we illustrate many concepts and design procedures with them [36]. This model of kinematics is better than a car or tank models in case of social and theatrical robots. All these robots start from the concept of the so-called ***Braitenberg Vehicles***, presented in a book [32] which deeply influenced Brook's approach [15] to robot design and control.

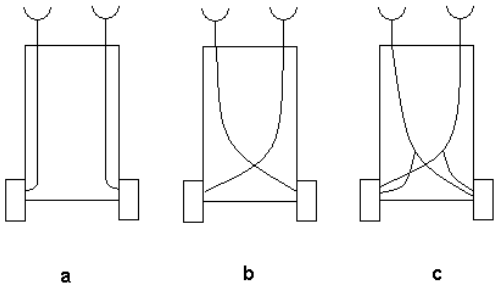


Fig. 1. The simplest Braitenberg Vehicles with analog control, (a) each sensor is connected to the motor on the same side, (b) each sensor connected to the motor on opposite side, (c) both sensors connected to both the motors.

Valentino Braitenberg wrote a revolutionary book [32]. In this book he describes a series of thought experiments. It is shown in these experiments how simple systems (the vehicles, our robots) can display complex life-like behaviors far beyond those which would be expected from the simple structure of their 'brains'. He describes the law called the *"law of uphill analysis and downhill invention"*. The law explains that it is far easier to create machines that exhibit complex behavior than it is to try and build the structures from the behavioral observations. By connecting simple motors to sensors, crossing wires and making some of them inhibitory, we can construct simple robots that could demonstrate behaviors similar to fear, aggression, love, affection, and other.

Original Braitenberg vehicles use just analog signals or Boolean logic, but we have generalized the controllers that exist between sensors and effectors to multiple-valued, fuzzy, probabilistic and quantum logics. Also by adding memory to these combinational logias, our controllers are generalized to state machines. Finally, we create ***hybrid automata controllers***. These automata used in our paper can be treated as standard Moore automata with every internal state corresponding to certain motion, behavior type or emotion. The input states come from sensors (vision). Outputs are motion signals such as velocities given to wheels. Some set of differential equations (such as Fokker-Planck) correspond to every internal state. They generate wheel velocities for each wheel individually. Thus our hybrid automata are more general than standard automata because in standard automaton we have a constant output signal and in hybrid automaton we have a dynamical process that continues until the automaton remains in its given internal state.

Let us explain what we understand by robot emotions and behaviors. The vehicles from Fig. 1 have two sensors and two motors, right and left. These vehicles can be controlled by the way the sensors are connected to the motors. Braitenberg defines three different basic ways we could possibly connect the two sensors to the two motors: (1) Each sensor connected to the motor on the same side, (2) Each sensor connected to the motor on the opposite side, (3) Both sensors connected to both the motors. Type (1) vehicle will spend more time in places where there is less of the stuff that excites its sensors and will speed up when it is exposed to higher concentrations. If the source of the light (for light sensors) is directly ahead, the vehicle may hit the source unless it is deflected from its course. If the source is to one side, one of the sensors, the one nearer to the source, is excited more than the other. The corresponding motor turns faster. As a consequence, the vehicle will turn away from the source. Turning away from the source (called in [32] a shy behavior) is illustrated at left in Fig. 2a.

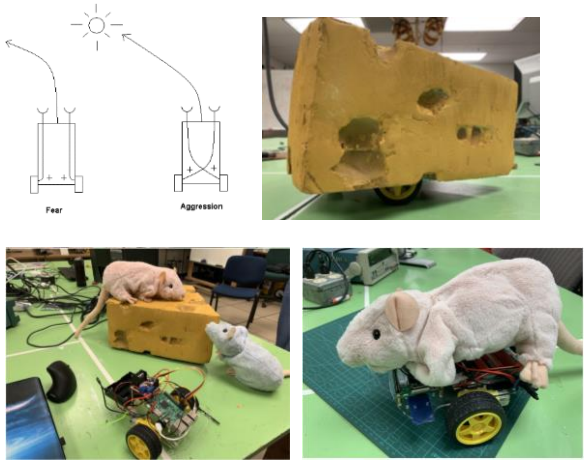


Fig.2.(a) Left Vehicle avoids light while right vehicle follows light.(b) A cheese robot, (c) red and blue mice and the base robot, (d) blue mouse.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

We can observe another type of vehicle, type (2) vehicle with positive motor connection. No change if the light is straight ahead, a similar reaction as seen in type (1). If it is to a side, then we observe the change. Here, the vehicle will turn towards the source and eventually hit it. As long as the vehicle stays in the vicinity of the source, no matter how it stumbles and hesitates, it will hit the source frontally, in the end. If the two vehicles are let loose in an environment with sufficient stimulus sources, then their characters emerge. Their characters are quite opposite. The type (1) vehicle with positive connection will become restless in their vicinity and tends to avoid them, escaping until it safely reaches a place where the influence of the source is scarcely felt. The feeling of fear is attributed [32] to this vehicle. Vehicle of type (2) with positive connection turns towards the source of light. It resolutely turns towards them and hits the source with high velocity, as if it wanted to destroy them. The aggressive feelings displayed clearly. It is really amazing to observe complex behaviors of these simple robots and their teams. The chasing or escaping actions attributed to the state of the hybrid automaton can be executed in different ways resulting from parameter values of differential equations solved in these internal states and used to generate motions.

By controlling individually the left and right wheels with velocities ***v(t)*** being solutions to different SDS (as discussed in sections II, III, and IV) we create a high variety of robot behaviors. Braitenberg Vehicles have the so-called "*cart kinematics*", the most common and the easiest kinematics used in small robots. In such cart robots if the same velocity is given to both wheels – robot drives forward, if a higher velocity is given to the right wheel the robot keeps turning left, if the higher velocity is given to the left wheel the robot keeps turning right. If the left wheel does not move and the right wheel turns forward the robot turns around the center being the left wheel. If the right wheel turns with velocity ***v*** and the left wheel turns with velocity ***-v*** the robot rotates counterclockwise around its center of gravity. It was experimentally observed that, as explained in simple examples above, with sequences of even just constant value velocities, many interesting "mobile robot gaits" are created from their simple combinations [12]. When two sinusoids are given to left and right wheels the walking pattern is created, like using left and next right leg with body rotations. When the velocity of every wheel (Fig.1) varies individually with added noise, the robot can move in very many stochastic patterns.

Noise-like processes (like Perlin or Diffusion Drift) can be added on top of any deterministic motion generations. For instance, Fokker-Planck model has parameters like ***D, v*** and ***w*** of equations, and the animator or the supervising software can experiment with their values. Natural noise or mechanical disturbances are simulated with ***w***. These values can be found using Nature-Inspired Methods as well [14]. In another example, in one of experiments with another robot a motion was developed for a golf-playing robot arm but next applied to the neck-moving behavior of a humanoid - interesting and unexpected gesture was thus created. We found this "*general motion transfer*" property to be useful in several robot theatre applications. It was observed that the motions transferred from area to area and from robot to robot can be unexpected, interesting and innovative.

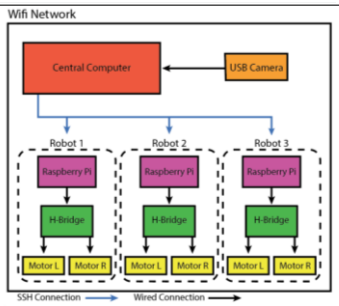


Figure 3. The Mice and Cheese System with ceiling camera and Wifi Network.

B. The Mice and Cheese Chasing Game

Fig. 2b presents the robot cheese, Fig. 2c two mice and the base mobile robot and Fig. 2d the blue mouse robot. These are the actors of our "game theatre" – in addition to moving the robots talk to represent their emotional states such as fear. These are all Braitenberg Vehicle robots playing the roles of two mice (blue and red) that chase the cheese robot in a stochastic game of "policemen and thief" type. Chasing games and fighting games are entertaining, like in a puppet theatre. The environment in this project is a large triangle decomposed to small triangular grids with only one escape door for the cheese. The robots start from random location each in a limited triangular grid. The cheese wins when it escapes from the game area through the door. The mice win when they catch the cheese; it means they surround it in such a way near the wall that the cheese cannot move to any neighbor triangle. Each mouse has collaborative motives (catch the cheese) and egoistic motives (being first to catch it). For the lack of space we do not explain the game algorithm or technical aspects of robots. Fig. 3 presents the entire system with the ceiling camera that permanently monitors the state of game and informs the control computer. The colors of the robots are for the ceiling camera to recognize robots' positions. Behaviors of robots are based on hybrid automata. The automata internal states correspond to Fokker-Planck equations with various parameters. With parameters $\mathbf{D} = \boldsymbol{w} = 0$ the automaton and the game are deterministic. However, with $\boldsymbol{w} \neq 0$ the game is stochastic and more difficult, but also more interesting to play and observe. Mice and cheese behave differently at every time. We treat this game as a simple example of a *game-based improvisational, interactive robot theatre.* Next section presents details of our realization. In this model each of the two wheels is controlled by a single-variable differential equation (the third wheel of the Braitenberg Vehicle is not controlled).

C. Fokker-Planck Equation for motion control.

An exact formalization of reliable behavior in the presence of stochasticity that unfolds in time and in space requires the model to be a dynamical system, contain stochasticity, have a continuous state space, and evolve in continuous time. The simplest possible system with all these features is a stochastic dynamical system with ***Brownian noise***.

$$dx(t) = v(x)\,dt + \sqrt{(2D)}\,dw \qquad (1)$$

where: $\boldsymbol{x}$ is a *one-dimensional state space*, $\boldsymbol{D}$ is the *diffusion constant* for *Brownian noise* $\boldsymbol{w}$, $\boldsymbol{v(x)}$ is called the *drift coefficient* and can depend on the state space. If one sets D = 0, the noise term vanishes and what one gets is a deterministic nonlinear dynamical system

$$dx(t) = v(x)dt \quad (2)$$

While this connection with deterministic dynamical systems make the system appealing, it is not clear how to perform an analysis of such a stochastic evolution. In order to do this, one needs a mapping to a more rigorous formulation in terms of *measure theoretic probability* [28]. In that formulation what is more natural is the notion of Markov processes. Markov processes are stochastic processes that satisfy the following independence relation

$$P(past,future|present)=P(past|present)P(future|present) \quad (3)$$

and can also be rewritten as

$$P(future \mid present, past) = P(future \mid present) \quad (4)$$

making it a statement about what knowledge is required to adequately albeit probabilistically know the future of the system at any given present moment. From this point of view, ordinary differential equation based dynamical systems are *Markov processes*. Both equations (3) and (4) can be made precise in measure theoretic terms and can be rewritten for any stochastic evolution on any kind of state space. In particular, when the state space is discrete and finite and the evolution is discrete. It becomes ***a discrete time Markov chain (DTMC)*** whose evolution is uniquely determined by the transition matrix

$$T_{ij} = P(j \mid i) \quad (5)$$

where ***P(j | i)*** is the probability of making a transition from state ***i*** to state ***j*** at any given discrete time instant. For simple state spaces like state spaces on real line ***R*** and line segment [0, 1], it is equivalent to the following ***partial differential equations (PDE)*** below. Equ. (6) is example of PDE defined by Fokker-Planck operators.

$$\frac{\partial c(x,t)}{\partial t} + v\frac{\partial c(x,t)}{\partial x} = D\frac{\partial^2 c(x,t)}{\partial x^2} \quad (6)$$

where *c(x, t)* is the probability density of finding the system in state ***x*** at time ***t***. Note that this PDE is deterministic, even though the underlying system is stochastic. This equation is the classic ***diffusion equation*** originally constructed to model evolution of heat conduction in extended finite temperature materials and is identical to ***Schrödinger's equation*** known from Quantum Mechanics. It captures probabilistic evolution of particle in the absence of drift ***v***. Just as ordinary DE requires some initial (boundary) conditions to make them completely and uniquely specified, PDEs require *boundary conditions*. Taken together, the equation encodes all the knowledge one can acquire about the system. In classical mechanics [20], the Lagrangian $L(q, \dot{q})$ is defined as a function of ***x*** and $\dot{x}$, the position and velocity of the system. In an equivalent Hamiltonian formulation, the Hamiltonian ***H(p, q)*** is defined as a function of ***x*** and ***q***, the position and momentum of the system. An important observable in SDS and stochastic processes in general is ***first-exit time***, the time it takes for the state of a system to cross a certain threshold. In higher dimensions, it makes sense to think of exit from a domain with domain boundary. One can associate with first-exit, a probability distribution function for the ***first-exit time*** or ***first-passage time (FPT)***, a random variable in this system.

$$u(x,T) = P\{\tau < T | \mathbf{x}(t=0) = x\} \quad (7)$$

where ***u(x, T)*** is the cumulative probability of a particle starting at ***x*** to reach the boundary before time ***T***. ***D*** is the domain and ***Bd(D)*** its boundary. The associated random variable is represented by ***τ***. ***L*** is the Fokker-Planck operator. The associated PDE for the cumulative probability is given by

$$\begin{aligned} \frac{\partial u(\mathbf{x},t,T)}{\partial t} + \mathcal{L}u(x,t,T) &= 0 \quad \text{for} \quad x \in \mathcal{D}, t < T \\ u(x,t,T) &= 1 \quad \text{for} \quad x \in Bd(\mathcal{D}), t < T \\ u(x,,t,T) &= 0 \quad \text{for} \quad x \in Bd(\mathcal{D}) \end{aligned} \quad (8)$$

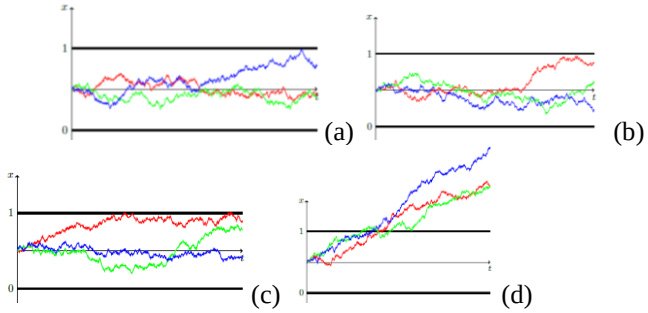


Fig.4. (a) Drift Diffusion Model (DDM) with no drift and small diffusion constant (b) DDM with very small drift and small diffusion constant, (c) DDM with small drift and larger diffusion constant, (d) DDM with large drift

Fig. 4 presents some basic examples of typical paths of SDS with constant diffusion coefficient and constant drift. The boundary points are 0 and 1. As the velocity (drift constant) increases, one can see that almost all paths exit boundary point 1. This is in contrast to the case where drift is zero. First-exit problems and analysis surrounding them are useful ways of thinking about transient nature of stochastic processes and have found their use in the sciences and in engineering disciplines [24]. In many non-equilibrium systems, traditional equilibrium analysis washes away details about non-asymptotic dynamics and patterns. In this paper, we identify transients and their timescales as ways to characterize biological reliability. In robotics terms one can think about ordinate as time axis and abscissa (denoted as ***x*** in Fig. 4) as a distance, velocity, radial velocity or rotation angle; the choice depending on robot design, kinematics and system of equations. Velocities from Fig. 4 are added to the fixed velocities of each wheel individually to create various gaits, behaviors and transitions to the neighbor triangular grids of the game.

V. CONCLUSION

As a part of a larger project *"teaching by games and improvisational robot theatres"* in this paper we sketched an innovative general methodology to create motions for all kinds of assistive and theatrical robots. It is based on the concept of universal motion processes that exist in Nature and are described by Fokker-Planck differential equations. Introduced robot motions generalize physiology and lift it to the systems level. In particular, by noting the surprisingly deterministic behavior of physiological processes the concept of reliable behavior is generalized and the models of nano and micro physiology, as well as atomic forces, are

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

lifted to the systems level and applied to "fairy-tale" robot programming – visualization of various dynamical processes. The concepts of stochastic differential systems and first-passage time to a boundary were illustrated for mobile robots and used to generate "*biological*" patterns of individual and swarm motions and behaviors with new applications in robotics, such as generating a variety of gaits for Braitenberg Vehicles and robot games.

The system approach in this paper has two aspects: **(1)** We designed a game for three robots and observed various interesting special cases of behaviors that result from unpredictable behaviors of diffusion models as well as theatrical values of generated motion patterns. **(2)** We discuss with young students the mathematical and system-level questions related to the design of robot motions that allow us to show the uniformity of several processes in the world and also, how to apply system reasoning to practical problems. The role of analogy and general patterns of motion is emphasized.

Future works.

(1) In quantum mechanics, the Euler-Lagrange PDE of Fokker-Planck is replaced by the Schrödinger equation where the evolution of wave function is determined by the Hamiltonian operator. The Hamiltonian is the generator of probabilistic evolution. Nano-scale Quantum Robots are presented in [31] and Quantumly-Controlled Braitenberg Vehicle robots are discussed in [36, 39]. Combining the quantum automata with Schrödinger equations we will be able to create new types of "quantum motions" that will generalize motions from this paper.

(2) In entertainment robotics, like here, the evaluation of Human-Robot-Interaction quality is subjective. Thus special procedures and methodologies have been developed to evaluate HRI [12]. We plan to apply them to similar "quantum biological motions" in the future.

REFERENCES

[1] B. Gros, Digital Games in Education: The Design of Games-Based Learning Environments, Journal of Research on Technology in Education, 2007, 40(1), pp. 23-38.

[2] M. Prensky, (2001). Digital game based learning. New York: McGraw Hill Press.

[3] K. Squre, (2005). Game-based learning: Present and future state of the field. Madison, TI: University of Wisconsin-Madison Press.

[4] K. Johnston, 10 Theatre Games Perfect For Drama Class, https://theatrenerds.com/10-theatre-games-perfect-drama-class/

[5] Drama Games for Kids, https://www.bbbpress.com/dramagames/

[6] Lecture By Professor Marek Perkowski - Hosted By Bytes of kitkatshttps://search.myway.com/search/video.jhtml?n=7876328d&p2=%5EY6%5Exdm277%5ETTAB03%5EUS&pg=video&pn=1&ptb=13CA0213-C765-4115-B916-FC7856E87135&qs=&searchfor=Perkowski+robot+video&si=EAIaIQobChMI4N6Az8at7gIVryx-Ch16KAfOEAEYASAAEgKu6_D_BwE&ss=sub&st=tab&tpr=sbt&trs=mv3

[7] Social robots for autism education, https://www.ucl.ac.uk/ioe/research-projects/2020/sep/social-robots-autism-education

[8] D. Xu, D. Blank, and D. Kumar, Games, robots, and robot games: complementary contexts for introductory computing education, GDCSE '08: Proc. 3rd Intern. Conf. on Game development in computer science education, Febr. 2008, pp. 66–70, https://doi.org/10.1145/1463673.1463687

[9] C. Barker, (2010) Theatre games: A new approach to drama training, books.google.com https://www.google.com/books/edition/Theatre_Games/qIlhT6Bu7YC?hl=en&gbpv=1&dq=theatre+games+in+education&pg=PR11&printsec=frontcover

[10] M. Jeon, M.F. Hosseini, J. Barnes, Z. Duford, R. Zhang, J. Ryan, and E. Vasey, Making live theatre with multiple robots as actors bringing robots to rural schools to promote STEAM education for underserved students, https://ieeexplore.ieee.org/abstract/document/7451798?casa_token=zZyr7W2aWaYAAAAA:0_Me4LEAahZCXn3eh7Opj4wavVRBUvihgAZkqT2V3O130O57UcbiuE4AqR2BJEzXEwM2uOgxw

[11] S.Shettleworth, *Cognition, Evolution, and Behavior*. Oxford Univ. Press, 2010.

[12] M. Sunardi, Synthesis of Expressive Behaviors for Humanoid Robots, Ph.D. Dissertation. ECE PSU, June 2020.

[13] B. Webb. Chapter 1: Using robots to understand animal behavior. volume 38 of Advances in the Study of Behavior, pp. 1-58. Academic Press, 2008.

[14] A. Slowik, H. Kwasnicka, Nature inspired methods and their industry applications – Swarm intelligence algorithms, IEEE Transactions on Industrial Informatics, Vol. 14, Issue 3, pp. 1004-1015.

[15] R.A. Brooks. Intelligence without representation. AI, 47(1-3):139-159, 1991.

[16] T. Bullock. How do Brains Work?: Papers of a Comparative Neurophysiologist. Contemporary Neuroscientists. Birkhauser Boston, 2013.

[17] J. Cartailler, Z. Schuss, and D. Holcman. Geometrical effects on nonlinear electrodiffusion in cell physiology. J. Nonlinear Sci, 27(6):1971-2000, Dec 2017.

[18] Y. Forterre, J. Skotheim, J. Dumais, and Lakshminarayanan Mahadevan. How the venus flytrap snaps. Nature, 433(7024):421, 2005.

[19] J. Gjorgjieva, G. Drion, and E. Marder. Computational implications of biophysical diversity and multiple timescales in neurons and synapses for circuit performance. Current Opinion in Neurobiology, 37:44-52,2016.

[20] K. Perlin, "An image synthesizer," ACM Siggraph Computer Graphics, vol. 19, no. 3, pp. 287–296, 1985.

[21] S. Grillner and A.M. Graybiel. *Microcircuits: The Interface Between Neurons and Global Brain Function*. Dahlem workshop reports. MIT Press, 2006.

[22] D.R. Hofstadter and Fluid Analogies Research Group. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* Penguin Press Science Series. Penguin Books, 1998.

[23] A. Bruderlin and L. Williams, "Motion signal processing," in Proceedings of the 22nd annual conference on Computer graphics and interactive techniques – SIGGRAPH '95, 1995, pp. 97–104, ISBN: 0897917014.

[24] M. Unuma, K. Anjyo, and R. Takeuchi, "Fourier principles for emotion-based human figure animation, "in Proc.eedingsofthe 22 ndannual conferenceon Computer Graphics and interactive techniques, 1995, pp. 91–96.

[25] J.H. Holland, K.J. Holyoak, R.E. Nisbett, and P.R. Thagard. *Induction: Processes of Inference, Learning, and Discovery*. MIT Press, 1989.

[26] R. Venkatachalapathy, Systems Isomorphisms in Stochastic Dynamic Systems, PSU, Systems Science, Dissertation, September 2019

[27] K. Immelmann, C. Beer. *A Dictionary of Ethology*. Harvard Univ Press, 1989.

[28] O. Kallenberg. Foundations of Modern Probability. Probability and Its Applications. Springer New York, 2002.

[29] R. Laban and F. C. Lawrence, *Effort: Economy of human movement*. Macdonald & Evans, 1979.

[30] P. Langley, H. Simon, G. Bradshaw, and J. Zytkow. Scientific Discovery: Computational Explorations of the Creative Processes. MIT Press, 1987.

[31] P. Benioff, Quantum robots and environments, Phys.Rev. A**58**, 893, 1998.

[32] V. Braitenberg, *Vehicles, experiments in synthetic psychology*, Cambridge, Mass. MIT Press, 1986.

[33] P. Paoletti and L. Mahadevan. A proprioceptive neuromechanical theory of crawling. Proc.Royal Soc. B: Biological Sciences, 281(1790):20141092, 2014.

[34] S. Redner. *A Guide to First-Passage Processes*. Cambr. Univ. Press, 2001.

[35] M. Perkowski, T. Sasao, J.-H. Kim, M. Lukac, J. Allen, S. Gebauer, Hahoe KAIST Robot Theatre: learning rules of interactive robot

behavior as a multiple-valued logic synthesis problem, ISMVL'05, 19-21 May 2005.

[36] A. Raghuvanshi, Y. Fan, M. Woyke, and M. Perkowski, Quantum Robots for Teenagers, Proc. ISMVL07, page 18, May 13 - 16, 2007.

[37] H.A. Simon. Models of Discovery: and Other Topics in the Methods of Science. Boston Studies in the Phil. History of Science. Springer Netherl. 2012.

[38] https://www.bing.com/videos/search?q=motion+acquisition+robot+humanoid&qpvt=motion+acquisition+robot+humanoid&FORM=VDRE

[39] M. Perkowski, "Quantum Robots. Now or Never?" Main address of the conference, *Proc. 5th Nat. Conf. Inform. Techn*, Gdansk, Poland, May 20-23, 2007.

[40] N. Tinbergen. On aims and methods of ethology. Zeitschrift fr Tierpsychologie, 20(4):410-433

[41] M. Perkowski, and K.Liu, Binary, Multi-Valued and Quantum Board and Computer Games to Teach Synthesis of Classical and Quantum Logic Circuits, Proc. ISMVL 2021.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Modeling Multidimensional Communication Lattices with Moore Neighborhood by Infinite Petri Nets

Dmitry A. Zaitsev
Department of Information Technology
Odessa State Environmental University
Odessa, Ukraine
daze@acm.org

Tatiana R. Shmeleva
Department of Computer Science
State University of Intelligent
Technology and Telecommunications
Odessa, Ukraine
t.shmeleva@onat.edu.ua

Peyman Ghaffari
Department of Mathematics
University of Aveiro
Aveiro, Portugal
pgsaid@fc.ul.pt

***Abstract*—Multidimensional torus topology plays a key role as a topology of the communication system of supercomputers and clusters as well as networks on chip. An infinite Petri net model of multidimensional torus communication grid with Moore neighborhood and combined cut-through and store-and-forward switching device has been constructed, its place invariance proven based on solving infinite linear Diophantine systems of equations in parametric form. For specifying Moore neighborhood, that traverses all hypercube bounds of lesser dimension, we use designation of the switching device ports by the coordinate difference. It is mentioned that the obtained results are applicable for modeling brain and processes of spreading insects and viruses.**

Keywords—multidimensional torus; communication lattice; Moore neigborhood; cut through; store and forward; infinite Petri net

I. INTRODUCTION

A multidimensional torus plays a key role as the topology of communication system of high-performance supercomputers and clusters [1] lately. For instance, the most powerful, at present, computer Fugaku [2] uses Tofu Interconnect D [3] which topology can be generalized as 6D torus. Moreover, modern multicore processors use network-on-chip [4], to connect cores and other units, which topology also represents plane or multidimensional lattice.

An infinite Petri net [5],[6] is a convenient tool for modeling communication and computing lattices. Together with reenterable models in the form of colored Petri nets [7] they represent a toolset for verification of protocols and performance evaluation of networks, grids, and clouds.

Due to minimalistic concerns, majority of known communication lattices uses von Neumann neighborhood for connection of nodes [8]. Moore neighborhood possess a series of advantages such as lesser distance between nodes, greater number of minimal length alternative paths, and others though a lattice with Moore neighborhood contains essentially greater number of connections. Hardware is getting considerably cheaper and more compact that opens prospects for practical application of communication lattices with Moore neighborhood.

The work has been partially supported by EU COST Action CA 16227

In known works [5]-[7], models of communication lattices with von Neumann neighborhood have been studied. The goal of the present work is a generalization of two dimensional lattice model with Moore neighborhood [9] on multidimensional case that requires development of a new system for ports (neighbors) enumeration. Also we generalize realistic, combined cut-through and store-and-forward switching node model [10], on multidimensional lattice with Moore neighborhood.

II. ENUMERATING NEIGHBORS IN VON NEUMANN AND MOORE NEIGHBORHOODS

In *d*-dimensional space, a node (switching device) is represented by a unit-size hypercube and is uniquely specified by its coordinate *d*-vector with nonnegative integer components (we start enumeration from zero), *k* is the lattice size

$$\bar{i} = (i_0, i_1, ..., i_d), 0 \le i_j < k, 0 \le j < d \ .$$

In von Neumann neighborhood [8], node neighbors share common *(d-1)*-dimensional hypercubes. For instance, in *2*-dimensional case, neighbors share the square sides, in *3*-dimensional case, neighbors share the cube facets. Thus, in *d*-dimensional space there are *2d* neighbors because there are two opposite facets in each dimension.

In Moore neighborhood [8], neighbors of a node share *h*-dimensional hypercubes for $0 \le h < d$. Remind that, *0*-dimensional hypercube is a point, *1*-dimensional hypercube is a section, *2*-dimensional hypercube is a square, and *3*-dimensional hypercube is a cube. Thus, in *d*-dimensional space, there are $3^d - 1$ neighbors because we exclude the central node from a hypercube of size 3 in *d*-dimensional space.

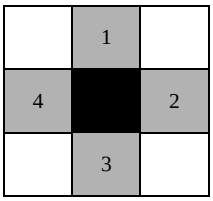
Let us consider enumeration of neighbor nodes (cells) with regard to the node (i_0, i_1) in Fig. 1a) and examples of enumeration of neighbors (ports that connect a node with its neighbors) for von Neumann (Fig. 1b) and Moore (Fig. 1c) neighborhoods on plane (2D). This simple clockwise system of

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

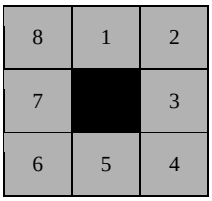
port enumeration has been applied in works dedicated to *2*-dimensional case only [5],[9],[10].

(i_0-1, i_1-1)	(i_0-1, i_1)	(i_0-1, i_1+1)
(i_0, i_1-1)	(i_0, i_1)	(i_0, i_1+1)
(i_0+1, i_1-1)	(i_0+1, i_1)	(i_0+1, i_1+1)

a) Coordinates of cells;



b) Von Neumann neighborhood;

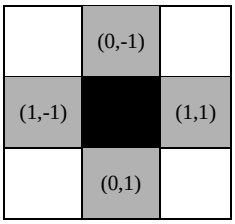


c) Moore neighborhood.

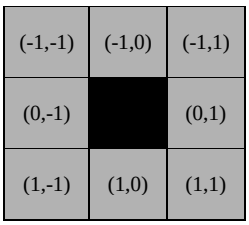
Figure 1. An example of neigborhoods on plane (2D) with enumeration of ports connecting neighbors.

For multidimensional grids, it is difficult to keep neighbors (ports) enumeration using a single number. Thus, for von Neumann neighborhood in multidimensional lattices, a special system of neighbors enumeration using 2 indices (m,r) has been applied [6], where *m* specifies the number of dimension and *r* specifies direction, represented in Fig. 2a) for *2*-dimensional case. Since only one coordinate changes in von Neumann neighborhood, for a given coordinate of node $\bar{i}$ and port number (m,r), the neighbor coordinate $\bar{i}'$ is calculated as

$$i'_j = \begin{cases} i_j, & j \neq m \\ i_j + r, & j = m \end{cases}, \quad 1 \le j \le d .$$



a) Von Neumann neighborhood index (m,r);



c) Moore neighborhood index $\bar{p}$.

Figure 2. Enumeration of ports connecting neighbors in multidimensional lattice, 2D example.

Here we offer to enumerate neighbors (ports) for Moore neighborhood in multidimensional lattices using their coordinate offset. Thus, the port index $\bar{p} = \bar{i}' - \bar{i}$ represents a *d*-vector with components which belong to the set {-1,0,1}, the neighbor indices are illustrated in Fig. 2 b). The neighbor coordinate $\bar{i}'$ is calculated as

$$\bar{i}' = \bar{i} + \bar{p} .$$

Thus, the neighbor index in von Neumann neighborhood has length 2 while in Moore neighborhood its length coincides with the node coordinate length and equals to *d*. Note that the plus operation in formulae for calculating $\bar{i}'$ is modulo *d* operation over nonnegative integer numbers as far as torus is concerned [11],[12]. Indeed, in torus on each dimension, we can reach given coordinate going ether clockwise or counter clockwise; usually we prefer the shortest path of two mentioned paths when their length is not equal.

For instance, in *4*-dimensional lattice of size 5, from (0,0,0,0) to (0,0,3,0), there are two following paths:

clockwise (0,0,0,0) → (0,0,1,0) → (0,0,2,0) → (0,0,3,0) and counter clockwise (0,0,0,0) → (0,0,4,0) → (0,0,3,0); we prefer the second path because it is the shortest (2 compared to 3).

We illustrate the neighbor indices with Table 1 where von Neumann neighborhood, included in Moore neighborhood, is written in bold font. Moore neighborhood is ordered on the decreasing dimension of the hypercube connecting neighbors: *(d-1)*-cube – one coordinate changes, *(d-2)*-cube – two coordinates change etc.

TABLE I. INDICES OF NEIGHBORS IN MOORE NEIGHBORHOOD

Number of dimensions	Moore neighbor indices (with included von Neumann neighborhood highlighted in bold)	Von Neumann neighbor indices
2	**(-1,0), (1,0), (0,-1), (0,1),** (-1,-1), (-1,1), (1,1), (1,-1)	(0,-1), (0,1), (1,-1), (1,1)
3	**(-1,0,0), (1,0,0), (0,-1,0), (0,1,0), (0,0,-1), (0,0,1),** (-1,-1,0), (1,-1,0), (-1,1,0), (1,1,0), (-1, 0,-1), (1,0,-1), (-1,0,1), (1,0,1), (0,-1,-1), (0,1,-1), (0,-1,1), (0,1,1), (-1,-1,-1), (1,-1,-1), (-1,1,-1), (1,1,-1), (-1,-1,1), (1,-1,1), (-1,1,1), (1,1,1)	(0,-1), (0,1), (1,-1), (1,1), (2,-1), (2,1)
4	**(-1,0,0,0), (1,0,0,0), (0,-1,0,0), (0,1,0,0), (0,0,-1,0), (0,0,1,0), (0,0,0,-1), (0,0,0,1),** (-1,-1,0,0), (1,-1,0,0), (-1,1,0,0), (1,1,0,0), (-1,0,-1,0), (1,0,-1,0), (-1,0,1,0), (1,0,1,0), (-1,0,0,-1), (1,0,0,-1), (-1,0,0,1), (1,0,0,1), (0,-1,-1,0), (0,1,-1,0), (0,-1,1,0), (0,1,1,0), (0,-1,0,-1), (0,1,0,-1), (0,-1,0,1), (0,1,0,1), (0,0,-1,-1), (0,0,1,-1), (0,0,-1,1), (0,0,1,1), (-1,-1,-1,0), (1,-1,-1,0), (-1,1,-1,0), (1,1,-1,0), (-1,-1,1,0), (1,-1,1,0), (-1,1,1,0), (1,1,1,0), (-1,-1,0,-1), (1,-1,0,-1), (-1,1,0,-1), (1,1,0,1), (-1,-1,0,1), (1,-1,0,1), (-1,1,0,1), (1,1,0,1), (-1,0,-1,-1), (1,0,-1,-1), (-1,0,1,-1), (1,0,1,1), (-1,0,-1,1), (1,0,-1,1), (-1,0,1,1), (1,0,1,1), (0,-1,-1,-1), (0,1,-1,-1), (0,-1,1,-1), (0,1,1,1), (0,-1,-1,1), (0,1,-1,1), (0,-1,1,1), (0,1,1,1), (-1,-1,-1,-1), (1,-1,-1,-1), (-1,1,-1,-1), (1,1,-1,-1), (-1,-1,1,-1), (1,-1,1,-1), (-1,1,1,-1), (1,1,1,-1), (-1,-1,-1,1), (1,-1,-1,1), (-1,1,-1,1), (1,1,-1,1), (-1,-1,1,1), (1,-1,1,1), (-1,1,1,1), (1,1,1,1)	(0,-1), (0,1), (1,-1), (1,1), (2,-1), (2,1), (3,-1), (3,1)

Note that we have: for 2 dimensions, 8 and 4; for 3 dimensions, 26 and 6; for 4 dimensions, 80 and 8 nodes in Moore and von Neumann neighborhood, respectively.

III. Distance and Routing within Lattices

Multidimensional torus topology essentially simplifies routing allowing us to use local rules only for packets delivery. For von Neumann neighborhood, local packet forwarding rules have been studied in [11]. Here we develop simple rules for lattices with Moore neighborhood.

At first, let us consider distance between nodes. In von Neuman neighborhood, the distance is specified by taxicab or Manhattan metrics where the distance is equal to the sum of distances on coordinates

$$d(\bar{i},\bar{i}') = \sum_{j=0}^{d-1} \left| \bar{i}' - \bar{i} \right| .$$

In Moore neighborhood, the distance corresponds to the maximal absolute value among the coordinate differences

$$d(\bar{i},\bar{i}') = \max_{j=0}^{d-1} \left| \bar{i}' - \bar{i} \right| .$$

To have formulae valid for both hypercube and hypertorus, we use regular signs of plus and minus though for hypertorus, modulo *k* operations should be applied. Note that there are two ways of computing the coordinate difference in hypertorus with regard to two possible directions of movement studied in [11].

Thus, in the best case, distance in Moore neighborhood is *d* times shorter than in von Neumann neighborhood, and, in the worst case, it is the same that together with huge amount of alternative shortest paths justifies application of Moore neighborhood.

In [11], we developed a series of local rules for fast packet switching in the lattice nodes, a dedicated simulator *ts* has been developed to justify rules statistically. The rules can be implemented as microprograms of switches that not only increases performance considerably because of routing table absence but facilitates to cybersecurity aspects of the lattice functioning.

The basic principle for packet forwarding rules is decreasing the distance at each hop. For von Neumann neighborhood, any coordinate with nonzero difference can be chosen. The same solution works in Moore neighborhood as well though we create a shorter path applying first "diagonal" moves with the maximal possible number of coordinates changing. Thus, for the best case with the same absolute value *v* for all coordinates, we achieve the destination node in *v* moves while in von Neuman neighborhood, $d \cdot v$ moves are required going alongside each coordinate separately.

For instance, in *4*-dimensional lattice of size 5, from (0,0,0,0) to (2,2,2,2) for von Neumann neighborhood, it will require $4 \cdot 2 = 8$ moves, say

(0,0,0,0) → (1,0,0,0) → (2,0,0,0) → (2,1,0,0) → (2,2,0,0) →

(2,2,1,0) →(2,2,2,0) →(2,2,2,1) →(2,2,2,2),

while for Moore neighborhood, it will require 2 moves only

(0,0,0,0) → (1,1,1,1) → (2,2,2,2) via port (1,1,1,1).

We offer the following simple rule for packet forwarding in a node within the Moore neighborhood lattice: *find the maximal subset of nonzero coordinates and choose port which index contains nonzero coordinates corresponding to the subset, preserving signs (absolute values equal to 1).* It can be proven in constructive way (or by contradiction) that the rule provides the packet delivery on the shortest path.

IV. Composing Lattice with Moore Neigborhood

Here we compose infinite Petri net model of multidimensional lattice with Moore neighborhood for the first time. Earlier, models of two dimensional and multidimensional lattices with von Neumann neighborhood [5]-[7] and two dimensional lattice with Moore neighborhood [9] that uses simple clockwise enumeration of ports (Fig. 2b) have been studied. As it is illustrated by Table 1, Moore neighborhood has essential differences compared with von Neumann neighborhood that is included into Moore neighborhood.

A. Model of Switching Node

The basic function of packet switching devices is the packet forwarding that represents redirection of an arrived (from some input port) packet to the output port [13]. Modern packet switching devices implement full duplex mode of packet transmission between connected ports of neighboring devices having separate receiving and sending tracts, for instance through the frequency separation. Moreover, they possess facilities for flow control preventing overflow of a device with packets and mass drop of packets. Usually, such devices are classified into switches and routers based on the fact whether they use plain or structured (hierarchical) addresses, respectively. Traditionally switching devices maintain a table that assigns output port to an address (a set of addresses). For instance, there are Ethernet and MPLS switches and IP routers. Using regular lattice communication structures allows us to abandon maintaining address tables and to apply simple packet forwarding rules in a node [11] which can be implemented as microprogram code embedded into device.

Packet switching devices use two basic approaches [13] to the packet forwarding: store-and-forward (SF) when arrived packed is stored entirely into internal buffer of switch from where it is transmitted into the output port; cut-through (CT) when arrived packet header is stored into the port buffer and, after analyzing the header, the packet is directly forwarded to the output port. Usually, ports have buffers with capacity to store one packet. The internal buffer capacity is also limited.

When developing models of lattices, we model the port receiving and sending tracts separately. A pair of places represents the tract buffer: one place – to store a packet, represented by a token, and other place – to specify the buffer size. Such places are called complimentary because each time when we take a token from one of them, we put a token into another that keeps constant the sum of their markings.

Traditionally, we use compound names of vertices where the first letter "p" means place and "t" means transition; the

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

second letter “i’ means input, “o” means output, and “b” means buffer; the third letter “l” means limitation of buffer size. Using elementary Petri nets, we do not specify the packet header having an elementary token as a model of the entire packet. Besides, tokens in places with suffix “l” represent the buffer capacity measured in the number of packets. In figures obtained in Tina [14] modeling system, we use TEX-like notation to specify indices: “_” for lower indices and “^” for upper indices, also, to specify names with strokes, we duplicate the corresponding symbol, for instance *pp* stands for $\bar{p}'$.

Forwarding packets based on store-and-forward, cut-through, and combined principles is specified by Fig. 3, the upper index of the lattice node omitted. Using compound lower index means a set of elements specified by the valid range of the second component. Let us denote

$$K=\{x\mid 0\le x<k\}, D=\{y\mid 0\le y<d\}, H=\{-1,0,1\}.$$

Then we can specify indices set of the entire lattice and the Moore neighborhood as follows, respectively

$$L=K^d,\ M=H^d\setminus 0^d.$$

Then we write concisely

$$\bar{i}\in L,\ \bar{p},\bar{p}'\in M,\bar{p}'\ne\bar{p}.$$

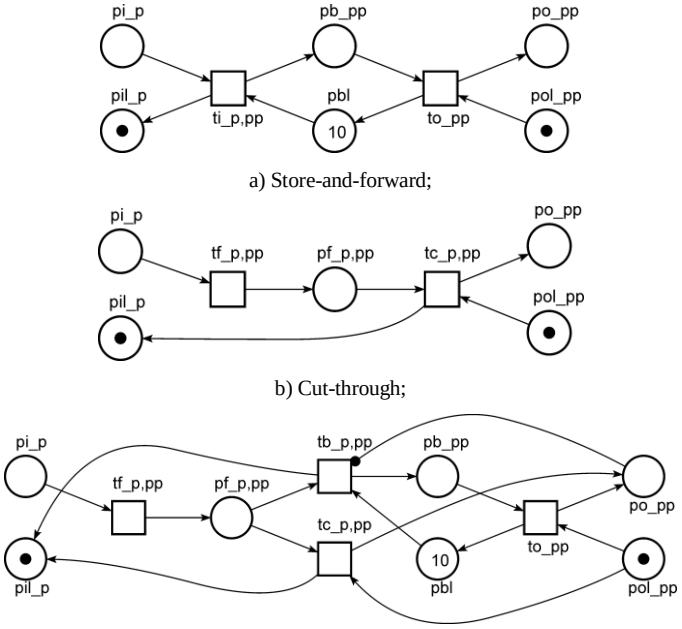


Figure 3. Basic models of packet forwarding

In Fig. 3a), unconditional alternative choice of a transition from the set specified by $ti_{\bar{p},\bar{p}'}$, within the range of index component $\bar{p}'$, implements nondeterministic packet forwarding decision. As a result, a packet is stored in the internal buffer section $pb_{\bar{p}'}$ that corresponds to the output port $\bar{p}'$. Then a packet is transmitted to the output port buffer (represented by a pair of complementary places $po_{\bar{p}'}$ and $pol_{\bar{p}'}$) by transition $to_{\bar{p}'}$ on the port availability. For flexible utilization of the buffer space, the buffer is represented as a set of places $pb_{\bar{p}}$ corresponding to all ports and the buffer size limit *pbl* which represent a complementary set preserving the total number of tokens. Complementarity of the mentioned sets of places is provided by the way of using them: each time we take a token from one, we put a token to the other and vice versa. Note that the switching store-and-forward device model represents state-of-art balance between simplicity and usefulness approved my manifold projects and publications [5],[6].

In Fig. 3b), there is no internal buffer. At first, an unconditional nondeterministic choice of the output port is implemented by the set of transitions $tf_{\bar{p},\bar{p}'}$, the choice is indicated by the corresponding place $pf_{\bar{p},\bar{p}'}$; here “f” stands for forwarding. Then a packet is transmitted directly to the output port buffer by transition $tc_{\bar{p},\bar{p}'}$ on the port availability; here “c” stands for “cut-through”. When for a definite given number of dimensions all the packet forwarding paths are taken together, we obtain complete specification of a packet forwarding device. For *2*-dimensional case, drawing the device model is possible and it is represented for CT device in Fig. 4; for dense Moore neighborhood it looks rather tangled with implicitly drawn sets of alternative forwarding vertices for all ports.

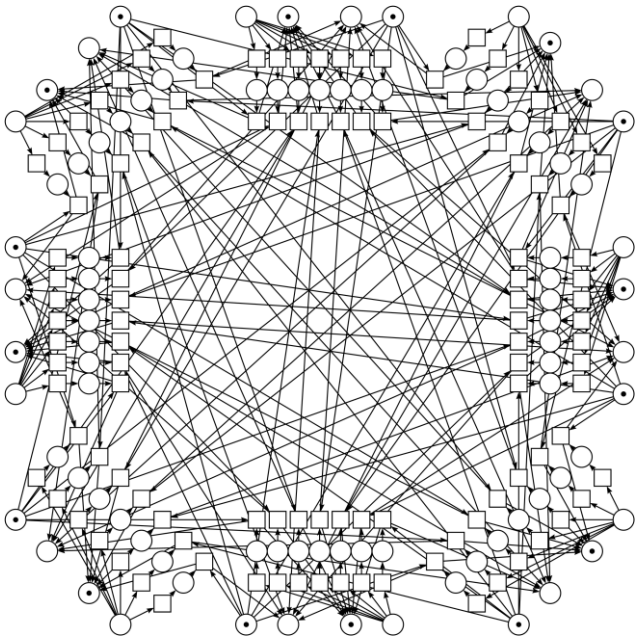


Figure 4. An example of cut-through node with Moore neighborhood, 2D.

A combined cut-through and store-and forward model is shown in Fig. 3c). It is the most realistic model for modern generation of switches. After nondeterministic forwarding decision, the choice between SF and CT modes is based on the output port availability represented by a pair of alternative

transitions $tb_{\bar{p},\bar{p}'}$ and $tc_{\bar{p},\bar{p}'}$. We use a read arc, with bold circle at the end of it, to check the token presence in place $po_{\bar{p}'}$ which can be interpreted here as an abbreviation of a pair of arcs having opposite directions.

Figures only illustrate while for formal specification of our models we use parametric expressions (PE) [5],[6] which represent a parameterized multiset rewriting system [15]. Each line specifies a transition by enumerating its input and output places separated by an arrow symbol ("$\rightarrow$"), then parameters specification follows, brackets group specifications. Let us consider CTSF device because it combines CT and SF techniques. Its formal specification is represented by PE (1).

$$\left(\begin{pmatrix} \begin{pmatrix} tf_{\bar{p},\bar{p}'} : pi_{\bar{p}} \rightarrow pf_{\bar{p},\bar{p}'} \\ tb_{\bar{p},\bar{p}'} : pf_{\bar{p},\bar{p}'}, pbl, po_{\bar{p}'} \rightarrow pb_{\bar{p}'}, pil_{\bar{p}}, po_{\bar{p}'} \\ tc_{\bar{p},\bar{p}'} : pf_{\bar{p},\bar{p}'}, pol_{\bar{p}'} \rightarrow po_{\bar{p}'}, pil_{\bar{p}} \end{pmatrix}, \bar{p}' \in M, \\ to_{\bar{p}} : pb_{\bar{p}'}, pol_{\bar{p}'} \rightarrow po_{\bar{p}'}, pbl \end{pmatrix}, \bar{p} \in M \right) \quad (1)$$

Note that PE (1) depends on a single parameter *d* that defines completely set *M*. Here to describe the output port, we use vector $\bar{p}$ which range enumerates all ports.

B. Composition of Lattice

In early works [5],[6], to minimize the model size, we composed lattice by merging the corresponding contact places of the neighboring cells. It leads to duplicate names of contact places, and sophisticates the description. In recent works [7],[12], we compose a lattice by connecting the output port with the input port of neighboring device by a dedicated transition that models the packet transmission channel as it is shown in Fig. 5. Suffix "l" here corresponds to "link". Transition $tl^{\bar{i}}_{\bar{p}'}$ indices are chosen with respect to the output port. Functions $rev(\bar{p})$ and $nei(\bar{i},\bar{p})$ stand for reversed port index and neighboring device index, respectively. They are defined as follows

$$rev(\bar{p}) = -\bar{p},$$

$$nei(\bar{i},\bar{p}) = \bar{i}' = \bar{i} + \bar{p},$$

where for torus, plus means addition modulo *k*.

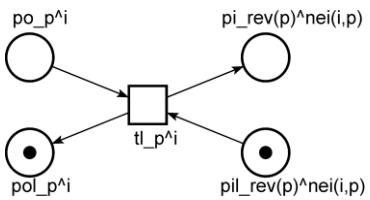
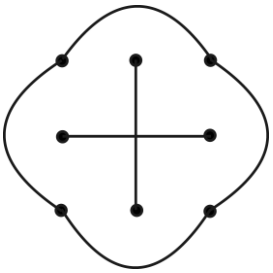


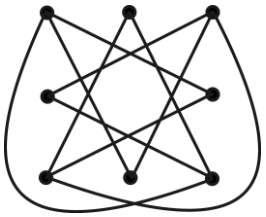
Figure 5. Scheme of nodes connection.

An example of obtained 3x3 lattice (*d=2*, *k=3*) is shown in Fig. 6 with the border connections omitted to not tangle the construct, the corresponding references to target ports specified.

In Fig. 7, the border connections are represented separately for straight (Fig. 7a), and slanting (Fig. 7b) connections to illustrate closing square into torus for Moore neighborhood.



a) straight connections (von Neumann neighborhood);



b) slanting connections (corner ports);

Figure 6. Am example of 3x3 torus with Moore neighborhood, concise connection scheme for border ports only.

Let us specify a lattice with Moore neighborhood and CTSF switching device by a PE. For this purpose, we situate models of communication device (1), supplied with the upper index of cell, into cells of lattice and connect them by transitions according to Fig. 5. As a result, we obtain PE (2).

$$\left(\begin{pmatrix} \begin{pmatrix} \begin{pmatrix} tf^{\bar{i}}_{\bar{p},\bar{p}'} : pi^{\bar{i}}_{\bar{p}} \rightarrow pf^{\bar{i}}_{\bar{p},\bar{p}'} \\ tb^{\bar{i}}_{\bar{p},\bar{p}'} : pf^{\bar{i}}_{\bar{p},\bar{p}'}, pbl^{\bar{i}}, po^{\bar{i}}_{\bar{p}'} \rightarrow pb^{\bar{i}}_{\bar{p}'}, pil^{\bar{i}}_{\bar{p}}, po^{\bar{i}}_{\bar{p}'} \\ tc^{\bar{i}}_{\bar{p},\bar{p}'} : pf^{\bar{i}}_{\bar{p},\bar{p}'}, pol^{\bar{i}}_{\bar{p}'} \rightarrow po^{\bar{i}}_{\bar{p}'}, pil^{\bar{i}}_{\bar{p}} \end{pmatrix}, \bar{p}' \in M, \\ to^{\bar{i}}_{\bar{p}} : pb^{\bar{i}}_{\bar{p}'}, pol^{\bar{i}}_{\bar{p}'} \rightarrow po^{\bar{i}}_{\bar{p}'}, pbl^{\bar{i}} \\ tl^{\bar{i}}_{\bar{p}} : po^{\bar{i}}_{\bar{p}}, pil^{\bar{i}+\bar{p}}_{-\bar{p}} \rightarrow pi^{\bar{i}+\bar{p}}_{-\bar{p}}, pol^{\bar{i}}_{\bar{p}} \end{pmatrix}, \bar{p} \in M \end{pmatrix}, \bar{i} \in K \right) \quad (2)$$

The initial marking of grid with one token in each of places $pil^{\bar{i}}_{\bar{p}}$ and $pol^{\bar{i}}_{\bar{p}}$, *a* packets in each section of the internal buffer, and available internal buffer capacity *b* is specified by PE (3).

$$\left(\begin{pmatrix} \left(a \cdot pb^{\bar{i}}_{\bar{p}}, pol^{\bar{i}}_{\bar{p}}, pil^{\bar{i}}_{\bar{p}}\right), \bar{p} \in M \\ b \cdot pbl^{\bar{i}} \end{pmatrix}, \bar{i} \in K \right) \quad (3)$$

The number of places, transitions, and arcs of the obtained model are calculated using technique [12] and specified by Table 2 depending on two given parameters *d* and *k*. For instance, for *d=3*, *k=4*, the model consists of 49984 places, 128128 transitions, 512512 arcs.

Note that, since we consider the entire infinite (countable) set of models specified by (2) for infinite (countable) range of natural parameters *d* and *k*, we call the formalism an infinite Petri net.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

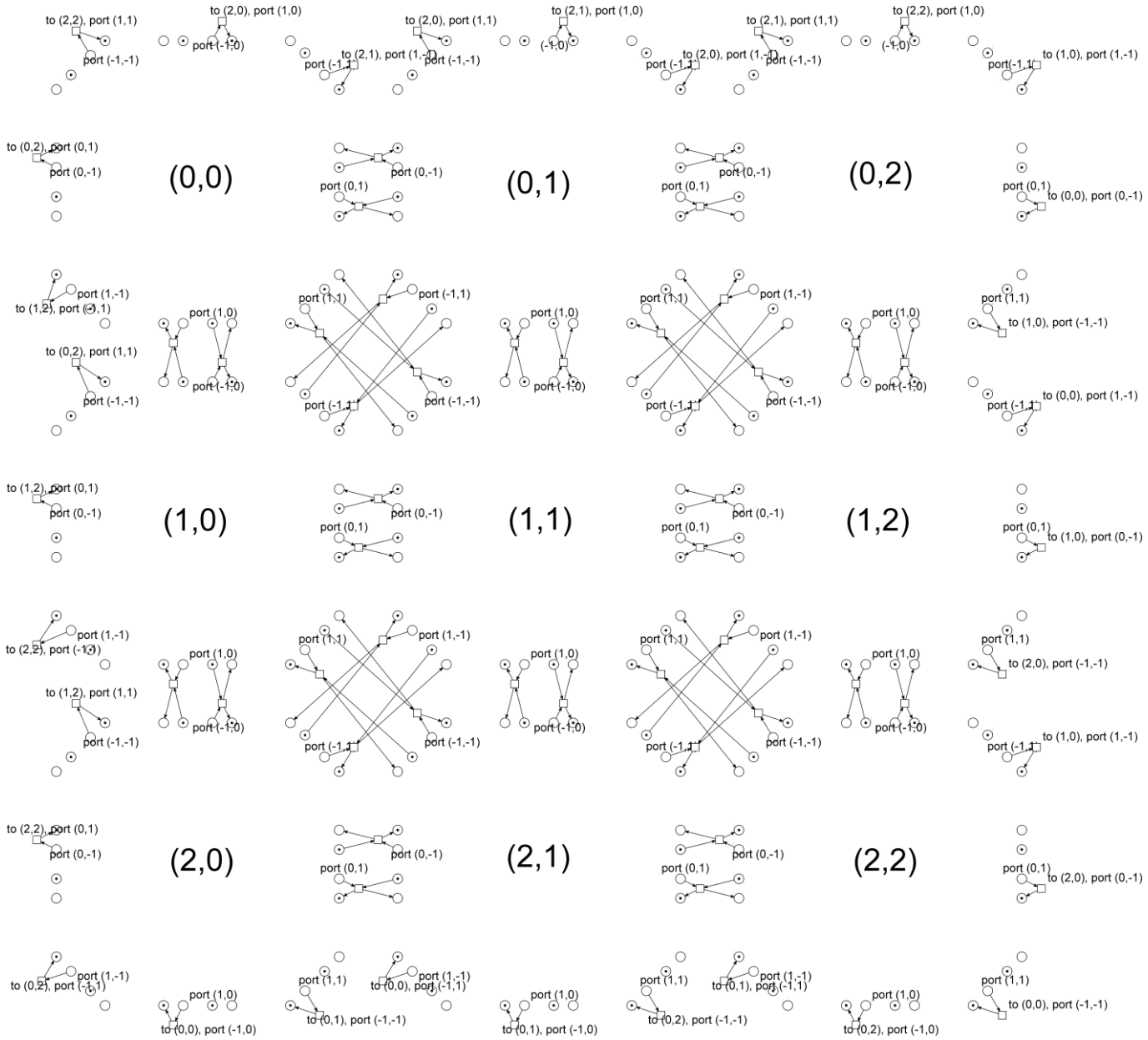


Figure 7. An example of 3x3 torus with Moore neighborhood connection scheme for internal ports with references for border ports.

TABLE II. FORMULAE FOR NUMBER OF VERTICES AND ARCS IN LATTICE

Element	Number
Place	$Npl = (3^d + (3^d + 2)(3^d - 1))k^d$
Transition	$Ntr = (3^d - 1)(3 \cdot 3^d - 4)k^d$
Arc	$Narc = (3^d - 1)(12 \cdot 3^d - 16)k^d$

V. PROVING PROPERTIES OF LATTICE

An ideal telecommunication (networking) system model should possess, after Berthelot and Diaz, such properties as conservativeness and liveness [16]. Remind that a conservative net is bounded [16]; conservativeness means that there is a natural vector of place weights such that the weighted sum of tokens is constant for any reachable marking. Here we concentrate on structural conservativeness of constructed models that is also called *p*-invariance (place invariance) [16],[17]. *t*-invariance (transition invariance), which is a necessary condition for liveness of bounded nets, is studied by similar methods using dual PE [6] composed with respect to places.

To prove *p*-invariance of lattice model (2), we apply methodology [3],[4] for composing and solving in parametric form infinite systems of linear Diophantine equations. An infinite system for computing *p*-invariants (3), unknowns are traditionally called *x* with suffix corresponding to the vertices notation, represents the balance of incoming and outgoing arcs of all transitions. The system is created directly on the model

specification (2) adding the unknown symbol "x" and replacing the arrow symbol by the equality symbol, variables moved to the left side. The sings can be reversed though we prefer to indicate with a minus sign that the place marking is decreased by the transition and to indicate with a plus sign that the place marking is increased.

$$\left(\left(\left(\begin{array}{l} -xpi^{\bar{i}}_{\bar{p}} + xpf^{\bar{i}}_{\bar{p},\bar{p}'} = 0 \\ -xpf^{\bar{i}}_{\bar{p},\bar{p}'} - xpbl^{\bar{i}} + xpb^{\bar{i}}_{\bar{p}'} + xpil^{\bar{i}}_{\bar{p}} = 0 \\ -xpf^{\bar{i}}_{\bar{p},\bar{p}'} - xpol^{\bar{i}}_{\bar{p}'} + xpo^{\bar{i}}_{\bar{p}'} + pil^{\bar{i}}_{\bar{p}} = 0 \end{array}\right), \begin{array}{l}\bar{p}' \in M, \\ \bar{p}' \neq \bar{p}\end{array}\right), \begin{array}{l} -xpb^{\bar{i}}_{\bar{p}'} - xpol^{\bar{i}}_{\bar{p}'} + xpo^{\bar{i}}_{\bar{p}'} + xpbl^{\bar{i}} = 0 \\ -xpo^{\bar{i}}_{\bar{p}} - xpil^{\bar{i}+\bar{p}}_{-\bar{p}} + xpi^{\bar{i}+\bar{p}}_{-\bar{p}} + xpol^{\bar{i}}_{\bar{p}} = 0 \end{array}, \bar{p} \in M\right), \bar{i} \in K\right) \quad (4)$$

We solve system (4) in parametric form using our heuristic technique [5],[6] and obtain its parametric solution (5), although further we prove in a formal way that (5) is a solution to (4).

$$\left(\begin{array}{c} ((pf^{\bar{i}}_{\bar{p},\bar{p}'}, \bar{p}' \in M), pi^{\bar{i}}_{\bar{p}}, pil^{\bar{i}}_{\bar{p}}), \bar{p} \in M, \bar{i} \in K \\ ((pb^{\bar{i}}_{\bar{p}}, \bar{p} \in M), pbl^{\bar{i}}), \bar{p} \in M, \bar{i} \in K \\ (po^{\bar{i}}_{\bar{p}}, pol^{\bar{i}}_{\bar{p}}),\ \bar{p} \in M, \bar{i} \in K \\ ((pbl^{\bar{i}}), ((pil^{\bar{i}}_{\bar{p}}, pol^{\bar{i}}_{\bar{p}}),\ \bar{p} \in M), \bar{i} \in K) \\ (((pf^{\bar{i}}_{\bar{p},\bar{p}'}, \bar{p}' \in M), pb^{\bar{i}}_{\bar{p}}, pi^{\bar{i}}_{\bar{p}}, po^{\bar{i}}_{\bar{p}}), \bar{p} \in M, \bar{i} \in K) \end{array}\right) \quad (5)$$

Expression (5), in essence, represents a sparse matrix specified in parametric form, only nonzero elements listed; all nonzero elements are equal to unit since there is no multiplier before the corresponding place name in accordance with the notation [12].

Using technique [5],[6] of direct substitution of each parametric solution to each parametric equation to obtain a correct equality, we prove Lemma 1 in a constructive way.

Lemma 1. Each row of (5) is a solution of (4).

Proof. Let us substitute row 3 of parametric solution (5)

$$(po^{\bar{i}}_{\bar{p}}, pol^{\bar{i}}_{\bar{p}}),\ \bar{p} \in M, \bar{i} \in K$$

into system (4). For equations 1 and 2, we at once obtain the correct equality

$$0 = 0$$

because variables $xpo^{\bar{i}}_{\bar{p}}$ and $xpol^{\bar{i}}_{\bar{p}}$ do not enter these equations and values of other variables are equal to zero in row 3 of (5). For equations 3, 4, and 5, we obtain

$$-xpo^{\bar{i}}_{\bar{p}} + xpol^{\bar{i}}_{\bar{p}} = 0,$$
$$0 = 0.$$

After removing absent (zero value) variables, the obtained equation can contain variables $xpo^{\bar{i}}_{\bar{p}}$ and $xpol^{\bar{i}}_{\bar{p}}$ with opposite signs; also it can use other index, say $\bar{p}'$ but the same for both variables.

In a similar way, we obtain the correct equalities for other 4 parametric solutions (rows) of (5). ■

The fact solution (5) lists each place of model (2) proves Theorem 1.

Theorem 1. Model (2) is a place invariant Petri net for any given number of dimensions *d* and lattice size *k*.

Proof. Let us compose a sum of rows 4 and 5 of sparse parametric matrix (5)

$$((pbl^{\bar{i}}), ((pil^{\bar{i}}_{\bar{p}}, pol^{\bar{i}}_{\bar{p}}),\ \bar{p} \in M), \bar{i} \in K) +$$
$$(((pf^{\bar{i}}_{\bar{p},\bar{p}'}, \bar{p}' \in M), pb^{\bar{i}}_{\bar{p}}, pi^{\bar{i}}_{\bar{p}}, po^{\bar{i}}_{\bar{p}}), \bar{p} \in M, \bar{i} \in K) =$$
$$((((pf^{\bar{i}}_{\bar{p},\bar{p}'}, \bar{p}' \in M), (pi^{\bar{i}}_{\bar{p}}, pil^{\bar{i}}_{\bar{p}}, pb^{\bar{i}}_{\bar{p}}, pol^{\bar{i}}_{\bar{p}}, po^{\bar{i}}_{\bar{p}}),\ \bar{p} \in M),$$
$$pbl^{\bar{i}}), \bar{i} \in K).$$

The obtained solution (*p*-invariant) lists each place of the lattice (2) and lists it only once (with omitted multiplier equal to 1). Thus, the obtained net is a *p*-invariant Petri net. ■

Based on the definitions of Petri net boundedness and conservativeness considered at the beginning of the present section, we formulate the following Corollary.

Corollary. Model (2) is a bounded and structurally strictly conservative Petri net for any given number of dimensions *d* and lattice size *k*.

Strict conservativeness means preserving the sum of tokens [16],[17]. Structural kind of conservativeness means that it holds for any given initial marking being characteristic to the Petri net structure (its graph).

From practical point of view, supposing hardware and software implementation of communication lattices according to its parametric specification, boundedness means that there will be no overflow of storage while conservativeness means that there is no invalid transformation of information when it appears from nowhere or disappears.

Since we obtained results which are valid for any lattice having the specified structure, we say that infinite Petri net (2) possesses the mentioned properties. In this context, we perceive an infinite Petri net as an abstraction represented by an infinite set of nets for an infinite countable range of parameters.

VI. Ad-hoc Software to Generate Models and Check their Properties

The composed constructs are rather abstract and sophisticated. That is why we employ double check of obtained results developing ad-hoc software to generate models and invariants for given parameters, and to compare them with invariants obtained for instances of the model by traditional technique provided by modeling system Tina [14].

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

A. Plot of Computational Experiment

In Fig. 8, a general scheme of computational experiments is presented, early developed generators of Petri net models are put on GitHub (https://github/dazeorgacm).

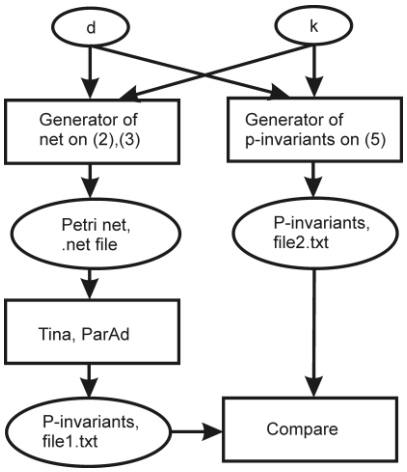


Figure 8. Scheme of computational experiments to check solutions of infinite system for *p*-invariants.

On the one hand, we generate a series of nets for inputted parameters and apply modeling system Tina [14] for computing *p*-invariants. On the other hand, we generate invariants as obtained parametric solutions (5) of infinite system for the same series of parameters values. Finally, for each set of parameters, we compare files of invariants, preliminary sorting each line and then sorting lines (in lexicographical order) since the order of place names can be different.

In series for *d* from 2 to 5 and *k* from 2 to 5, all *p*-invariants coincide that additionally acknowledges correctness of obtained parametric solutions. Invariants of big models, generated for upper values of parameters, have been computed by ParAd software [18] that takes advantages of the modern parallel architecture using distributed nodes with multicore processors.

Let us consider the general design of ad-hoc software.

B. Generator of Models

A generator of multidimensional torus communication lattice with a Moore neighborhood and a combined cut-through and store-and-forward switching node *(ht-gen-ctsf-moore.c*) is designed based on the lattice specification by parametric expressions (2) and (3). For each specified range of parameters, a separate loop is created, similar ranges from (2) and (3) are united for brevity.

Among basic routines, we mention routines which enumerate indexes of hypertorus nodes to organize a loop (while) on all $\bar{i}$: *start_i()* initializes index; *next_i()* calculates the next index and returns zero (false) if all the nodes have been enumerated (there is no next node). Listings of the mentioned routines follow:

```
void start_i(int *i, int d) {
  int j;
  for( j=0; j<d; j++ ) i[j] = 0;
}

int next_i(int *i, int d, int k) {
   int j=d-1, go=1, loop=1;
   while( go )
   {
     (i[j])++;
     if( i[j] >= k )
     {
       if( j == 0 ) { loop=0; go=0; }
       else
       {
         i[j]=0;
         j--;
       }
     }
     else go=0;
   } /* while go */
   return loop;
} // next_i
```

We consider *d*-vectors as numbers having *d* digits in the positional numbering system with radix *k*. To obtain the next element, we increment such number and propagate the possible carry.

For enumeration of port indices, we process *d*-vectors for *k=3* (range 0,1,2) because we have 3 options from set {-1,0,1}. To obtain the corresponding port index, we encode 1 to -1 and 2 to 1 preserving the order.

We simply print lines according to .net format of modeling system Tina [14] where a line, specifying a transition, starts from “tr” prefix and a line, specifying the place marking, starts from “pl” prefix. The rest of line corresponds to the format of parametric expression.

For printing parametric expression lines, we employ three routines which print a node name with indices *print_node_i()*, *print_node_p_i()*, and *print_node_p_pp_i()* which print names of the following formats $name^{\bar{i}}$, $name^{\bar{i}}_{\bar{p}}$, and $name^{\bar{i}}_{\bar{p},\bar{p}'}$, respectively. Listing of *print_node_p_i()* follows:

```
int print_node_p_i(char * name, int * i, int
*p, int d)
{
  int u;
  printf( "{" );
  printf( "%s", name );
  printf( "_" );
  for( u=0; u<d; u++) if(u==0) printf( "%d",
p[u] ); else printf( ",%d", p[u] );
  printf( "^" );
  for( u=0; u<d; u++) if(u==0) printf( "%d",
i[u] ); else printf( ",%d", i[u] );
  printf( "}");
} // print_node_p_i
```

According to Tina [14] .net file format, we put in curly brackets complex names of nodes. Condensed code of the main loop, with nested loops, follows:

```
start_i(i, d);
loop=1;
while( loop )
{
```

```
      loop_port=1;
      start_i(p1,d); p1[d-1]=1;
      while( loop_port )
      {
        vec_to_off(p1,p,d); // 0,1,2 => 0,-1,2
        /* output port */
        printf( "tr " );
        print_node_p_i("to",i,p,d); BLANK;
        print_node_p_i("pol",i,p,d); BLANK;
        print_node_p_i("pb",i,p,d);
        printf( " -> " );
        print_node_p_i("po",i,p,d); BLANK;
        print_node_i("pbl",i,d); NEW_LINE;
        ...
        loop_forward_port=1;
        start_i(pp1,d); pp1[d-1]=1;
        /* foward port */
        while( loop_forward_port )
        {
           vec_to_off(pp1,pp,d);
           if( !vect_eq(p,pp,d) ){
                ...
           }
           loop_forward_port=next_i(pp1, d, 3);
       } /* while (on forward ports pp) */
        ...
       loop_port=next_i(p1, d, 3);
      } /* while (on ports p) */
       ...
      loop=next_i(i,d,k);
} /* while loop */
```

Routine *vec_to_off()* encodes *d*-digital number with radix 3 as neighbors offset (port number) in Moore neighborhood and routine *vect_eq()* compares two vectors. Note that after initializing the port index with routine *start_i()*, we exclude the current node from its neighborhood by explicit assignment, for instance for port *p* by *p1[d-1]=1*. Inside the double loop, we print a line corresponding to transition $po_{\bar{p}}^{\bar{i}}$, printing other lines of the generated model omitted and replaced by ellipsis. Simple macros BLANK and NEW_LINE print space and carriage return symbols, respectively.

Program *ht-gen-ctsf-moore* uses the command line interface. An example of its launch to generate and store in file *ht-4536.net* 4 dimensional lattice of size 5 with 3 packets in each section of the internal buffer and remained available capacity of the internal buffer equal to 6 follows

```
> ht-gen-ctsf-moore 4 5 3 6 > ht-4536.net
```

An example of a line generated by the considered example of code for *2*-dimrnsional grid of size 2 for the parameter value $\bar{p} = (-1,1)$ and $\bar{i} = (2,1)$ follows

```
tr {to_-1,1^2,1} {pol_-1,1^2,1} {pb_-1,1^2,1}
-> {po_-1,1^2,1} {pbl^2,1}
```

Loaded into Tina (graphical editor *nd*) [14], the net can be automatically visualized, for instance visualization of 3x3 lattice is shown in Fig. 8.

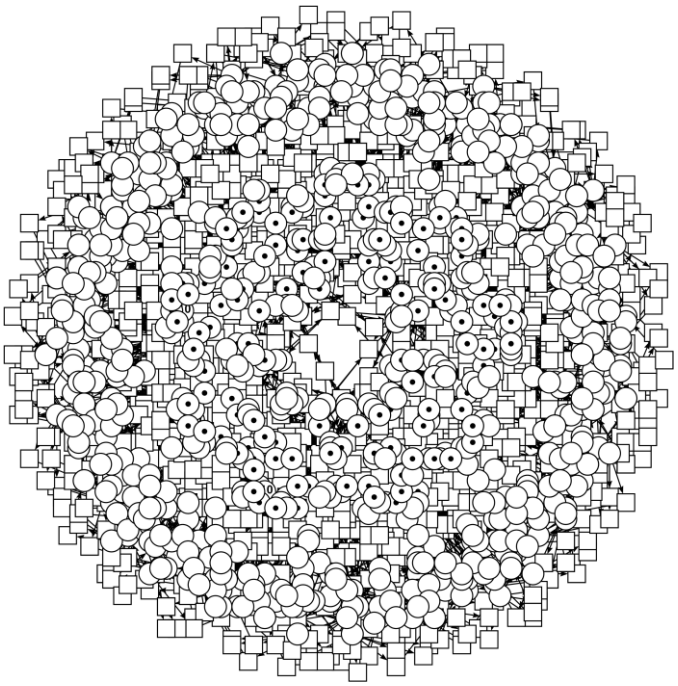


Figure 9. Automatical visualization of 3x3 lattice in Tina environment.

Density of connections for Moore neighborhood does not allow us to perceive the lattice structure, even for low dimension plane lattice, that justifies explanation of Moore neighborhood for 3x3 lattice with two separate figures (Fig. 6 and Fig. 7).

C. Generator of Invariants

A generator of *p*-invariants *ht-p-inv-gen-ctsf-moore.c* that prints invariants (5) of the lattice model (2) is based on the above described (in the previous subsection) principles of parametric expressions encoding.

Note that in PE (5) first 3 lines describe a set of invariants each, while each of the last 2 lines describes only one (rather long) invariant. The difference is specified by using brackets within parametric expression.

An example of a command line to compute p-invariants of net *ht-4536.net* and store them into file *ht-4536-p-inv.txt* follows

```
> p-inv-gen-ht-ctsf-moore ht-4536.net > ht-
4536-p-inv.txt
```

Let us overview the program structure. Since we consider each invariant as a (sparse) matrix (5) row, the program represents a sequence of nested loops, a loop per row. For instance, for the first row of (5)

$$((pf_{\bar{p},\bar{p}'}^{\bar{i}}, \bar{p}' \in M), pi_{\bar{p}}^{\bar{i}}, pil_{\bar{p}}^{\bar{i}}), \bar{p} \in M, i \in K,$$

the most inner loop on $\bar{p}'$ generates the line elements while the outer loops on $\bar{p}$ and $\bar{i}$ generate lines. Thus, for $\bar{p} = (-1,-1)$ and $\bar{i} = (0,0)$, we print the following line

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

```
{pf_(-1,-1),(-1,0)^0,0} {pf_(-1,-1),(-1,1)^0,0}
{pf_(-1,-1),(0,-1)^0,0}  {pf_(-1,-1),(0,1)^0,0}
{pf_(-1,-1),(1,-1)^0,0}  {pf_(-1,-1),(1,0)^0,0}
{pf_(-1,-1),(1,1)^0,0} {pi_-1,-1^0,0}
{pil_-1,-1^0,0}
```

Rows 2 and 3 of (5) are treated in the same way as row 1, while rows 4 and 5 generate one (very long) invariant each. For instance, for row 4 of sparse matrix (5)

$$((pbl^{\bar{i}}),((pil_{\bar{p}}^{\bar{i}}, pol_{\bar{p}}^{\bar{i}}),\ \bar{p} \in M),\bar{i} \in K),$$

we obtain a line that lists $3^2 \cdot (1+2 \cdot (3^2-1)) = 153$ following places

```
{pbl^0,0} {pbl^0,1} {pbl^0,2} {pbl^1,0}
{pbl^1,1} {pbl^1,2} {pbl^2,0} {pbl^2,1} {pbl^2,2}
{pil_-1,-1^0,0} {pil_-1,-1^0,1} {pil_-1,-1^0,2}
{pil_-1,-1^1,0} {pil_-1,-1^1,1} {pil_-1,-1^1,2}
{pil_-1,-1^2,0} {pil_-1,-1^2,1} {pil_-1,-1^2,2}
{pil_-1,0^0,0} {pil_-1,0^0,1} {pil_-1,0^0,2}
{pil_-1,0^1,0} {pil_-1,0^1,1} {pil_-1,0^1,2}
{pil_-1,0^2,0} {pil_-1,0^2,1} {pil_-1,0^2,2}
{pil_-1,1^0,0} {pil_-1,1^0,1} {pil_-1,1^0,2}
{pil_-1,1^1,0} {pil_-1,1^1,1} {pil_-1,1^1,2}
{pil_-1,1^2,0} {pil_-1,1^2,1} {pil_-1,1^2,2}
{pil_0,-1^0,0} {pil_0,-1^0,1} {pil_0,-1^0,2}
{pil_0,-1^1,0} {pil_0,-1^1,1} {pil_0,-1^1,2}
{pil_0,-1^2,0} {pil_0,-1^2,1} {pil_0,-1^2,2}
{pil_0,1^0,0} {pil_0,1^0,1} {pil_0,1^0,2}
{pil_0,1^1,0} {pil_0,1^1,1} {pil_0,1^1,2}
{pil_0,1^2,0} {pil_0,1^2,1} {pil_0,1^2,2}
{pil_1,-1^0,0} {pil_1,-1^0,1} {pil_1,-1^0,2}
{pil_1,-1^1,0} {pil_1,-1^1,1} {pil_1,-1^1,2}
{pil_1,-1^2,0} {pil_1,-1^2,1} {pil_1,-1^2,2}
{pil_1,0^0,0} {pil_1,0^0,1} {pil_1,0^0,2}
{pil_1,0^1,0} {pil_1,0^1,1} {pil_1,0^1,2}
{pil_1,0^2,0} {pil_1,0^2,1} {pil_1,0^2,2}
{pil_1,1^0,0} {pil_1,1^0,1} {pil_1,1^0,2}
{pil_1,1^1,0} {pil_1,1^1,1} {pil_1,1^1,2}
{pil_1,1^2,0} {pil_1,1^2,1} {pil_1,1^2,2}
{pol_-1,-1^0,0} {pol_-1,-1^0,1} {pol_-1,-1^0,2}
{pol_-1,-1^1,0} {pol_-1,-1^1,1} {pol_-1,-1^1,2}
{pol_-1,-1^2,0} {pol_-1,-1^2,1} {pol_-1,-1^2,2}
{pol_-1,0^0,0} {pol_-1,0^0,1} {pol_-1,0^0,2}
{pol_-1,0^1,0} {pol_-1,0^1,1} {pol_-1,0^1,2}
{pol_-1,0^2,0} {pol_-1,0^2,1} {pol_-1,0^2,2}
{pol_-1,1^0,0} {pol_-1,1^0,1} {pol_-1,1^0,2}
{pol_-1,1^1,0} {pol_-1,1^1,1} {pol_-1,1^1,2}
{pol_-1,1^2,0} {pol_-1,1^2,1} {pol_-1,1^2,2}
{pol_0,-1^0,0} {pol_0,-1^0,1} {pol_0,-1^0,2}
{pol_0,-1^1,0} {pol_0,-1^1,1} {pol_0,-1^1,2}
{pol_0,-1^2,0} {pol_0,-1^2,1} {pol_0,-1^2,2}
{pol_0,1^0,0} {pol_0,1^0,1} {pol_0,1^0,2}
{pol_0,1^1,0} {pol_0,1^1,1} {pol_0,1^1,2}
{pol_0,1^2,0} {pol_0,1^2,1} {pol_0,1^2,2}
{pol_1,-1^0,0} {pol_1,-1^0,1} {pol_1,-1^0,2}
{pol_1,-1^1,0} {pol_1,-1^1,1} {pol_1,-1^1,2}
{pol_1,-1^2,0} {pol_1,-1^2,1} {pol_1,-1^2,2}
{pol_1,0^0,0} {pol_1,0^0,1} {pol_1,0^0,2}
{pol_1,0^1,0} {pol_1,0^1,1} {pol_1,0^1,2}
{pol_1,0^2,0} {pol_1,0^2,1} {pol_1,0^2,2}
{pol_1,1^0,0} {pol_1,1^0,1} {pol_1,1^0,2}
{pol_1,1^1,0} {pol_1,1^1,1} {pol_1,1^1,2}
{pol_1,1^2,0} {pol_1,1^2,1} {pol_1,1^2,2}
```

To sort and compare files of invariants, obtained on the one hand by Tina or ParAd and on the other hand generated by program *ht-p-inv-gen-ctsf-moore*, we use script in Scheme language described in [12].

The coincidence of invariants in all computational experiments not only provides a double check of the obtained results, it also allows us to hypothesize that our ad-hoc technique [5],[6] for solving infinite Diophantine systems of linear algebraic equations in parametric form creates basis solutions.

VII. Conclusions

In the present paper, an infinite Petri net model of a hypertorus communication lattice with Moore neighborhood, that uses combined cut-through and store-and-forward switching node, has been developed and analyzed. For enumeration of device ports, the coordinate difference has been applied. Some aspects of routing within multidimensional torus with Moore neighborhood have been considered, simple packet switching rules proposed. To prove boundedness and conservativeness of the model, an infinite linear Diophantine system for finding *p*-invariants has been composed and solved in parametric form. From practical point of view, *p*-invariance means a balanced system using storage of limited capacity.

The obtained results can be also applied in systems biology [19] where cell-to-cell communication resembles functioning of packet switching network, especially for modeling human brain where manifold connections between neurons [20] could be structured using multidimensional lattices with dense Moore neighborhood. Since Petri nets simulate cellular automata, the developed techniques is applicable for modeling processes of spreading insects [21] and viruses [22], and for estimating efficiency of counter measures. Modeling controllable nuclear fusion [23] also looks a prospective application area because of using toroid constructs.

References

[1] P. Raj, S. Koteeswaran, Novel Practices and Trends in Grid and Cloud Computing, USA: IGI Global, 2019.

[2] J. Dongarra, Report on the fujitsu fugaku system. Tech Report ICL-UT-20-06, University of Tennessee, Knoxville, Oak Ridge National Laboratory, University of Manchester, USA (2020).

[3] Y. Ajima, S. Sumimoto, T. Shimizu, "Fujitsu Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers," Computer, vol. 42, no. 11, pp. 36-40, Nov. 2009.

[4] N.E. Jerger, T. Krishna, L.S. Peh, On-Chip Networks, Morgan & Claypool Publishers, 2017.

[5] D.A. Zaitsev, I.D. Zaitsev, and T.R. Shmeleva, "Infinite Petri Nets: Part 1, Modeling Square Grid Structures," Complex Systems, 26(2), 2017, 157-195. DOI: 10.25088/ComplexSystems.26.2.157

[6] D.A. Zaitsev, I.D. Zaitsev, and T.R. Shmeleva, "Infinite Petri Nets: Part 2, Modeling Triangular, Hexagonal, Hypercube and Hypertorus Structures," Complex Systems, 26(4), 2017, 341-371. DOI: 10.25088/ComplexSystems.26.2.341

[7] D.A. Zaitsev, T.R. Shmeleva, Retschitzegger R. "Spatial Specification of Grid Structures by Petri Nets, Micro-Electronics and Telecommunication Engineering," in Proceedings of 4th ICMETE 2020, Lecture Notes in Networks and Systems, vol. 179, 2021.

[8] J. Kari, Theory of cellular automata: A survey, TCS, vol. 336. no. 1-3, 3-33.

[9] T.R. Shmeleva, A.A. Kostikov, "Verification of Square Lattices with Dedicated Channels by Infinite Petri Nets," in 2020 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kharkiv, Ukraine, 6-9 Oct. 2020.

[10] T.R. Shmeleva, I.V. Stetsenko, "Modeling Unconditional Forwarding Decision within Switching Lattices," Springer, Current Trends in Communicationand Information Technologies, Lecture Notes in

Networks and Systems, vol.212, Chapter 10, 2021. DOI 10.1007/978-3-030-76343-5

[11] D.A. Zaitsev, S.I. Tymchenko, N.Z. Shtefan, "Switching vs Routing within Multidimensional Torus Interconnect," in 2020 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC&ST2020), October 6-9, 2020, Kharkiv, Ukraine.

[12] D.A. Zaitsev, T.R. Shmeleva, R.N. Guliak, "Analyzing Multidimensional Communication Lattice with Combined Cut-through and Store-and-forward Switching Node," Lect. Notes in Networks, Syst., Vol. 201, Raghvendra Kumar et al. (Eds): Next Generation of Internet of Things, 978-981-16-0665-6, 498624_1_En, (Chapter 58).

[13] Cut-Through and Store-and-Forward Ethernet Switching for Low-Latency Environments, Cisco White Paper, 2008.

[14] B. Berthomieu and F.Vernadat, "Time Petri Nets Analysis with TINA," in Proceedings of 3rd Int. Conf. on The Quantitative Evaluation of Systems (QEST 2006), 2006, IEEE Computer Society.

[15] I. Cervesato, E.S.L. Lam, "Modular Multiset Rewriting," in Davis M., Fehnker A., McIver A., Voronkov A. (eds) Logic for Programming, Artificial Intelligence, and Reasoning (LPAR 2015), Lecture Notes in Computer Science, vol 9450, 2015, Springer, Berlin, Heidelberg.

[16] Z. W. Li, and M. C. Zhou, Deadlock Resolution in Automated Manufacturing Systems, Springer, 2010.

[17] Diaz M. Petri Nets: Fundamental Models, Verification and Applications, John Wiley & Sons, 2013.

[18] D. Zaitsev, S. Tomov and J. Dongarra, "Solving Linear Diophantine Systems on Parallel Architectures," IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 5, pp. 1158-1169, 1 May 2019, doi: 10.1109/TPDS.2018.2873354.

[19] S. Janowski, B. Kormeier, T. Töpel et al "Modeling of Cell-to-Cell Communication Processes with Petri Nets Using the Example of Quorum Sensing," Stud Health Technol Inform. 2011;162:182-203. PMID: 21685572.

[20] J.F. Peters, S. Ramanna, Z. Suraj, M. Borkowski, "Rough Neurons: Petri Net Models and Applications," in: Pal S.K., Polkowski L., Skowron A. (eds) Rough-Neural Computing. Cognitive Technologies, 2004, Springer, Berlin, Heidelberg.

[21] G. Ortigoza, F. Brauer and A. Lorandi, "Mosquito-borne diseases simulated by cellular automata: A review," International Journal of Mosquito Research 2019; 6(6): 31-38.

[22] E. Burkhead, J. Hawkins, "A cellular automata model of Ebola virus dynamics," Physica A 438 (2015) 424–435.

[23] Richard D. Gill, Plasma Physics and Nuclear Fusion Research, Academic Press, 2013.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Improving resource allocation system for 5G networks

Astrakhantsev A.A.
Dept. of Infocommunication Networks
Igor Sikorsky Kyiv Polytechnic Institute
Kyiv, Ukraine
aastrakhantsev@its.kpi.ua

Globa, L.S.
Dept. of Infocommunication Networks
Igor Sikorsky Kyiv Polytechnic Institute
Kyiv, Ukraine
lgloba@its.kpi.ua

Novogrudska, R.L.
Intelligent network tools department
National Center "Junior Academy of Sciences of Ukraine"
Kyiv, Ukraine
rinan@ukr.net

Skulysh M.A.
Dept. of Infocommunication Networks
Igor Sikorsky Kyiv Polytechnic Institute
Kyiv, Ukraine
mskulysh@gmail.com

Stryzhak O.Ye.
National Center "Junior Academy of Sciences of Ukraine"
Kyiv, Ukraine
stryzhak@man.gov.ua

***Abstract*— 5G Networks must provide high-quality communication for various devices and users, but unfortunately today, when building such networks, energy efficiency and security factors are not taken into account, provided that the specified quality of service is ensured. In this paper proposes a comprehensive approach to servicing hybrid telecommunications services in 5G networks, which allows flexible management of information and communication system resources involved in servicing the load with a given level of QoS. The main advantage of the approach is to take into account the relationship between service quality and resource allocation processes, taking into account energy efficiency and productivity of computing processes. The results of the approach will allow to take into account the characteristic features of the next generation of communication networks, will provide a minimum delay and a given level of service quality when providing services to users.**

Keywords—next-generation networks; telecommunication services; architecture; resource; allocation system; security.

I. INTRODUCTION

5G networks should provide quality communication for a variety of devices and users. The flexibility and scalability of 5th generation smart grids will be ensured through the transition to software-managed networks, which widely use distributed cloud and fog computing, the efficiency of which significantly affects the quality of services to the end user. Now, there are no comprehensive technological solutions that would simultaneously take into account the effectiveness of both the communication and information component, as the only factor in providing quality services at the user level and on the scale of modern digital enterprises.

This goal of this research is creating an integrated approach for infocommunications management, which will coordinate quality of communication, use distributed cloud computing and provide high level of security.

The structure of the paper is the following: Section 2 describes research chalanges. Section 3 shows characteristic features of next-generation networks arcitecture. In section 4 architecture of the proposed intelligent resource allocation system is depicted. Section 5 gives the description of intelligent management tools that are used in proposed system. Basic notions on information protection in proposed system are given in Section 6. Section 7 presents conclusions and plans for future work.

II. RESEARCH CHALLENGES

The intelligent control system for the service of hybrid telecommunications services in 5G networks solves the problem of organizing management processes at a fundamentally new level according to the capabilities published in 5G RPP architecture [1], which will prepare for the implementation of Smart Connectivity as a platform for next generation Internet, with usage of flexible connection infrastructure, while facilitating the management of processing and storage of user data, providing quality service and reducing energy costs.

This requires a comprehensive approach to providing a quality of service process and intelligent management tools that take into account the assessment of requirements and needs for efficient, seamless and secure interoperability with computing resources (e.g. distributed data centers, edge computing) and a set of innovative devices.

Also relevant is the task of developing comprehensive mechanisms that would take into account the relationship between the quality of end-user service and the processes of allocation of computing resources between virtual entities, taking into account energy efficiency and productivity of computing processes [2]. Currently, virtualization of network functions (NFV) still needs to address issues related to the implementation of virtualized network elements [3], and issues related to the scalability of the network to a large number of IoT devices with limited resources [4].

The efficiency and quality of management of the process of servicing hybrid telecommunications services in 5G networks is significantly influenced by the distribution of tasks in the computer nodes of the 5G information and communication network, which is characterized by

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

significant energy consumption [5]. This phenomenon encourages mobile operators and Internet service providers to build complex network architectures with the inclusion of new features and extensions that are more difficult to manage and that function inefficiently [6]. The variability of the load, which is recorded in modern information and communication systems, affects the placement of virtual machines, which occurs constantly and in real time, so there is a need to optimize their placement [7]. Intelligent tools will integrate elements of the information and communication environment, including physical networks, SDN networks, cloud and fog computing nodes, repositories of information resources and services, adhering to a single criterion of quality of service of hybrid telecommunication services in 5G networks.

III. NGN ARCHITECTURE

Modern communication systems require the organization of processing large arrays of information in a distributed heterogeneous environment. Most work on load processing in data centers does not take into account that the amount of load that arrives is pulsating and often changes significantly during the day.

Due to the variability of the load experienced by modern systems, the placement of virtual machines in a heterogeneous information and communication environment must be constantly optimized in real time. Given the difficulty in predicting peak loads, the system must use a combination of dynamic resource allocation and request management to respond in a timely manner to load changes. Dynamic resource allocation allows you to allocate additional resources for application software services, such as servers, to cope with the increase in workload, while managing the process of providing them allows the load management node to temporarily reject redundant requests while additional resources are allocated. In modern communication systems, the problem of quality of service is modified.

If earlier it was a question of a limited network resource, limited possibilities of a network of access, and also resources of switching nodes, now the system has conditionally unlimited technical resources. New technologies for software management of software defined radio (SDR), software defined network (SDN), as well as technologies for network functions virtualization (NFV) reduce the task of organizing information and communication environment to three main tasks:

- tasks of placement of transceivers

- the task of organizing a distributed software package, which reproduces the structure and logic of the information and communication network,

- tasks of intelligent control of a complex system.

The software package that provides the information and communication network is a complex distributed system based on a heterogeneous cloud environment, the computing resources of which are relatively infinite.

In fig. 1 shows the features of the organization of 6G networks presented in [8]. To provide all subsystems of the communication system, it is necessary to provide intelligent management of the infrastructure, an illustrative image of which can be seen in fig. 2.

The key aspects of next generation networks (Fig. 1) include:

Intelligence makes all configurations of services and network connections, which are mainly manual today, automated. This reduces time to market for new services and improves the quality with less risk of error.

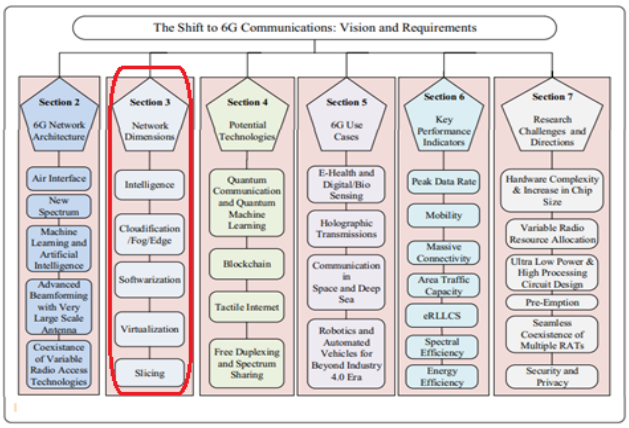


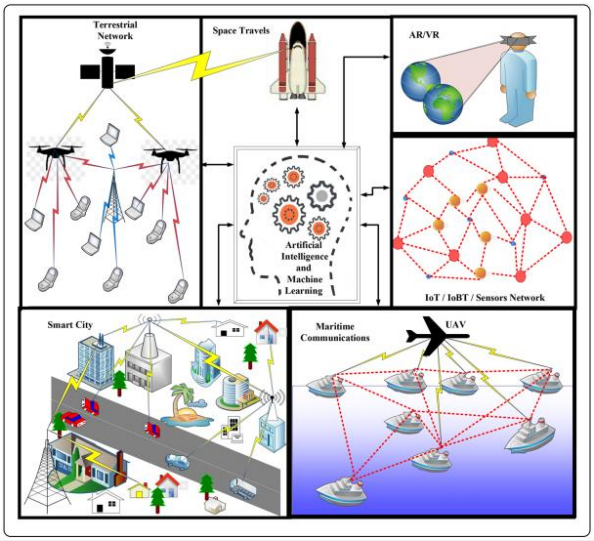
Fig. 1. Features of construction of 6G systems

Cloudification is the conversion and/or migration of data and application programs to make better use of cloud computing.

Network softwarization is an overall transformation trend about designing, implementing, deploying, managing, and maintaining network equipment and/or network components through software programming.

Virtualization enables the on-demand instantiation of functions in a format easier to load-balance, scale up/down, and allow for the movement of functions dynamically across distributed hardware resources in the network. The approach to the effective usage of virtual resources based on a dynamic approach to their location is presented on paper [9].

Slicing enables mobile network operators to provide dedicated virtual networks with functionality specific to the service or customer over a common network. The approach to slicing realization based on optimization of the efficiency of mobile networks by forming and mapping slices of a multi-service communication network is presented on paper [10].



978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Fig. 2. A depiction of space-air-ground-sea based integrated next-generation communication system with a wide range of applications

The operation of the system with intelligent control is provided by the network of the communication operator, the core of which is a geographically distributed network of data centers. Each of them is connected to communication channels that deliver the primary information of subscribers, which requires conversion at the lowest level.

The services provided to mobile users are heterogeneous and require constant monitoring of service quality indicators, as different types of services are differently sensitive to delays and losses in the process of information flow. That is why to organize a flexible information and communication system, the operator must implement solutions using software in different parts of the service system. To ensure control of service quality indicators at each stage it is necessary to control and manage not only the service process, but also the organization of access network resources, which includes radio resource, physical / virtual computing resources, and network core resource, which includes physical and virtual resources of data centers and telecommunications networks that are used to organize the work of distributed data centers.

Therefore, it is necessary to implement and improve service resource allocation systems in all areas of the next generation network.

IV. ARCHITECTURE OF THE PROPOSED INTELLIGENT RESOURCE ALLOCATION SYSTEM

The 5G networks currently being developed by the scientific community envisage the transformation of the communication system, when the functions of the network subsystems will be partially performed as application software components.

Given all the features of the next generation of communication networks, as well as the flexible opportunities offered by the virtualization of network functions and the level of development of cloud computing technology, this study proposes a 5G network architecture that summarizes the main trends of next generation networks. The proposed architecture of the next generation mobile network is shown in fig. 3, the feature of which is the ability to apply intelligent controls for all units and subsystems of the network.

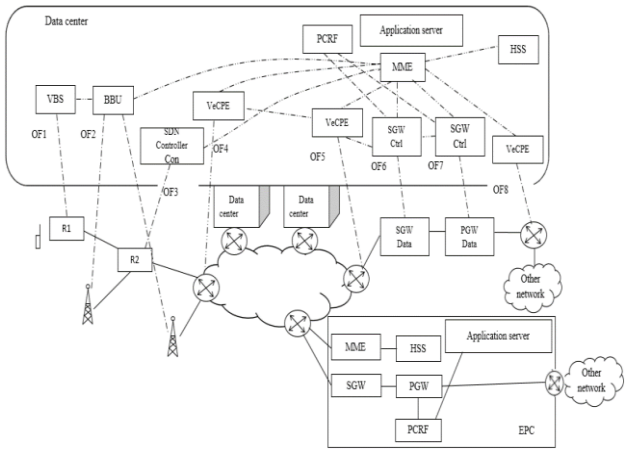


Fig. 3. Provider core network structure using software-controlled routers

Fig. 3 shows a scenario where a mobile subscriber communicates with the repeater R1, which converts the radio signal into an optical one, then the signal reaches the repeater R2 controlled by the SDN controller, which is also located in the data center. Once in the data center, the signal is processed by the virtual base station. Next, according to LTE technology, the flow is sent to the core of the operator for further processing.

The BBU subsystem (modulating signal generation unit) is based on the technology of software-configured networks and virtualized network functions. It is a system that not only supports virtual base stations, but can also be used for 2G / 3G / 4G / Pre5G hybrid solutions.

Maintenance in the kernel determines the further direction of the data flow. If the stream is directed to the internal network of the operator, then immediately in the data center it is sent for service to the appropriate virtual base station, and then sent to the subscriber through the repeaters R2 and R1. If its purpose lies outside the operator's local network, then the stream is routed to a virtual router of the local border, which is located immediately in the data center, after maintenance in which the stream goes to external networks.

According to the specification ETSI GS NFV 001 v.1.1.1 (01/2015) [11] all computing functions that accompany the transmission process will be performed in data centers with cloud infrastructure, virtualization is provided, including base stations, which will reduce the energy used associated with the dynamic allocation of resources and load balancing. In addition to base stations for cloud-based access networks (Cloud-RANs), it is planned to create a base frequency processing resource (BBU) to combine various base stations, both virtualized and non-virtualized, into a single virtualized environment. Next, the virtualization of network functions according to the specification is proposed for a router located on the boundary of the local network of the operator. The router performs the functions of flow classification, routing management and providing network barrier protection (Firewall).

Thus, the proposed solution for the provider's network core structure can be used as a basic architecture for the deployment of new intelligent service management systems in next-generation networks. This provides opportunities for the introduction of artificial intelligence in communications management systems. The data center element combines a group of data centers connected through a secure network into a single logical service space. Ensuring quality service to end users significantly depends on the organization of processes in such a heterogeneous data center, built on the concept of cloud computing.

V. FEATURES AND CHARACTERISTICS OF THE PROPOSED INTELLIGENT RESOURCE ALLOCATION SYSTEM

The intelligent system must take into account all the features of different hybrid telecommunications services, configure the optimal infrastructure to ensure a minimum delay for each service. That is why the development of

intelligent resource allocation system (IRAS) involves the integration of intelligent management tools for maintenance of hybrid telecommunications services of various natures in the information and communication environment of the 5th generation.

Such tools take into account the assessment of requirements and needs for efficient, seamless and secure interoperability with computing resources (eg distributed data centers, edge computing) and a range of devices. Intelligent management tools will integrate elements of the information and communication environment, including physical networks, SDN networks, cloud and fog computing nodes, storage of information resources and services, adhering to a single criterion of service quality of hybrid telecommunications services in 5G networks. The use of intelligent management tools in the design process of IRAS will integrate modern technologies and architectures, provide a minimum delay and a given level of service quality when providing services to end users.

The introduction of intelligent management tools in the structure of IRAS implements the tasks:

- semantic linking of network information resources,

- analytical support of search and analysis of large amounts of distributed information,

- integration of customer service models,

- energy efficient allocation of resources,

- control of input load for use of resources of service subsystems,

- planning and forecasting user needs in future periods.

In the research, intelligent management tools are software and hardware developed using modern models and methods of artificial intelligence. Nowadays specialized technologies based on artificial intelligence are used to develop information systems of various kinds and purposes. Artificial intelligence technologies also offer a wide range of specialized models, methods and technological means to store, operate, analyze and optimize the great amound of information [12]. Model and methods of artificial intelligence, that are the basis of IRAS, are presented in fig. 4.

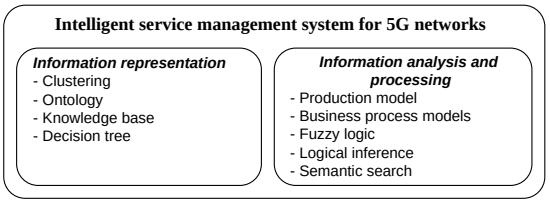


Fig. 4. Intelligent models and methods for IRAS

Each element of the set of intelligent models and methods shown in fig. 4 performs a certain function in IRAS, namely:

1. Clustering is used in the method of energy efficient allocation of resources, which increases the energy efficiency of input load processing by distributed computing systems while ensuring a high level of computing performance and compliance with service quality requirements [13].

2. Ontology integrates all available components of information and communication system and intelligent components of process control of hybrid telecommunication services and allows to dynamically form service scenarios in 5G networks [14].

3. The knowledge base acts as a central repository and stores information operated by IRAS, expert rules, meta-descriptions of services and business processes.

4. Decision trees implement mechanisms for intelligent planning and forecasting of user needs in subsequent periods of time.

5. Rules of production models allow to implement mechanisms of support of decision-making (based on expert rules) for definition of volumes of input loading for use of resources of subsystems of service that will allow to provide service with the set indicators of QoE.

6. Business process models implement the formation of a sequence of services in the dynamic formation of scenarios of customer service in 5G networks, as well as improve the quality of service through automated calculation of quality assessment of their provision.

7. Methods of fuzzy logic are the basis of the algorithm for estimating the current status of providing services by telecom operator. Such algorithm is based on the integral quality index of providing services. Object is achieved by methods of fuzzy logic for collaborative accounting of the impact of clear and fuzzy parameters [15].

8. The mechanisms of logical inference allow us to draw conclusions about the information space of the ontological model, which ensures the implementation of the connectivity of all elements of IRAS, information and services.

9. Semantic search provides a search for large amounts of distributed information based on ontology, knowledge base and decision trees.

An experimental analysis of energy-efficient approaches to distributed workload processing in communications networks was focused on the PCPB-2 (Power Consumption and Performance Balance) algorithm proposed by the authors [23]. Proposed approach allows to obtain an energy efficiency gain of up to 8.9% compared to the widely used Backfill energy-efficient scheduling algorithm.

Among the means of artificial intelligence there is a group of means of knowledge representation [16], one of which is ontological modeling. The ontological approach is a powerful intellectual tool for designing and modeling systems that operate in different subject areas [17, 18]. It is ontological models that allow to structure and systematize disparate information in order to further analyze, process and search [19].

The application of the ontological approach in the design of IRAS allows to increase the efficiency of integration of all elements of the information and communication system in the process of managing the service of hybrid telecommunication services in 5G networks. The use of the ontological model will allow structuring and systematizing the data and services of IRAS in order to control and analyze the operation of the telecommunications network, service maintenance

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

processes and resource use. The ontological model as a basic intelligent component of IRAS allows to dynamically form scenarios of service maintenance in 5G networks. Due to the application of the ontological approach to the formation of work processes, the evaluation of the quality of service provision by the telecommunications provider will save labor, time and financial costs for the implementation of such processes.

Thus, the use of intelligent management tools will allow to develop an intelligent system of control and management of information and communication resources and flows, which due to the original variable depending on the load architecture is able to flexibly and efficiently handle large amounts of information flows. The introduction of intelligent management tools in the structure of IRAS will allow to systematize the protocols of interaction of elements of the hybrid information and communication network and maintenance procedures of hybrid telecommunication services. This will give the following advantages: service scenarios of hybrid telecommunication services are formed in accordance with the current state of the telecommunication and computer system; loading of information and communication system resources is carried out in accordance with the policies of using available resources and dynamic attraction of additional resources. The peculiarity of this approach is the ability to automatically select the best structure of the telecommunications network and computing resources that fully meet the needs of modern information and communication system serving digital production, IoT systems, autonomous control systems for mobile objects, integrated automated emergency response systems. human health monitoring systems, etc.

VI. ENSURING THE PROTECTION OF INFORMATION IN THE PROPOSED SYSTEM

A. New security requirements and scalability

With the digital transformation of industries, we are entering the cyber-physical domain through robots, sensors and autonomous cyber-physical processes, which will further introduce new dimensions of attack vectors and vulnerability through these connected digital systems [20]. Novel types of attack, as well as new privacy and cybersecurity regulations, may take many industries by surprise. The expectations of the public and governments regarding security and privacy are very high. At the same time, information security is a top concern among enterprises on a digital transformation journey. The IoT, therefore, needs to be secure from the start, protecting personal information, company secrets, and critical infrastructure. Regulators need to walk a fine line between protecting privacy, safeguarding national security, stimulating economic growth, and benefiting society at large. To succeed with the transformation that 5G brings about, industries need to gather competence and understand new threats and how to mitigate them.

When discussing security, confidentiality protection – often forwarded by need of end-to-end encryption specifically – is the first security dimension that comes up. While confidentiality certainly is important, it is just one of the many dimensions needed to ensure a trustworthy system. To build truly trustworthy systems it is important to take a holistic view and not only focus on individual parts in isolation. For example, interactions between user authentication, traffic encryption, mobility, overload situations, and network resilience aspects as well as secure supply chains need to be considered together. It is also important to understand relevant risks and how to appropriately deal with them (fig. 5).

An integrated approach with end-to-end capabilities covering device chipsets, security properties in the networks, security management of data, software and timely patching as well as threat intelligence is necessary.

In this approach, standards and certifications have clear roles. Given the scale of billions of connected devices and processes to be handled by mobile networks, the role of advanced analytics alongside data integrity protection and timely software patching is also critical. Addressing the challenges on a global scale by thinking ahead and creating built in security capabilities becomes a necessity.

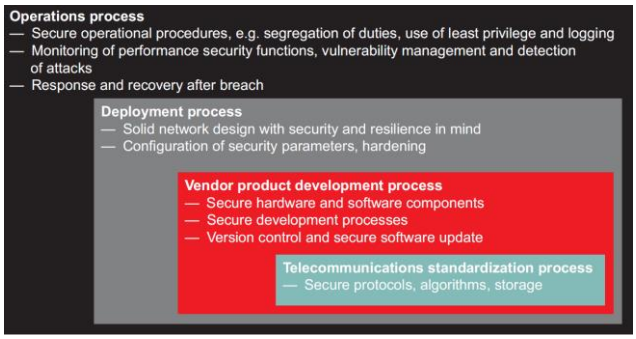


Fig. 5. A holistic approach to security

B. Service provider responses to new security challenges

The GSM (2G) standard, used as a base also for the 3G and 4G standards, anticipated the security issues from its beginning and introduced robust Confidentiality and Authentication mechanisms that still last at a scale of 7 billion units/users. Similarly, the new 5G standard defines capabilities to support the expansion beyond Mobile Broadband with mission critical, enterprise applications and society critical applications etc. in a built-in fashion [21, 22].

The world is going from having to secure primarily business data, to having to secure all things connected, whole business processes and everything in between.

With the inherent security of the 4G/5G networks we are creating an overall system security that can scale to many new use cases, connections and devices. With a solid underpinning of standardized 3GPP security including privacy & data protection, we have added analytics, intelligence and capabilities to orchestrate security from device to cloud.

Based on 5G technologies and standards, service providers will be well suited to address some of the emerging security challenges [20]:

- Ensure that no one can move devices in a factory without permission – Security manager monitors security threats and finds anomalies through machine learning
- Block remote access to devices to prevent unauthorized actuator commands – a device and data management system control that only authorized users can actuate commands

- Make sure that no unwanted software is installed on devices to modify behavior – software signing ensures only trusted software can be installed on device and machines

- Prevent outside eavesdropping – 3GPP symmetric communication encryption makes eavesdropping very hard

- Keep unwanted devices from connecting to factory floor network – Hardware Credentials makes it impossible for a rogue device to connect to a factory network

C. 5G standards drive network security

When the mobile network evolves into 5G, the network security and trustworthiness will automatically evolve. Ericsson believes in an industry wide approach to security and has actively contributed to the 3GPP security work and other standardization bodies. The main benefits from this work have now crystallized into five key elements to support the evolving use cases for 5G [20] (fig. 6):

Fig. 6. Five properties that contribute to the trustworthiness of the 5G system

1. Resilience

Enhanced security in the architecture for 5G. E.g. Base station function has been separated into two functions, a central and a distributed unit, where security sensitive functionality like user plane encryption terminates in the central unit. Network slicing can isolate groups of functions and users. E.g. public safety can have its own slice of the network. Furthermore, an enterprise can select to have less secure IoT devices in one slice, that would not affect any other slice with more sensitive data traffic. Efficient usage of cloud technology makes it possible to handle dynamic resilience as well as to create small, isolated network partition, i.e. a network slice, thus only smaller units of the network will be hit by possible attack vectors.

2. Communication security

The strong and well-proven security algorithms from the 4G system are reused. The 5G system includes protection against eavesdropping and modification attacks. Signaling traffic is encrypted and integrity protected. User plane traffic is encrypted and can be integrity protected. Mobility in the 5G system also inherits security features from the 4G system, such as separation of keys for specific purposes, backward and forward security for keys at handovers and idle mode mobility, and secure algorithm negotiation.

User plane integrity protection is a new feature that is valuable for small data transmissions, particularly for constrained IoT devices. Another related new feature is user plane security policy. There are many other new security features, like automatic recovery from malicious security algorithm mismatches, security key separation between core network functions, fast synchronization of security contexts in access and core networks, secure steering of roaming, etc.

3. Identity Management

At its heart, the 5G system has secure identity management for identifying and authenticating subscribers, roaming or not, ensuring that only the genuine subscribers can access network services. It builds on strong cryptographic primitives and security characteristics that already exist in the 4G system. Examples of primitives include strong cryptographic algorithm sets and key generation functions, and mutual authentication between device and network.

One of the most valuable new security features in the 5G system is the new authentication framework that allows other types of credentials, that do not require physical SIM cards, such as certificates, pre-shared keys and token cards. This feature enables mobile operators to flexibly choose authentication credentials, identifier formats and authentication methods for subscribers and IoT devices. Another valuable new security feature is the ability of a subscriber's operator to determine the presence of the subscriber during an authentication procedure – even when roaming. This feature enables the subscriber's operator to mitigate potential fraud and prevent security and privacy attacks against the subscriber or operator.

4. Privacy

Addressing the issue of privacy has been a high priority in the 5G system from the beginning so that subscribers' privacy is included by design. The devices and the network mutually authenticate each other and use integrity-protected signaling. This setup makes it unfeasible for an unauthorized party to decrypt and read the information that is communicated over the air. The most important new privacy feature in the 5G system is enabling a home operator to conceal a subscriber's long-term identifier, roaming or not, while simultaneously complying with regulatory duties. When enabled, this feature makes active attacks and the infamous IMSI catchers ineffective in a 5G-only system. Another new privacy feature is that the 5G system enforces a stricter policy for update of temporary identifiers. This guarantees that temporary identifiers are refreshed regularly, which makes passive attacks impractical.

5. Security assurance

In 3GPP, security assurance is a means to ensure that network equipment meets security requirements and is implemented following secure development and product lifecycle processes. This assurance is especially important for mobile systems, as they form the backbone of the connected society and are even classified as critical infrastructure in some jurisdictions. The NESAS (network equipment security assurance scheme) is an initiative from 3GPP and GSMA to create a security assurance scheme suitable to the telecom equipment lifecycle. NESAS aims to meet the needs of many national and international cybersecurity regulations, such as the EU cybersecurity certification framework.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The increase in vulnerability of 5G networks is caused by the presence and simultaneous interaction of both physical components of the network and virtualized (software) elements. To increase the security of next-generation networks, 5 components should be used: resilience, communication security, identity management, which will prevent major threats and ensure the security of the network and data.

VII. CONCLUSIONS AND FUTURE WORK

The presented research proposes a comprehensive approach to the maintenance of hybrid telecommunications services in 5G networks, which allows flexible management of information and communication system resources involved in servicing the load with a given level of QoS entering the network and its scalability for timely response to load. The 5G network architecture is proposed, the feature of which is the ability to use intelligent controls for all blocks and subsystems of the network, which can be recommended for usage as a basic architecture for the deployment of new intelligent control systems for service processes for the future generations networks. The architecture includes a group of intelligent controls, the use of which in the design process SeMaS will take into account the characteristics of the next generation of communication networks, provide a minimum delay and a given level of service quality when providing services to end users.

The main threats and vulnerabilities inherent in next-generation networks (their physical and virtual components), as well as the information transmitted by them, have been identified. An approach is proposed, which includes 5 components, including resilience, communication security, identity management, which allows to cover the main threats and ensure the security of the network and the data transmitted and processed in it. Future research will focus on the implementation (implementation of software and hardware) and testing the proposed approach using intelligent network management tools in test mode on the training samples of the data of the operator.

REFERENCES

[1] View on 5G Architecture [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2019/07/5G-PPP-5G-Architecture-White-Paper_v3.0_PublicConsultation.pdf

[2] A.P Singh, S. Nigam, and N.K Gupta, “A study of next generation wireless network 6G,” Int.J. Innovative Res. Computer Commun. Eng., vol. 4, no. 1, pp. 871–874, 2007.

[3] A. Al-Dulaimi, X. Wang, and I. Chih-Lin, “5G Networks: fundamental requirements, enabling technologies, and operations management,” Wiley, New Jersey, 2018.

[4] A. Mourad, R. Yang, P.H. Lehne, and A. De La Oliva, “A baseline roadmap for advanced wireless research beyond 5G,” Electronics, vol. 9, no. 2, pp. 351, 2020.

[5] A. Pouttu, “6Genesis-taking the first steps towards 6G,” in: Proc. IEEE Conf. Standards Communications and Networking, 2018

[6] H. Viswanathan, “Mogensen Communications in the 6G era,” IEEE, Access 8:57063–57074, 2020

[7] J. Gozalvez, “Tentative 3GPP timeline for 5G [mobile radio],” IEEE Vehicular Technol Magazine, vol. 10, no. 3, pp. 12–18, 2015.

[8] M.W. Akhtar, S.A. Hassan, R. Ghaffar, H. Jung, S. Garg, and M.S. Hossain, “The shift to 6G communications: vision and requirements,” Human-centric Computing and Information Sciences, vol. 10, no 1, pp. 1-27, 2020.

[9] L. Globa, M. Skulysh, and E. Siemens, “Conditionally Infinite Telecommunication Resource for Subscribers,” In M. Ilchenko et al., Advances in Information and Communication Technology and Systems. MCT 2019, LNNS 152, 2021, Springer, pp. 206–216, https://doi.org/10.1007/978-3-030-58359-0_11.

[10] L. Globa, S. Sulima, M. Skulysh, and A. Zhuravel, “An approach for virtualized network slices planning in multiservice communication environment,” Information and Telecommunication Sciences, vol. 1, pp. 37–44, 2019.

[11] ETSI: Network Functions Virtualisation (NFV); Infrastructure Overview. (ETSI GS NFV-INF 001 V1.1.1 (2015-01)). [Online]. Available: https://www.etsi.org/deliver/etsi_gs/NFV-INF/001_099/001/01.01.01_60/gs_NFV-INF001v010101p.pdf

[12] R. Russell, J. Stuart, and P. Norvig, “Artificial Intelligence: A Modern Approach,” 34d ed., New Jersey: Prentice Hall, 2020.

[13] L. Globa and N. Gvozdetska, "Comprehensive Energy Efficient Approach to Workload Processing in Distributed Computing Environment," 2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Odessa, Ukraine, 2020, pp. 1-6, doi: 10.1109/BlackSeaCom48709.2020.9235010..

[14] L.S. Globa, R.L. Novogrudska, and A.V. Koval, “Ontology Model of Telecom Operator Big Data,” in Proceedings of IEEE International Black Sea Conference on Communication and Netwoorking (BlackSeaCom), June 2018, pp. 1-5, doi:10.1109/BlackSeaCom.2018.8433710.

[15] L. Globa, I. Svetsynska, and E. Volvach, “Computation of providing services integral quality index,” Information and Telecommunication Sciences, no. 1, pp. 34-42, 2018.

[16] P. Tanwar, T. Prasad, and K. Dutt, “A Tour Towards the Various Knowledge Representation Techniques for Cognitive Hybrid Sentence Modeling and Analyzer,” International Journal of Informatics and Communication Technology (IJ-ICT, vol. 7, no. 3, pp. 124-134, 2018, DOI:10.11591/ijict.v7i3.pp124-134.

[17] M. Rosing, W. Laurier, and S. Polovina, “The Value of Ontology,” The Complete Business Process Handbook, vol. 1, pp.91-100, 2018.

[18] L. Globa, R. Novogrudska and O. Oriekhov, "Method of heterogeneous information resources structuring and systematizing for Internet portals development," in Eurocon 2013, July 2013, pp. 319-326, DOI: https://doi.org/10.1109/EUROCON.2013.6625003

[19] M. Uschold, J. Bateman, M. Bennett, R. Brooks, M. Davis, A. Dima, at al., “Making the case for ontology,” Applied ontology, vol. 7, pp. 373-373, 2012,doi:10.3233/AO-2012-0110.

[20] 5G security for transformed industries by Ericsson. 2018 [Online]. Available: https://www.ericsson.com/en/security

[21] 3GPP TS 33.401, “3GPP System Architecture Evolution (SAE); Security architecture”. [Online]. Available: https://itectec.com/archive/3gpp-specification-ts-33-401/

[22] UMTS Security Awareness. 3GPP/PCG#13 Meeting. Seoul, Korea. 6 October 2004. [Online]. Available: http://www.3gpp.org/ftp/PCG/PCG_13/DOCS/PDF/PCG13_19.pdf

[23] L. Globa and N. Gvozdetska, "Comprehensive Energy Efficient Approach to Workload Processing in Distributed Computing Environment," 2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Odessa, Ukraine, 2020, pp. 1-6, doi: 10.1109/BlackSeaCom48709.2020.9235010.

Simulated annealing metaheuristic with greedy improvement for road segments selection problem

1st Dobroslav Grygar
Department of Mathematical Methods and Operations Research
University of Zilina
Zilina, Slovakia
dobroslav.grygar@fri.uniza.sk

2nd Michal Kohani
Department of Mathematical Methods and Operations Research
University of Zilina
Zilina, Slovakia
michal.kohani@fri.uniza.sk

***Abstract*—Battery-assisted trolleybuses represent the potential for reducing emissions produced by public transport. This vehicle is a trolleybus equipped with a battery as an additional power source. The battery is charged in motion using overhead contact wires. The vehicle is then able to overcome road segments with no wires installed. To deploy such technology into some area, we need to solve location problem on digraph edges. This paper focuses on using metaheuristic simulated annealing with greedy improvement for designing overhead contact line infrastructure for battery-assisted trolleybuses. Results obtained by metaheuristic will be compared to results obtained by the exact approach.**

***Index Terms*—Simulated annealing, greedy heuristic, battery-assisted trolleybus, location problem on edges.**

I. Introduction

The major cause of global warming is greenhouse gases. They include carbon dioxide, methane, nitrous oxides and in some cases chlorine and bromine-containing compounds [1]. Electromobility is the current solution for greener and more sustainable transport. Public transport which is using a road network can be almost fully electrified. There are already vastly deployed existing solutions, such as trolleybuses or electro buses. A combination of these vehicles is called a battery-assisted trolleybus. It is a vehicle based on a standard trolleybus equpped with an additional battery. Using this battery vehicle can overcome road segments with no overhead contact lines installed [2]. Such vehicle is in Figure 1.

This freedom of movement allows us to solve an interesting optimization problem. In this case to propose a minimal overhead contact lines network, for the needs of this type of vehicle. It is a location problem on digraph edges [2].

This paper complements and expands our previous work by introducing a metaheuristic approach for solving such a problem. This time we used Simulated Annealing with greedy improvement. Results obtained by this heuristic will be compared to results obtained by the exact approach. Computational tests were performed using multiple data instances based on the road network and public transport schedules of Žilina, Slovakia [2].

Fig. 1. Battery-assisted trolleybus used in Žilina on road with no overhead contact lines [3]

II. State of the art

Papers related to the deployment of battery-assisted trolleybuses can be divided into several categories. First of all, we mention those that deal with vehicles themselves. Vehicle characteristics are described, for example, by [4], [5] and [6]. The authors deal with the characteristics of these vehicles, which they observed in a test deployment on selected schedules.

Since battery-assisted trolleybuses contain a battery, it is necessary to know its parameters. The articles [5], [7], [8] and [9] are valuable. Authors mentioned, for example, the usable capacities of batteries, charging speed, or performance in extreme cold or at high temperatures.

The economy of public transport is an important part of planning. We are pleased to state that of all the electrically powered public transport vehicles with batteries, battery-assisted trolleybuses are the most advantageous. This is confirmed by the authors of the articles [10], [11] or [12].

The solution of optimization problems associated with the complex deployment of battery-assisted trolleybuses has not been previously investigated. The authors of the articles [13], [14] dealt with the optimization of induction lines for vehicle charging.

In our previous publications, we have summarized the limiting factors of vehicles that affect the optimization in

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

[15]. Next, we focused on the transition based model of the problem [16]. Then we explored the possibilities of location-based linear models and compared different formulations [17].

Metaheuristic Simulated Annealing is used to solve various optimization problems. For example, the authors [20] or [21] described the algorithm and tested performance by solving selected problems.

III. Problem definition

Traditionally location problem means selecting nodes in a graph according to some rules. For example locations of fire stations, warehouses or other types of civic amenities [23], [22]. The problem of designing the infrastructure for the operation and charging of battery-assisted trolleybuses is a different story. It can be defined as a location problem on the edges. So, edges in the road network digraph need to be selected. The resulting edges are recommended for building overhead contact wires. The main task is the reduction of the building cost of overhead contact wires [2]. The illustration of such a network is in figure 2.

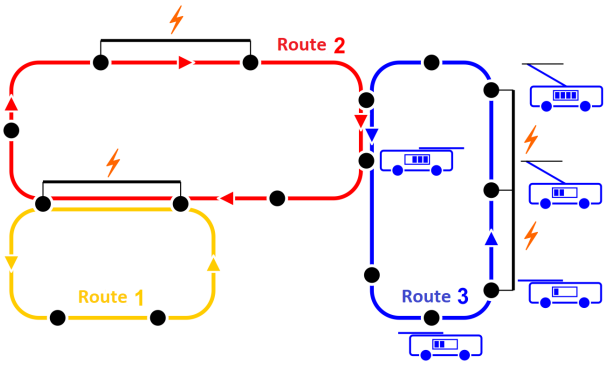


Fig. 2. Illustration of battery-assisted trolleybus routes, charging and battery state.

IV. Mathematical model

Let variable $y_i \in 0, 1$ decide on building overhead contact wires at road segment i. Then z_{rj+1} is a decision variable modelling battery state of vehicle schedule r after passing segment j. Further, input data describing the problem are defined: B_{max} is a maximal allowed state of the battery in kWh and B_{min} is a minimal allowed state of the battery in kWh. CO is energy consumption in kWh/m and CH is charging constant in kWh/m. Moreover, m is vehicle schedules count, n is the number of road segments, D_i is the length of road segment i and $a(r, i)$ is the index of road segment j in vehicle schedule r, while $q(r)$ is the count of road segments in vehicle schedule r [2], [17].

$$min \sum_{i=1}^{n} D_i \cdot y_i \tag{1}$$

$$z_{r,1} = B_{max} - B_{min} \quad r = 1..m. \tag{2}$$

$$z_{r,j+1} - z_{r,j} \leq -CO \cdot D(a_{r,j}) \cdot (1 - y(a_{r,j})) + CH \cdot D(a_{r,j}) \cdot y(a_{r,j}) \quad for \quad r = 1..m; \quad j = 1..q(r). \tag{3}$$

$$z_{r,j+1} \leq B_{max} - B_{min} \quad for \quad r = 1..m; \quad j = 1..q(r). \tag{4}$$

$$y_i \in \{0, 1\} \quad for \quad i = 1..n. \tag{5}$$

$$z_{r,j} \geq 0 \quad for \quad r = 1..m; \quad j = 1..q(r). \tag{6}$$

The objective function (1) consists of the minimization of the total building cost of overhead wires. Constraint (2) sets the initial available state of battery for the vehicle before the start of the schedule. Constraint (3) counts charge or discharge after the vehicle passes road segment *j* in schedule *r*. Then constraint (4) limits the maximally available state of the battery. Obligatory constraints are numbered as (5) and (6) [2], [17].

V. Simulated annealing with greedy improvement

Metaheuristics make it possible to find relatively good solutions to optimization problems. Unlike basic heuristics, they can leave the local minimum, which allows a larger set of admissible solutions to be explored [20].

Simulated Annealing is a metaheuristic based on real-world phenomena. It is based on a method of metal processing, where the metal heated to a certain temperature is gradually cooled [20], [24].

Greedy heuristic starts from an admissible solution in which all or some road segments are covered by the overhead contact lines. The algorithm gradually removes (if possible) those road segments which are least frequently used by public transport vehicles. The algorithm ends if all segments have been checked. This approach is based on the assumption that it is reasonable to cover road segments often used by vehicles [2], [25].

For a better overview of implementation, we present the flowchart of the implemented metaheuristic algorithm 3.

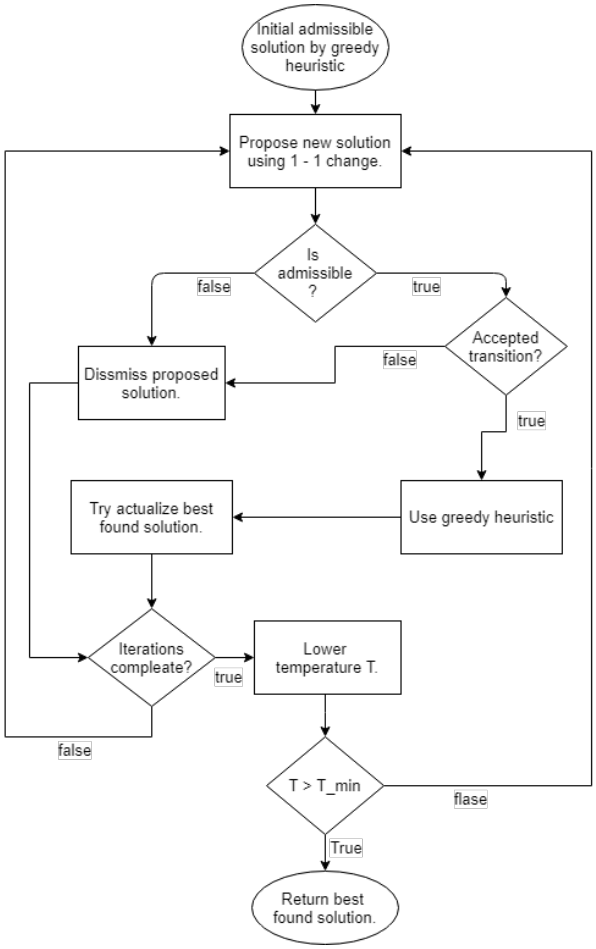


Fig. 3. Flowchart algorithm

In individual iterations of the Simulated Annealing heuristics, we use the 1 - 1 exchange operation to obtain neighbours solutions. This operation removes one cowered road section and adds another from the set of not cowered ones. If such a solution is admissible and at the same time is accepted according to the formula:

$$p(x', x, T) = e^{-(f(x')-f(x))/T},$$

then, the greedy heuristic is applied to this solution. This ensures a gradual reduction in the number of road segments covered. Also, it allows searching for other subsets of solutions. Metaheuristic ends when the temperature is reduced to the minimum level [2].

VI. Data and instances

The road network of the city of Žilina was selected as digraph testing data. Žilina is city located in the north-western part of Slovakia with eighty-one thousand citizens. We used the schedules of buses and trolleybuses operating in working days.The schedules were provided by transport company (DPMŽ). Road segments used by public transport vehicles from OpenStreetMap geographic data, were selected. Then we generated data files using simulation software (OptSim) [2].

We have prepared nubmer of differently sized instances. In the simplest case, it was one bus schedule, then 5 schedules, or schedules of all buses of trolleybuses, and finally the schedules of all vehicles (buses and trolleybuses combinated) in the city. The parameters of instances can be found in the table I.

TABLE I
Benchmark instances parameters.

Instances	**Vehicles count**	**Individual segmens count**	**Average schedule segments count**	**Totoal lenght of used segments (m)**
Bus 1	1	215	426,0	40 809
Tbus 1	1	129	1146,0	17 449
Bus 5	5	486	954,2	100 041
Tbus 5	5	254	946,2	37 469
Tbus ZA	30	299	1320,4	45 174
Bus ZA	30	753	764,2	209 445
ZA	60	779	1042,3	212 618
ZA two d.	60	779	2084,6	212 618

It is well known that the capacity of electric vehicle batteries and energy consumption is affected by the ambient temperature. Therefore, we prepared three scenarios, spring, summer and winter one. Energy consumption is selected based on data from observations of battery-assisted trolleybus operation [5].

TABLE II
Scenarios and vehicle parameter.

Scenario	Max battery state (kWh)	Min battery state (kWh)	Charging (kWh/m)	Power consuption (kWh/m)
spring	10	40	0.0026	0.0013
summer	10	40	0.0026	0.0023
winter	10	30	0.0026	0.0023

VII. Numerical experiments

Computational experiments were performed on a workstation with the following hardware specifications: Processor Intel Core i5-7200U 2.5Ghz with 3.10Ghz turbo boost (two cores and four threads), paired with 16 GB of DDR4 2133MHz RAM. For obtaining exact results we have solved the linear model using Xpress IVE Optimizer solver. Metaheuristic results were obtained using our software tool developed in C# programming language [2].

Metaheuristics are sensitive to the selection of input parameters values. Initial experiments showed that the parameter T (temperature) has the greatest influence. Therefore, the values of the other parameters were fixed on the values in the table IV and we performed a series of experiments, using 10 runs of heuristic. We used a spring dataset with all vehicles in Žilina. The results of experiments with different values of the T parameter are in the table III.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE III
INFLUENCE OF T PARAMETER VALUE ON SA PERFORMENCE

T	OV (m) min.	OV (m) avg.	Time (s) avg.
50	27601	27889	21.70
100	27198	27443	20.03
200	27189	27504	18.77
500	27037	27271	14.92
750	26909	27209	15.56
1000	26923	27215	16.84

Based on experiments with parameter T value setting, we state that the value of this parameter has a major influence on the quality of the obtained solutions. Using low values, the vast majority of transitions to neighbouring solutions were rejected, therefore a majority of admissible solutions was not explored. Increasing the value of the parameter made sense up to the value $T = 750$. We have chosen this value to perform additional experiments. At $T = 1000$, the improvements have not been so significant. In addition to that, we also observed the influence of this parameter on computational time. The shortest time was achieved at $T = 500$. At the selected value of $T = 750$, the calculation time is still favourably short. The table IV contains the resulting values of the parameters, which we used for the input setting of metaheuristics, for all the following experiments [2].

TABLE IV
INPUT PARAMETERS AND VALUES USED IN HEURISTIC

Parameter	Value
T	750
T_min	0.001
alpha	0.9
iterations before temperature change	30

The tables V, VI and VII contain the results of the best one of 10 runs of optimizing datasets using Simulated Annealing. This is because metaheuristic is affected by random generator. The results obtained heuristically are then compared with the results obtained by the exact approach. If the result is marked with "*" then the optimal result was not obtained in time [2].

TABLE V
RESULTS COMPARISON OF EXACT AND SA APPROACH - SPRING SCENARIO.

Instance	OV (m) Exact	OV (m) SA	Diff. OV (%)	Time (s) Exact.	Time (s) SA
Bus 1	6859	7008	2.17	1.35	0.14
Tbus 1	4911	4928	0.35	0.73	0.22
Bus 5	15109	16327	8.06	6.44	1.00
Tbus 5	6655	7333	10.19	3.01	0.91
Tbus ZA	8895	9710	9.16	67.25	7.76
Bus ZA	23198	25607	10.38	842.98	6.93
ZA	24434	26909	10.13	3254.21	16.09
ZA dva dni	*26976	28125	4.26	15585.50	22.41

TABLE VI
RESULTS COMPARISON OF EXACT AND SA APPROACH - SUMMER SCENARIO.

Instance	OV (m) Exact	OV (m) SA	Diff. OV (%)	Time (s) Exact.	Time (s) SA
Bus 1	12308	13050	6.03	0.44	0.11
Tbus 1	7458	7505	0.63	0.45	0.25
Bus 5	25898	29217	12.82	6.72	1.05
Tbus 5	10669	12305	15.33	3.56	0.83
Tbus ZA	13653	14786	8.30	138.44	7.75
Bus ZA	42518	46347	9.01	157.91	6.83
ZA	*44471	47000	5.69	15591.10	17.89
ZA dva dni	*47100	48958	3.94	15610.40	21.86

TABLE VII
RESULTS COMPARISON OF EXACT AND SA APPROACH - WINTER SCENARIO.

Instance	OV (m) Exact	OV (m) SA	Diff. OV (%)	Time (s) Exact.	Time (s) SA
Bus 1	13912	14811	6.46	0.50	0.10
Tbus 1	7685	7755	0.91	0.90	0.20
Bus 5	*27590	30376	10.10	9727.90	0.90
Tbus 5	*11184	12573	12.42	5082.86	0.77
Tbus ZA	*14114	15075	6.81	5985.44	8.31
Bus ZA	*46455	49392	6.32	15063.40	6.16
ZA	*47760	50339	5.40	15598.00	16.61
ZA dva dni	*49684	51269	3.19	15600.40	21.15

VIII. CONCLUSION

Solving the problem of designing a minimal trolley network for operation and charging battery-assisted trolleybuses opens possibilities for more use of sustainable public transport vehicles. In this paper, we presented a mathematical model of the problem and applied metaheuristic Simulated Annealing with greedy improvement. Then we compared results found by exact approach and metaheuristic. The main benefit of this study is in the extension of available approaches to solving such problem [2].

Simulated Annealing metaheuristics used to solve the problem of deploying charging road segments is one of the possible approaches. The presented experiments show that metaheuristics allow us to find acceptable solutions to the problem in a short computational time. How close the optimum sits the solution depends on the instance as well as the quality of the input solution. In our case, the input solutions used were found by greedy heuristics. These solutions provided be a good input into metaheuristic and have been significantly improved with Simulated Annealing. Relatively good solutions were found for large datasets, which contained all vehicles in the city of Žilina. Compared to the exact approach, there was a significant saving of computational time. If such a problem is needed to be solved operationally, this time efficiency would provide an interesting alternative to slower but more accurate exact approach [2].

In future, we can explore the decomposition approach. It means dividing the road network digraph into smaller components. This could be used as a potentially beneficial way of solving similar problems. In this way, the problem could be solved as a set of subproblems.

ACKNOWLEDGMENT

Authors would like to thank VEGA 1/0689/19 - Optimal design and economically efficient charging infrastructure deployment for electric buses in public transportation of smart cities.

REFERENCES

[1] U. Shahzad, 2015. Global Warming: Causes, Effects and Solutions.

[2] D. Grygar, "Efektívne algoritmy na riešenie úlohy rozmiestnenia nabíjacích úsekov v dopravnej sieti," Dissertation thesis, University of Žilina, Faculty of Management Science and Informatics; Department of Mathematical Methods and Operations Research, 2021.

[3] DPMŽ. "Úspešná skúšobná prevádzka parciálnych trolejbusov," 2020. http://www.dpmz.sk/n684/. [online cit. 10.04.2021].

[4] M. Wołek, M. Wolanski, M. Bartłomiejczyk, O. Wyszomirski, K. Grzelec,and K. Hebel, "Ensuring sustainable development of urban public transport:A case study of the trolleybus system in gdynia and sopot (poland),"Journal ofCleaner Production, vol. 279, p. 123807, 2021.

[5] M. Bartłomiejczyk, "Practical application of in motion charging: Trolleybuses service on bus lines," 2017 18th International Scientific Conference on Electric Power Engineering (EPE), Kouty nad Desnou, 2017, pp. 1-6.

[6] M. Bartłomiejczyk, V. Stýskala, R. Hrbac, M. Połom. "Trolleybus with traction batteries for autonomous running," 2013.

[7] A. Montoya, Ch. Guert, J. Mendoza, J. Villegas. "The electric vehicle routing problem with nonli-near charging function," Transportation Research Part B: Methodologi-cal. 2017, 103, p. 87–110.issn0191-2615. Green Urban Transportati-on.

[8] D. Göhlich, A. Kunith and T. Ly, "Technology Assessment Of An-Electric Urban Bus System For Berlin," WIT Transactions on The BuiltEnvironment. 2014, 138, p. 13. ISBN: 9781845647780.

[9] A. Bruce, "New Trolleybus systems," The electric tbus group, http://www.tbus.org.uk/lifecycle.htm, 3.12.2020, [Online citation: 11.2.2021]

[10] A. Ritter, P. Elbert, Ch. Onder. "Energy Saving Potential of a Battery-Assisted Fleet of Trolley Buses," 2016. IFAC-PapersOnLine. 49. 377-384. 10.1016/j.ifacol.2016.08.056.

[11] O. Olsson, A. Grauers, S. Pettersson. "Method to analyze cost effectiveness of different electric bus systems," EVS29 Symposium Montréal, Québec, Canada, 2016.

[12] F. Bergk, K. Biemann, U. Lambrecht, R. Pütz, H. Landinger. "Potential of In-Motion Charging Buses for the Electrification of Urban Bus Lines Journal of Earth Sciences and Geotechnical Engineering," vol.6, no. 4, 2016, 347-362, ISSN: 1792-9040, Scienpress Ltd, 2016.

[13] I. Hwang, Y. Jang, J. Ko Y, M. S Lee, "System Optimization for Dynamic Wireless Charging Electric Vehicles Operating in a Multiple-Route Environment," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 6, pp. 1709-1726, June 2018. doi: 10.1109/TITS.2017.2731787

[14] Z. Chen, F. He, Y. Yin, "Optimal deployment of charging lanes for electric vehicles in transportation networks," Transportation Research Part B: Methodological, Volume 91, 2016, Pages 344-365, ISSN 0191-2615, https://doi.org/10.1016/j.trb.2016.05.018.

[15] D. Grygar, M. Kohani, R. Stefun, P. Drgona, "Analysis of limiting factors of battery assisted trolleybuses," Transportation Research Procedia, Volume 40,2019, Pages 229-235, ISSN 2352-1465.

[16] D. Grygar, M. Kohani. "Linear Model Adjustment and Approximate Approach for Creating Minimal Overhead Wires Network for Vehicle Schedules," 2020. In Proceedings of the 9th International Conference on Operations Research and Enterprise Systems - Volume 1: ICORES, ISBN 978-989-758-396-4, ISSN 2184-4372, pages 187-193.

[17] D. Grygar, M. Kohani. "Location based linear models for digraph edgeselection for overhead contact wires planning," 2021. MIST - Mathematics in Science and Technologies.

[18] M. Rogge, S. Wollny, D.U. Sauer, "Fast Charging Battery Buses for the Electrification of Urban Public Transport—A Feasibility Study Focusing on Charging Infrastructure and Energy Storage Requirements," Energies 2015, 8, 4587-4606.

[19] A. Kunith, R.Mendelevitch, D.Goehlich, "Electrification of a city bus network - An optimization model for cost-effective placing of charging infrastructure and battery sizing of fast-charging electric bus systems," International Journal of Sustainable Transportation, 2017, 11:10, 707-720.

[20] R.W. Eglese, "Simulated annealing: A tool for operational research," European Journal of Operational Research, Volume 46, Issue 3, 1990, Pages 271-281, ISSN 0377-2217.

[21] H. S. Park, C. Won Sung, "Optimization of steel structures using distributed simulated annealing algorithm on a cluster of personal computers. Computers and Structures," 2002, 80(14-15), 1305-1316.

[22] J. Janacek, L. Buzna. "An acceleration of Erlenkotter–Körkel's algoriths for uncapacitated facility location problem," Annalsof Operations Research. 2008, 164, no. 1, p. 97–109. Issn: 0254-5330.

[23] P. Czimmermann, "Location problems in transportation networks," 2016 Communications –Scientific Letters of the University of Zilina, 18(3), 50-53.

[24] M. Mašurik, "Heuristiky a metaheuristiky," Bratislava, Slovenská republika 2015. Bakalárska práca. Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra informatiky.

[25] V. Stozhkov, V. Boginski, O.A. Prokopyev, et al. "A simple greedy heuristic for linear assignment interdiction," Ann Oper Res 249, 39–53 (2017).

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

AlphaZero with Real-Time Opponent Skill Adaptation

Marek Baláž
Faculty of Management Science and Informatics
University of Žilina
Žilina, Slovakia
Email: Marek.Balaz@fri.uniza.sk

Peter Tarábek
Faculty of Management Science and Informatics
University of Žilina
Žilina, Slovakia
Email: Peter.Tarabek@fri.uniza.sk

***Abstract*—Reinforcement learning based methods achieved super-human score in many complex games. Ability to play on super-human level can be impractical when playing against casual players as the skill gap can be too big for the game to be enjoyable and challenging. In this paper, we propose modification of AlphaZero method that allows us to adapt agent to weaker opponent skill level during a single game. We added another output head to the neural network that predicts remaining game length. Based on this prediction, we added new action selection mechanism to Monte Carlo Tree Search. This mechanism allows us to make trade-off between original and new action selection strategy. The results of experiments show that the proposed modifications reduce the gap between strong and weak agents by increasing the number of draws which is our primary measurement of adaptability.**

***Index Terms*—reinforcement learning, AlphaZero, Monte-Carlo tree search, real-time adaptation, game balancing**

I. Introduction

In 2016, algorithm from the field of reinforcement learning (RL) [1] defeated the best world player Lee Sedol in game GO [3]. This success can be considered as one of the greatest achievements of machine learning given the huge number of possible board positions. Today, training an agent to achieve super-human score is possible for many games and simulators [20]. When casual player plays competitive game against algorithm on super-human level, game could feel mismatched. As the skill gap between an agent and an opponent is very large, playing the game can end up being too difficult and unexciting. In this paper we propose modification of popular algorithm AlphaZero [5] which aims to adapt game difficulty to weaker opponents during a single game. The idea is composed of three parts:

- Before AlphaZero training - we modify neural network by adding third output head predicting remaining length of the game.
- During AlphaZero training - the new output head learns to predict remaining length of the game from dataset of self-played games.
- After AlphaZero training - we modify AlphaZero decision making by modifying existing Monte-Carlo tree search (MCTS) selection strategy by adding new term called prolonging strategy. This strategy uses remaining game length prediction to adapt to an opponent skill level during a single game.

We tested two methods to combine existing selection strategy with proposed prolonging strategy: alpha and ratio trade-off. Alpha trade-off depends on the current game step and the remaining length of the game. Ratio trade-off is based on the quality of actions performed by agent and opponent. This quality is evaluated by auxiliary MCTS with original selection strategy.

Modified AlphaZero was trained on a simulator of 5x5 Tic-Tac-Toe game. The first player who connects four symbols in a line wins the game. Tic-Tac-Toe is a zero-sum game with two players in which each player can play higher amount of different strategies. Small dimensions of game allow us to test the proposed method against a larger variety of opponents. These opponents are represented by different neural network architectures and MCTS parameters. We trained 460 different weaker agents to test the proposed modification of AlphaZero. We split weaker agents into three groups according to their performance against tested agent with original MCTS selection mechanism. Subsequently, we tested parameters of proposed method against opponents from all three groups. Experimental results show the ability of the proposed method to adapt to opponent skill level during a single game measured as an increase in the number of draws.

Although proposed modifications are implemented in AlphaZero algorithm, these modifications can be applied to other model-based reinforcement learning algorithms using MCTS. AlphaZero was chosen, because it has achieved state-of-the-art in board games and does not require to learn model of simulator for planning such as MuZero algorithm [6].

II. Related work

Algorithm of reinforcement learning (RL) is called agent. This agent performs actions in environment. For each action a_t at time t, agent receives reward r_t. Reward could be positive, negative or zero. Decision making is based on the current state of environment s_t and the agent's policy π. Policy is a function that maps perceived states of environment to actions.

RL agents can be divided into two main categories: model-free and model-based [1]. Model-free agents cannot know how will environment change after the action has been performed,

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

whereas policy function of model-based agents uses model of environment for planning. Model is represented by Markov-decision process [2] consisting of state transition function and reward function.

Model-based algorithms have achieved super-human level in many zero-sum games [8], since AlphaGo defeated best player in game GO [3]. Successors of AlphaGo are AlphaGo Zero [4], AlphaZero and MuZero. In AlphaGo, neural network model learns from replays of expert games before the RL training phase, whereas AlphaGo Zero learns itself only via methods of RL. AlphaZero is a new generation of AlphaGO Zero agent with tuned parameters, better network architecture and is aimed on games GO, Chess and Shogi. AlphaGo, AlphaGo Zero and AlphaZero agents use copy of environment as model, whereas MuZero learns model of environment via deep neural network [6]. Learning this model is much more time and computation consuming.

A. AlphaZero

AlphaZero's policy is formed by deep neural network, MCTS [7] and model of environment [5]. Neural network (with parameters θ) predicts two outputs: probability distribution of actions $P(s,\cdot|\theta)$ in state s and value of state $V^{\pi}(s|\theta)$ under policy π. Value of state represents weighted mean of possible future rewards according to current policy π. Gao et al. [14] proposed to add third output head to neural network called action-value. They argued that this modification could lead to more efficient MCTS.

In general, most of the board-game engines use tree search for planning in which nodes represent states and edges possible actions in these states [9], [10], [11]. AlphaZero uses MCTS in each state during the game episode. Each node stores information whether it is a terminal or a non-terminal node. Terminal nodes only hold terminal reward (-1 for lose, 0 for draw and 1 for win). Non-terminal nodes hold:

- State of environment s
- Probability distribution of actions predicted by neural network $P(s,\cdot)$ (illegal actions are set to zero)
- Vector representing number of visits of each action $N(s,\cdot)$
- Vector of Q values for each action $Q(s,\cdot)$ representing mean values of future rewards

MCTS performs following four phases several times in each state:

- Phase of selection - Tree is traversed from root to leaf node in order to find not performed action or terminal node. During this traversal, action in each node is selected according to the variant of PUCT (Predictor + Upper Confidence bounds applied to Trees) algorithm [3], [12] which defines trade-off between exploration and exploitation described by Equation (1) and (2) [5].
- Phase of simulation - Selected action will be executed in model of environment in parent state s_{t-1}. After the action is executed, agent receives information about new state s_t.
- Phase of expansion - If new state is non terminal, neural network will predict outputs for new state. Otherwise, terminal reward will be observed. In both cases, new node is created according to these values.
- Phase of back propagation - Q values of traversed nodes (during the phase of selection) are updated according to Equation (3). Value G represents predicted value of state $V^{\pi}(s|\theta)$ in non-terminal nodes or terminal reward in terminal nodes. After the Q values are updated, the number of visits $N(s,a)$ is increased by one for all selected actions.

$$U(s_t,a) = c_{puct} \times P(s_t,a) \times \frac{\sqrt{\sum_{b\in A} N(s_t,b)}}{N(s_t,a)} \tag{1}$$

$$a_t = \underset{a}{\operatorname{argmax}} \left[Q(s_t,a) + U(s_t,a)\right] \tag{2}$$

$$Q(s_t,a_t) = \frac{N(s_t,a_t) \times Q(s_t,a_t) + G}{N(s_t,a_t)+1} \tag{3}$$

Probability distribution of actions is computed by Equation (4) after predefined number of MCTS cycles are performed. AlphaZero creates a dataset from self-play games in which agent takes actions by obtained probability distribution. In the field of RL, the dataset is called replay buffer. Exploration is supported by adding Dirichlet noise [13] to the predicted probability distribution. After each game episode, agent stores data to experience replay buffer. Stored data consists of a set of reached states, a set of probability distributions obtained from MCTS and an information about the result of the game.

$$\pi(s_t,\cdot) = \frac{N(s_t,\cdot|s_t = s_{root})}{\sum_{b\in A} N(s_t,b)} \tag{4}$$

During the training, agent chooses random batch of size N from replay buffer. Batch consists of triplet information $(s,\pi(s,\cdot),z)$, where z represents result for current game state. Loss function is computed based on the predicted probability distribution and the state value shown in Equation (5). Left part consists of mean square error between predicted value $V^{\pi}(s_n|\theta)$ and game result z_n. Right part consists of cross entropy between probability distribution obtained from MCTS $\pi(s_n,\cdot)$ and predicted probability distribution $P(s_n,\cdot|\theta)$.

$$L(\theta) = \frac{1}{N}\sum_{n=0}^{N}\left[(V^{\pi}(s_n|\theta) - z_n)^2 - \pi(s_n,\cdot) \times logP(s_n,\cdot|\theta)\right] \tag{5}$$

B. Adaptation of RL agent

There are two different meanings of term adaptability in reinforcement learning zero-sum games.

In the first case, agent plays against stronger opponent in order to adapt to his level and defeat him. After adaptation, agent usually plays against another stronger opponent to further improve [16]. In complex environments with many different strategies agent can also play against more opponents

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

with different strategies in parallel fashion [17]. The goal of RL agent is to maximize collected rewards during environment episode. This often results in super-human score [20].

On the other hand, there are cases in which trained agent tries to adapt to weaker opponent to balance the difficulty of the game. Some researches model opponent's behavior via another neural network [18] or function that evaluates agent and opponent skill level [19]. Both recommended publications need to learn another predictive model at the end of the agent training. Furthermore, these approaches were tested in shooter games with less action space and smaller consequences of actions than in the case of the board games.

III. Proposed AlphaZero modification

The common approach to train AlphaZero agent is by collecting positive rewards for winning the game. During the training, AlphaZero increases probabilities of actions which are part of the winning strategies and decreases other actions. At the end of the training, almost all actions (except actions in winning strategies) have very small probabilities. This results in behavior where the agent tries to win the game as soon as possible. We call it original strategy. This non-existing difference in probabilities for majority of actions makes the adaptability to opponent skill level difficult.

Our approach uses new strategy based on remaining length of the game. Therefore, we modified neural network by adding a third output head $M^{\pi}(s,\cdot)$. This new output predicts probability distribution of the remaining length of the game (number of actions) assuming the best policy π. Loss of the third output is composed of cross entropy and one-hot encoded vector of possible remaining game lengths. This modification is inspired by [22], where remaining length of the game was predicted as a regression task. They use this modification solely to improve the training.

The main idea is to first obtain agent with superior score, called strong agent. We use standard approach to train AlphaZero agent in order to obtain sufficiently high score. After the training, we modify selection phase in MCTS to provide adaptation ability against worse performing agents, called weak agents or opponents. Each non terminal node receives new vector of L values for each action $L(s,\cdot)$. $L(s,a)$ value represents expected number of steps from state s to the end of the game under the best policy π when using action a. Expected length is updated during the back propagation phase, similar to Equation (3), by prediction of remaining length of the game (Equation (6)). During the MCTS back propagation we need to back propagate also the information about the remaining length of the game. In a case of non-terminal node, M is set to prediction $M^{\pi}(s,\cdot)$. In a case of terminal node, M is set to 0. The value of M is then increased by 1 for each transition when traversing back to the parent node.

$$L(s_t,a_t) = \frac{N(s_t,a_t) \times L(s_t,a_t) + M}{N(s_t,a_t) + 1} \tag{6}$$

MCTS phase of selection uses variant of PUCT algorithm to find the best action in current state. We suggest different strategy for action selection called prolonging strategy.

During the phase of selection, agent uses original MCTS action selection algorithm for nodes which represent opponent actions. In nodes that represent agent actions, agent use proposed action selection method. This method consists of the original action selection strategy combined with the new prolonging strategy defined by Equation (7). Parameter β is used to amplify the importance of actions resulting in long game. $L_{norm}(s_t,a)$ equals to $L(s_t,a)$ divided by the number of remaining steps. This normalization step forces the values to be in $[0,1]$ interval. For environments where the remaining number of steps is not possible to compute or is very high, we can normalize vector itself to $[0,1]$ interval. This new selection strategy aims for adaptability to opponent skill level by prolonging the game.

$$C(s_t,a) = Q(s_t,a) + \beta \times L_{norm}(s_t,a) \tag{7}$$

The overall action selection method combines both, original and prolonging strategies using a trade-off mechanism. We proposed and tested two trade-off methods: alpha and ratio trade-off. Alpha trade-off method is based on the current game step and the total game length. Trade-off is represented by parameter α computed by Equation (8). Overall action selection method is defined by Equation (10), where $B(s_t,a)$ represents original selection strategy (Equation (9)) and C represents proposed prolonging strategy (Equation (7)). At the beginning of the game, agent will play with goal to prolong the game. Therefore, opponent should not be defeated. During the rest of the game, agent's strategy will gradually change from prolonged to original as the game progresses. Although parameter α defines trade-off between strategies based on the current step, overall influence of prolonging strategy during the whole game is amplified by parameter β.

$$\alpha = \frac{current\ number\ of\ step\ in\ episode}{number\ of\ all\ steps\ per\ episode} \tag{8}$$

$$B(s_t,a) = Q(s_t,a) + U(s_t,a) \tag{9}$$

$$a_t = \underset{a}{\operatorname{argmax}} \left[\alpha \times B(s_t,a) + (1-\alpha) \times C(s_t,a)\right] \tag{10}$$

Ratio trade-off method takes into account evaluation of strong agent and opponent actions. Therefore, we use auxiliary MCTS which independently simulates action selection using best policy for both agent and opponents moves. During the simulation, auxiliary MCTS uses the same neural network as strong agent with original action selection algorithm to compute probability distribution. Subsequently, the probability distribution is used to evaluate quality of actions performed in real game. Strong agent stores sum of last W evaluations of his and opponent's actions defined as a and o. Trade-off is computed by Equation (11) as a ratio between sums of these evaluations. If $o > a$, agent will prefer original selection strategy. Otherwise, the prolonging strategy is preferred. Second

trade-off approach is more computationally intensive than first one as it requires auxiliary MCTS.

$$a_t = \underset{a}{\operatorname{argmax}} \left[\frac{o}{a+o} \times B(s_t, a) + \frac{a}{a+o} \times C(s_t, a) \right] \quad (11)$$

IV. Experiments

The proposed method, environment and training were implemented using PyTorch library.

A. Environment

The proposed method was tested on modification of zero-sum game called Tic-Tac-Toe. This modification uses 5x5 board and player who connects four symbols in line wins. Therefore, maximal length of game episode is 25 steps (actions). We use data augmentation during the training. Each finished game is augmented by combinations of rotation and horizontal/vertical flip to produce additional seven games. Subsequently, all eight games are added to the replay buffer. Simulation environment is optimized to run on GPU.

B. Agent architecture and training

We trained two types of agents: strong and weak. Strong agent aims for best performance and uses proposed modifications for opponent skill adaptations. Weak agents aim for variety of skill levels and are used to asses adaptation capabilities of strong agent. To obtained strong agent we tested different parameters such as architecture of neural network, count of MCTS cycles, influence of noise, training length, and etc. Best performing agent uses 200 MCTS cycles, collects 1000 games during one training iteration and stores 1000000 triplets in replay buffer. Noise parameters are set according to recommendation from AlphaZero paper [5]. Neural network consists of convolutions, residual blocks [21] and fully connected layers. Each convolution uses kernel size 3x3, stride 1 and padding set to 1. Each hidden layer is followed by ReLU activation function.

Neural network is composed of:

- Convolution with 96 kernels
- 3× Residual block with 96 kernels
- Policy head:
  - Convolution with 32 kernels
  - Fully connected layer with 256 neurons
  - Fully connected layer with 25 neurons
- Value head:
  - Convolution with 32 kernels
  - Fully connected layer with 128 neurons
  - Fully connected layer with 3 neurons
- Remaining length head:
  - Convolution with 32 kernels
  - Fully connected layer with 256 neurons
  - Fully connected layer with 25 neurons

We modified value head from predicting expected state value to classify expected result of game using 3 classes (win, draw, lose) to improve decision making [22].

We trained weak agents in similar fashion. We use smaller neural network architectures consisted of four convolution layers with 128 kernels. Furthermore, we use less MCTS cycles ranging from 50 to 100.

C. Results

Strong agent played against 460 different weak agents twice (once as a first player and once as a second player). Agent did not lose a single game while playing in best mode. To better assess proposed modifications, we decided to split games into three groups:

- Group 1 consists of games in which strong agent won within 10 moves.
- Group 2 is composed of games in which strong agent won after 10 moves.
- Group 3 is a hardest group as it consists of games which end up in a draw.

We tested proposed action selection strategy using both alpha and ratio trade-off methods. We reported results for different values of β parameter for both methods.

1) Alpha Trade-off: Results for individual groups are presented in Table I-III. First row represents the results for unmodified strong agent (α is fixed to 1). Column $Average\ steps$ represents average length of wins, loses and all games (including draws). Results show that β has similar effect on different groups of opponents. As the parameter β increases, number of wins decreases. Majority of these games are flipped from wins to draws. In the first group consisting of weakest opponents (Table I), strong agent lost few games when using higher value of β. But he was still able to flip majority of games into draws. On the other hand, strong agent with β around 5.0 took draw in most of the games against weaker opponents (Table I and II). Games against opponents from third group usually end in a draw for all values of β parameter (Table III). We can see that parameter β allows us to adapt strong agent to opponent skill level for all three groups. Furthermore, we can use the same value of β in all tested situations. Dependence of draw ratio on β parameter for each group is depicted in Fig. 1.

TABLE I
Group 1 results using alpha trade-off

Beta	Wins	Losses	Draws	Average steps wins / loses / all
-	550	0	0	7.8 / - / 7.8
1.0	209	0	341	12.8 / - / 20.3
2.0	82	0	468	16.0 / - / 23.7
3.0	56	0	494	15.0 / - / 24.0
5.0	30	0	520	12.2 / - / 24.3
10.0	29	0	521	12.6 / - / 24.3
15.0	27	17	506	13.7 / 12.9 / 24.0
20.0	27	17	506	13.7 / 12.9 / 24.0
25.0	27	17	506	13.7 / 12.9 / 24.0

2) Ratio Trade-off: Agent played with different values of β and W parameters. Most of the games against opponents from the first and the second group ended in a draw (Table IV

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE II
GROUP 2 RESULTS USING ALPHA TRADE-OFF

Beta	Wins	Losses	Draws	Average steps wins / loses / all
-	123	0	0	12.5 / - / 12.5
1.0	60	13	50	12.4 / 18.0 / 18.1
2.0	46	0	77	14.5 / - / 21.0
3.0	27	0	96	18.1 / - / 23.5
5.0	18	0	105	16.8 / - / 23.8
10.0	17	0	106	16.6 / - / 23.8
15.0	17	0	106	16.6 / - / 23.8
20.0	17	0	106	16.6 / - / 23.8
25.0	16	1	106	19.0 / 9.0 / 24.0

TABLE III
GROUP 3 RESULTS USING ALPHA TRADE-OFF

Beta	Wins	Losses	Draws	Average steps wins / loses / all
-	0	0	247	- / - / 25.0
1.0	5	0	242	16.4 / - / 24.8
2.0	3	0	244	14.7 / - / 24.9
3.0	3	0	244	16.0 / - / 24.9
5.0	1	0	246	20.0 / - / 25.0
10.0	0	0	247	- / - / 25.0
15.0	0	0	247	- / - / 25.0
20.0	0	0	247	- / - / 25.0
25.0	0	1	246	- / 13.0 / 25.0

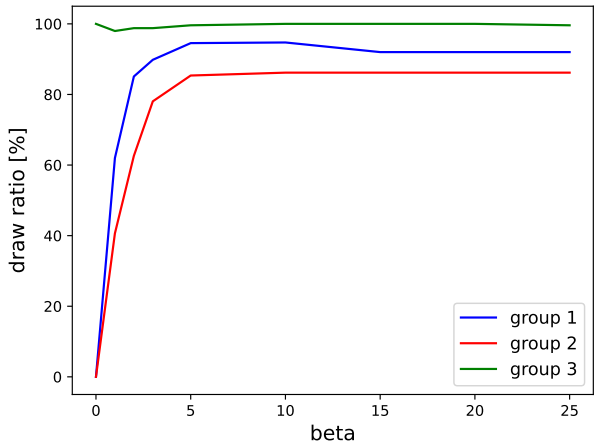


Fig. 1. Ratio between draws and all games for each group against parameter beta for alpha trade-off selection strategy.

TABLE IV
GROUP 1 RESULTS USING RATIO TRADE-OFF

Beta	W	Wins	Losses	Draws	Average steps wins / loses / all
-	-	550	0	0	7.8 / - / 7.8
3.0	1	54	0	496	16.0 / - / 24.1
3.0	4	40	0	510	13.6 / - / 24.2
3.0	8	40	0	510	13.6 / - / 24.2
3.0	25	41	0	509	13.8 / - / 24.2
5.0	1	30	0	520	11.6 / - / 24.3
5.0	4	30	0	520	11.6 / - / 24.3
5.0	8	30	0	520	11.6 / - / 24.3
5.0	25	30	0	520	11.6 / - / 24.3
10.0	1	30	15	505	11.6 / 13.0 / 23.9
10.0	4	30	15	505	11.6 / 13.0 / 23.9
10.0	8	30	15	505	11.6 / 13.0 / 23.9
10.0	25	30	15	505	11.6 / 13.0 / 23.9

TABLE V
GROUP 2 RESULTS USING RATIO TRADE-OFF

Beta	W	Wins	Losses	Draws	Average steps wins / loses / all
-	-	123	0	0	12.5 / - / 12.5
3.0	1	2	0	121	17.0 / - / 24.9
3.0	4	1	0	122	10.0 / - / 24.9
3.0	8	1	0	122	10.0 / - / 24.9
3.0	25	1	0	122	10.0 / - / 24.9
5.0	1	1	0	122	10.0 / - / 24.9
5.0	4	1	0	122	10.0 / - / 24.9
5.0	8	1	0	122	10.0 / - / 24.9
5.0	25	1	0	122	10.0 / - / 24.9
10.0	1	1	0	122	10.0 / - / 24.9
10.0	4	1	0	122	10.0 / - / 24.9
10.0	8	1	0	122	10.0 / - / 24.9
10.0	25	1	0	122	10.0 / - / 24.9

TABLE VI
GROUP 3 RESULTS USING RATIO TRADE-OFF

Beta	W	Wins	Losses	Draws	Average steps wins / loses / all
-	-	0	0	247	- / - / 25.0
3.0	1	0	0	247	- / - / 25.0
3.0	4	0	0	247	- / - / 25.0
3.0	8	0	0	247	- / - / 25.0
3.0	25	0	0	247	- / - / 25.0
5.0	1	0	0	247	- / - / 25.0
5.0	4	0	0	247	- / - / 25.0
5.0	8	0	0	247	- / - / 25.0
5.0	25	0	0	247	- / - / 25.0
10.0	1	0	0	247	- / - / 25.0
10.0	4	0	0	247	- / - / 25.0
10.0	8	0	0	247	- / - / 25.0
10.0	25	0	0	247	- / - / 25.0

and V). We can see that influence of parameter W for higher values of β is insignificant. We think it is caused by almost non-existing difference in evaluations of actions which are not part of the winning strategy. Value of parameter β is important to reach draw, especially in the first group. Group 3 consists of games that originally ended in draw. We can see (Table VI) that strong agent was able to achieve the same results despite the changes introduced in proposed action selection method.

V. CONCLUSION

Playing the game against better opponent with large skill gap can feel mismatched and unexciting. To tackle this problem we proposed modification to AlphaZero that introduce new adaptability mechanism. It allows us to adapt strong agent to opponent skill level during a single game. This adaptation is acquired by additional action selection strategy called prolonging strategy. The main idea is to promote adaptability by

making the game as even as possible. In our scenario, this is done by ending the game in a draw. We measured the adaptability in Tic-Tac-Toe game environment by counting the number of wins, draws and losses. We show that the proposed modifications allow strong agent to flip majority of games from wins to draws. Furthermore, we show stability of this behaviour across different groups of opponents. Both trade-off approaches were able to achieve draw in majority of games. Count of draws depends mainly on parameter β. Agent with alpha trade-off (for $\beta = 5$) reaches draw ratio 94.7% against group 1 and 86.2% against group 2. Agent with ratio trade-off (for $\beta = 5$) reaches draw ratio 94.5% against group 1 and 99.2% against group 2 . Ratio trade-off method shows slightly better results for group 2, but it requires auxiliary MCTS which makes it more computationally demanding. Both methods show worst results for group 1 that consists of weakest agents. These agents very often make almost random actions which in combination with MCTS makes adaptability tricky. We have approached the problem of adaptability to opponent skill level via predicting length of the game. We use this information to push the game outcome into a draw and use the number of draws as the adaptability criterion. In the future, it would be useful to explore other criteria to guide the adaptability procedure. Furthermore, we would like to test proposed methods in zero-sum games with very high or theoretically infinite number of moves such as Go or chess.

ACKNOWLEDGMENT

This work was supported by grant VEGA 1/0089/19 and Grant System of University of Žilina No. 1/2020 (8041).

REFERENCES

[1] Sutton, Richard S and Barto, Andrew G, *Reinforcement learning: An introduction*, 2018, MIT press
[2] van Otterlo, Martijn and Wiering, Marco, *Reinforcement Learning and Markov Decision Processes*, 2012, Springer Berlin Heidelberg
[3] Silver, D., Huang, A., Maddison, C. et al., *Mastering the game of Go with deep neural networks and tree search*, 2016, Nature
[4] Silver, D., Schrittwieser, J., Simonyan, K. et al., *Mastering the game of Go without human knowledge.*, 2017, Nature
[5] Silver, David and Hubert, Thomas, et al., *Mastering chess and shogi by self-play with a general reinforcement learning algorithm*, 2017, arXiv preprint arXiv:1712.01815
[6] Schrittwieser, Julian and Antonoglou, Ioannis, et al., *Mastering atari, go, chess and shogi by planning with a learned model*, 2019, arXiv preprint arXiv:1911.08265
[7] Browne, Cameron B. and Powley, et al., *A Survey of Monte Carlo Tree Search Methods*, 2012, IEEE Transactions on Computational Intelligence and AI in Games
[8] Alan Washburn, Kevin Wood, *Two-Person Zero-Sum Games for Network Interdiction*, 1995, Operations Research 43 (2) 243-251
[9] T. A. Marsland and F. Popowich, *Parallel Game-Tree Search*, 1985, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-7, no. 4
[10] Alan Washburn, Kevin Wood, *Progressive strategies for Monte-Carlo tree search*, 2008, New Mathematics and Natural Computation Vol. 04, No. 03
[11] Chaslot, Guillaume M. J., et al., *Parallel Monte-Carlo Tree Search*, Computers and Games 2008, Springer Berlin Heidelberg
[12] Rosin, C. D., *Multi-armed bandits with episode context*, 2011, Annals of Mathematics and Artificial Intelligence 61
[13] Hida, T. and et. al., *Dirichlet forms and white noise analysis*, 1988, Commun.Math. Phys., https://doi.org/10.1007/BF01225257
[14] Gao, Chao and Müller, Martin and Hayward, Ryan, *Three-Head Neural Network Architecture for Monte Carlo Tree Search*, 2018, Twenty-Seventh International Joint Conference on Artificial Intelligence
[15] Wang, Hui and Emmerich, Michael and Preuss, Mike and Plaat, Aske, *Alternative Loss Functions in AlphaZero-like Self-play*, 2019, IEEE Symposium Series on Computational Intelligence
[16] M. van der Ree and M. Wiering, *Reinforcement learning in the game of Othello: Learning against a fixed opponent and learning from self-play*, 2013, IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning
[17] Luís Filipe, TeófiloNuno Passos, et. al. *Adapting Strategies to Opponent Models in Incomplete Information Games: A Reinforcement Learning Approach for Poker*, Autonomous and Intelligent Systems 2012, Springer Berlin Heidelberg
[18] He, He and Boyd-Graber, Jordan, et. al. *Opponent Modeling in Deep Reinforcement Learning*, 2016, Proceedings of The 33rd International Conference on Machine Learning, Vol. 48
[19] Andrade, Gustavo and Ramalho, Geber, et. al. *Extending Reinforcement Learning to Provide Dynamic Game Balancing*, 2005, Workshop on Reasoning, Representation, and Learning in Computer Games
[20] Volodymyr Mnih, et. al. *Playing Atari with Deep Reinforcement Learning*, 2013, arXiv:1312.5602
[21] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, *Deep residual learning for image recognition*, 2016, Proceedings of the IEEE conference on computer vision and pattern recognition
[22] Gary Linscott and et. al., *Leela Chess Zero*, 2020, Available: https://lczero.org/

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Phishing Detection for Secure Operations of UAVs

Jan Bohacik
Department of Informatics
Faculty of Management Science and Informatics, University of Zilina
Zilina, Slovakia
Jan.Bohacik@uniza.sk

Abstract—**Unmanned Aerial Vehicles (UAV) or drones are used in various domains more and more, including military operations, monitoring, rescue of victims, and transport. Often, UAV resources are developed as web services so that they can be accessed anywhere on the Internet through the World Wide Web. However, this makes them vulnerable to phishing activities of criminals who may try to access these resources and other sensitive information. Therefore, the development of a phishing detection tool based on data mining is presented in this paper. It consists of a browser extension monitoring visited webpages and a backend communicating with the browser extension for the purposes of executing some specific tasks. The browser extension is implemented in JavaScript and the ReactJS framework, and it contains an implementation of classifications with a Bayesian network, decision tree, nearest neighbor classifier and neural network. The backend uses PHP, Python scripts and the Apache HTTP Server. In addition, a browser extension is implemented so that data about webpages can be collected and this data is used for the creation of data mining models. Experimental validation with 10-fold cross-validation and through the browsing of real-world websites show promising results in phishing detection.**

Keywords—UAV; security; phishing; data mining

I. INTRODUCTION

Flying vehicles which do not have a human pilot on board are called Unmanned Aerial Vehicles (UAV) or drones [2]. Traditionally, they were used for tasks in military operations whose capabilities were beyond the capabilities of many. But recent advances in technology and reductions in costs have allowed the spread of UAVs to other areas as well. And so, in addition to military operations as shown in [7], they are currently used in environmental and other monitoring [14], rescue of victims [12], and transport [4]. UAVs use waves on specific radio frequencies for the communication with the ground station so that they receive or send commands. This leads to a situation when the user is restricted to a location, which is not desirable for applications placed anywhere by any user. Therefore, concepts for the inclusion of UAVs into the cloud infrastructure have been developed recently [8]. This inclusion makes the UAV and its resources available to the client ubiquitously. This client could be a person who uses a browser on the Internet or another UAV. In general, cloud infrastructure has servers which are very powerful computers providing services in the cloud. The UAV is included into the cloud as a backend server which provides its services and resources. However, this is a common client-server system which may be vulnerable to phishing techniques in the same way as Internet banking for example.

Phishing is defined as the fraudulent attempt to obtain sensitive data or information by passing oneself off as a trustworthy entity in some digital communication [11]. This data or information may include usernames, passwords, or other sensitive details. It is often accompanied with the redirection of the user to a fake webpage that matches the look and feel of the legitimate one. Fake webpages are easy to make nowadays due to the existence of available user-friendly tools [5]. This is alarming for the users of UAVs who access them and their resources through webpages in a browser, and some anti-phishing approaches are required. Several anti-phishing approaches can be found and each is based on something different [13]. The oldest ones are blacklist-based approaches which check if webpages are in lists of forbidden URL links. However, new fake webpages can appear easily and so the functionality of them is limited these days. Heuristic-based approaches look at the URLs of webpages and apply heuristics for checking their features. Content-based approaches look for some terms in webpages or images in screenshots of webpages. KDD-based approaches where KDD is Knowledge Discovery in Data use attributes extracted from webpages for the creation and use of data mining classification models. There are also hybrid-based approaches which combine several of the approaches. The most complex and potentially most effective approaches are KDD-based ones and their merges in hybrid-based approaches. In addition, KDD-based approaches are actively researched at this moment [3][6][9][10]. Among them, the most used methods are classification methods. These methods use data about instances described by describing attributes and classified into the class attribute for the creation of a data mining model which is used for the classification of new instances. In this paper, a Bayesian network, a decision tree, a nearest neighbor classifier and a neural network are utilized in the development of a browser extension for monitoring visited webpages, of a browser extension for data collection and of a backed communicating with these extensions for the purposes of executing some specific tasks related to describing attributes.

The paper is organized in a way which is explained in this paragraph. The developed browser extension for monitoring visited webpages, browser extension for data collection and backed communicating with these extensions are presented in Section II. In Section III, the webpage data collected with the browser extension for data collection is described in detail. The creation of data mining models with the collected webpage data and carried experimental evaluation are analyzed in Section IV. Section V contains summarized conclusions of results.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

II. Browser Extensions And The Backend

There are two developed browser extensions which have been implemented for the purposes of phishing detection. In general, a browser extension is a module for an Internet browser and it is used for the customization of the browser. One of the developed extensions is for further advancement and for data mining specialists who can collect data about visited webpages with it and assign if particular webpages are phishing or legitimate. In addition, data mining specialists use this data collected about webpages for the development of data mining classification models. These models are utilized in the other developed browser extension which is used for monitoring visited webpages. A screenshot of this extension displayed in a browser is shown in Fig. 1. Currently, four types of created classification data mining models can be imported into the extension: a) a Bayesian network marked as Naive Bayes network in Fig. 1; b) a decision tree marked as j48 decision tree; c) a nearest neighbor classifier marked as NNge; and d) a neural network marked as MLP. During the importation, the models are transformed automatically from files containing representations taken from Weka into scripts in JavaScript, where Weka [15] is an open-source data mining tool. Both of the browser extensions are written in JavaScript and the frontend uses the ReactJS framework for the creation of their user interface.

Another important part of the implementation is the backend. It communicates with the browser extensions through the HTTP GET method and runs on a Linux server with PHP, Python scripts and the open-source Apache HTTP Server. For a successful execution of the backend, packages apache, sqlite, mysql, php7.*, and python have to be installed. Dependencies for Python scripts are in files 'requirements.txt' and they can be installed with command 'pip install -r requirements.txt'. Data about webpages are stored with the open-source MySQL relational database management system. Some tasks related to the data collection about webpages which cannot be done in the browser extensions are performed in the backend as well. In addition, the backed allows to collect data about thousands of webpages without the necessity of visiting each webpage in an Internet browser manually. This functionality uses some webpages which are available on the Internet, contain links to other webpages and allow determination if the webpages on these links are likely to be phishing or legitimate. For the creation of models in Weka, data about webpages has to be selected from the database in the backend. Then, the data is exported to the CSV format and loaded into Weka. The models from Weka are processed by the backend into files which are stored in the backend. When the extensions are started, these files are downloaded by them if updates are detected.

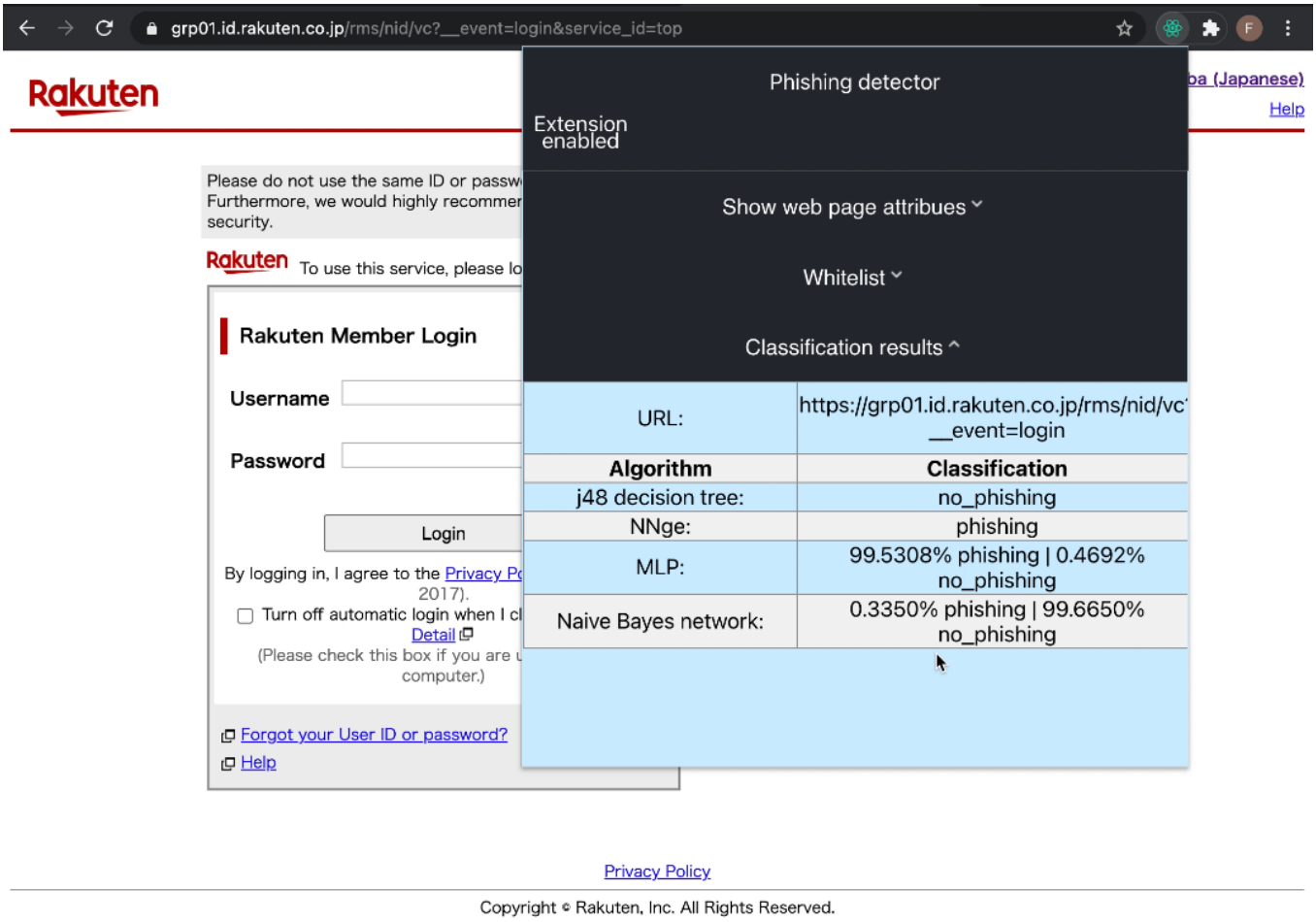


Figure 1. A screenshot of the developed browser extension for monitoring visited webpages.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

III. COLLECTED WEBPAGE DATA

A group of websites collected with the browser extension for data collection introduced in Section II is described here. This group contains data about 21755 webpages described by 29 attributes whose values were saved in the backend and value legitimate or phishing was assigned to each website on the basis of an inspection. The attributes had been inspired by [1]. Suppose a set $\boldsymbol{W}$ which represents all collected webpages is defined. In other words, $\boldsymbol{W}$ is a set of known instances. The cardinality of $\boldsymbol{W}$ is 21755. Suppose each webpage $\boldsymbol{w} \in \boldsymbol{W}$ is described by 29 attributes $A_k \in \boldsymbol{A} = \{A_1; \ldots; A_k; \ldots; A_{29}\}$. Symbol $A_k(\boldsymbol{w}) = a$ is employed for any numerical attribute A_k whose value is a for webpage $\boldsymbol{w}$. Symbol $A_k = \boldsymbol{P}$ for any numerical attribute A_k where $\boldsymbol{P}$ is a set of numbers represents possible numerical values of A_k. Symbol $A_k(\boldsymbol{w}) = a_{k,l}$ is employed for any categorical attribute A_k whose value is $a_{k,l}$ for webpage $\boldsymbol{w}$. Possible categorical values $a_{k,1}, \ldots, a_{k,l}, \ldots a_{k,l_k}$ for attribute A_k are represented by symbol $A_k = \{a_{k,1}; \ldots; a_{k,l}; \ldots; a_{k,l_k}\}$. The assigned value legitimate or phishing for each webpage $\boldsymbol{w} \in \boldsymbol{W}$ is represented by $D(\boldsymbol{w})$. Symbol d_1 is employed for a categorical value meaning legitimate and d_2 is for phishing. Symbol $D = \{d_1; d_2\}$ means d_1, d_2 are possible values for D. A description of the collected data is in Table I.

TABLE I. DESCRIPTION OF COLLECTED DATA

Particular Attribute	Type	Possible Values	Units
BaseURLLengthWithParams (A_1)	Numerical	4, 5, 6, …	count
AtSymbolInURL (A_2)	Categorical	*absent* ($a_{2,1}$)	N/A
		present ($a_{2,2}$)	
DashInDomain (A_3)	Categorical	*absent* ($a_{3,1}$)	N/A
		present ($a_{3,2}$)	
DoubleSlashInURL (A_4)	Categorical	*absent* ($a_{4,1}$)	N/A
		present ($a_{4,2}$)	
DNSRecord (A_5)	Categorical	*absent* ($a_{5,1}$)	N/A
		present ($a_{5,2}$)	
PercentageOfExternalPictures (A_6)	Numerical	[0; 100]	%
HostInURL (A_7)	Categorical	*absent* ($a_{7,1}$)	N/A
		present ($a_{7,2}$)	
HTTPSInDomain (A_8)	Categorical	*absent* ($a_{8,1}$)	N/A
		present ($a_{8,2}$)	
IFrame (A_9)	Categorical	*absent* ($a_{9,1}$)	N/A
		present ($a_{9,2}$)	
NonStandardPort (A_{10})	Categorical	*absent* ($a_{10,1}$)	N/A
		present ($a_{10,2}$)	
TextFieldInPopUpWindow (A_{11})	Categorical	*absent* ($a_{11,1}$)	N/A
		present ($a_{11,2}$)	
ShorteningService (A_{12})	Categorical	*absent* ($a_{12,1}$)	N/A
		present ($a_{12,2}$)	
SubmittingToMail (A_{13})	Categorical	*absent* ($a_{13,1}$)	N/A
		present ($a_{13,2}$)	
HTTPSProtocol (A_{14})	Categorical	*absent* ($a_{14,1}$)	N/A
		present ($a_{14,2}$)	
ServerFormHandler (A_{15})	Categorical	*absent* ($a_{15,1}$)	N/A
		present ($a_{15,2}$)	
IndexationByGoogle (A_{16})	Categorical	*absent* ($a_{16,1}$)	N/A
		present ($a_{16,2}$)	
WebsiteRanking (A_{17})	Numerical	[0; 10]	Open Page Rank
AllowOnMouseOver (A_{18})	Categorical	*absent* ($a_{18,1}$)	N/A
		present ($a_{18,2}$)	
AllowRightClick (A_{19})	Categorical	*absent* ($a_{19,1}$)	N/A
		present ($a_{19,2}$)	
IPAddress (A_{20})	Categorical	*absent* ($a_{20,1}$)	N/A
		present ($a_{20,2}$)	
NumberOfLinksPointingToPage (A_{21})	Numerical	0, 1, 2, …	count
NumberOfSubDomains (A_{22})	Numerical	0, 1, 2, …	count
NumberOfRedirections (A_{23})	Numerical	0, 1, 2, …	count
DomainAge (A_{24})	Numerical	0, 1, 2, …	day
PercentageOfExternalURLsInATags (A_{25})	Numerical	[0; 100]	%
PercentageOfExternalOrNoURLsInATags (A_{26})	Numerical	[0; 100]	%
PercentageOfExternalURLsInMetaLinkScriptTags (A_{27})	Numerical	[0; 100]	%
TimeToDomainExpiration (A_{28})	Numerical	0, 1, 2, …	day
WebTraffic (A_{29})	Numerical	0, 1, 2, …	Alexa Rank
Class (D)	Categorial	*legitimate* (d_1)	N/A
		phishing (d_2)	

Attribute *BaseURLLengthWithParams* (A_1) is related to the URL and represents the sum of the number of characters in the domain and the number of characters in the parameters after ?. $A_1 = \{4, 5, 6, \ldots\}$. Attribute $A_2 = AtSymbolInURL = \{a_{2,1}; a_{2,2}\} = \{absent; present\}$ gives information if the URL address of a webpage $\boldsymbol{w}$ contains the @ symbol. If it does not contain this

symbol, $A_2(\boldsymbol{w})$ = *absent*. Otherwise, $A_2(\boldsymbol{w})$ = *present*. The other attributes can be read from Table I in a similar way. In addition to the description in Table I, the collected data has been analyzed and the summary of this analysis is in Table II.

TABLE II. WEBPAGE DATA ANALYSIS

Attribute	Possible Values	Frequency	Median	Mode
A_1	4, 5, 6, …	N/A	22	21
A_2	*absent* ($a_{2,1}$)	21555	N/A	*absent*
	present ($a_{2,2}$)	200		
A_3	*absent* ($a_{3,1}$)	18708	N/A	*absent*
	present ($a_{3,2}$)	3047		
A_4	*absent* ($a_{4,1}$)	21701	N/A	*absent*
	present ($a_{4,2}$)	54		
A_5	*absent* ($a_{5,1}$)	4393	N/A	*present*
	present ($a_{5,2}$)	17362		
A_6	[0; 100]	N/A	0	0
A_7	*absent* ($a_{7,1}$)	21729	N/A	*absent*
	present ($a_{7,2}$)	26		
A_8	*absent* ($a_{8,1}$)	9	N/A	*present*
	present ($a_{8,2}$)	21746		
A_9	*absent* ($a_{9,1}$)	17894	N/A	*absent*
	present ($a_{9,2}$)	3861		
A_{10}	*absent* ($a_{10,1}$)	21752	N/A	*absent*
	present ($a_{10,2}$)	3		
A_{11}	*absent* ($a_{11,1}$)	20780	N/A	*absent*
	present ($a_{11,2}$)	975		
A_{12}	*absent* ($a_{12,1}$)	19532	N/A	*absent*
	present ($a_{12,2}$)	2223		
A_{13}	*absent* ($a_{13,1}$)	19004	N/A	*absent*
	present ($a_{13,2}$)	2751		
A_{14}	*absent* ($a_{14,1}$)	7491	N/A	*present*
	present ($a_{14,2}$)	14264		
A_{15}	*absent* ($a_{15,1}$)	17797	N/A	*absent*
	present ($a_{15,2}$)	3958		
A_{16}	*absent* ($a_{16,1}$)	8225	N/A	*present*
	present ($a_{16,2}$)	13530		
A_{17}	[0; 10]	N/A	0.76	0
A_{18}	*absent* ($a_{18,1}$)	15774	N/A	*absent*
	present ($a_{18,2}$)	5981		
A_{19}	*absent* ($a_{19,1}$)	21582	N/A	*absent*
	present ($a_{19,2}$)	173		
A_{20}	*absent* ($a_{20,1}$)	21690	N/A	*absent*
	present ($a_{20,2}$)	65		
A_{21}	0, 1, 2, …	N/A	0	0
A_{22}	0, 1, 2, …	N/A	0	0
A_{23}	0, 1, 2, …	N/A	0	0
A_{24}	0, 1, 2, …	N/A	0	0
A_{25}	[0; 100]	N/A	54	0
A_{26}	[0; 100]	N/A	58	100
A_{27}	[0; 100]	N/A	92	100
A_{28}	0, 1, 2, …	N/A	0	0
A_{29}	0, 1, 2, …	N/A	0	0
D	*legitimate* (d_1)	17662	N/A	*legitimate*
	phishing (d_2)	4093		

IV. EXPERIMENTAL EVALUTATION

The data mining models used in the browser extension for monitoring visited webpages from Section II and created with the collected webpage data from Section III are evaluated here. The models are created with Weka, which is an open-source data mining tool that is widely used for research and industrial applications [15]. An example of a data mining model produced by Weka is shown in Fig. 2. Normally, there are a model created on the basis of all data and some statistics related to its validation. Due to a low percentage of phishing webpages within the total number of webpages in the real world, 10-fold cross-validation is employed. This is similar to the medical field where 10-fold cross-validation is typical and where only a small percentage of patients within the whole universe has a particular medical issue. 10-fold cross-validation partitions the webpage data into ten subsets with equal numbers of cases with legitimate and phishing webpages. Nine subsets are used for the creation of a model and the model is then validated on the remaining subset, which is possible thanks to the availability of the values for D. This is repeated ten times for the purposes of using each subset for validation exactly once. The statistics from Weka is recomputed into sensitivity, specificity, accuracy and criterion. TP Rate for phishing in Weka is equal to sensitivity which expresses the ability to correctly identify phishing webpages. FP Rate for phishing in Weka is used for the computation of specificity as 1 – FP Rate. Specificity expresses the ability to correctly identify legitimate webpages. Ideally, sensitivity equals one and specificity equals one. When sensitivity is too low, the browser extension for monitoring visited webpages does not show any warnings for many dangerous phishing webpages. This would make the use of the extension meaningless. When specificity is too low, the extension shows many warnings even if the visited webpages are legitimate. This would make the use of the extension bothersome, especially when the high number of existing legitimate webpages is considered. Therefore, specificity should be very close to one. Both sensitivity and specificity are important and so the criterion defined as the average of sensitivity and specificity is used. Correctly classified instances in Weka correspond to accuracy, which is not very determinative in unbalanced situations.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

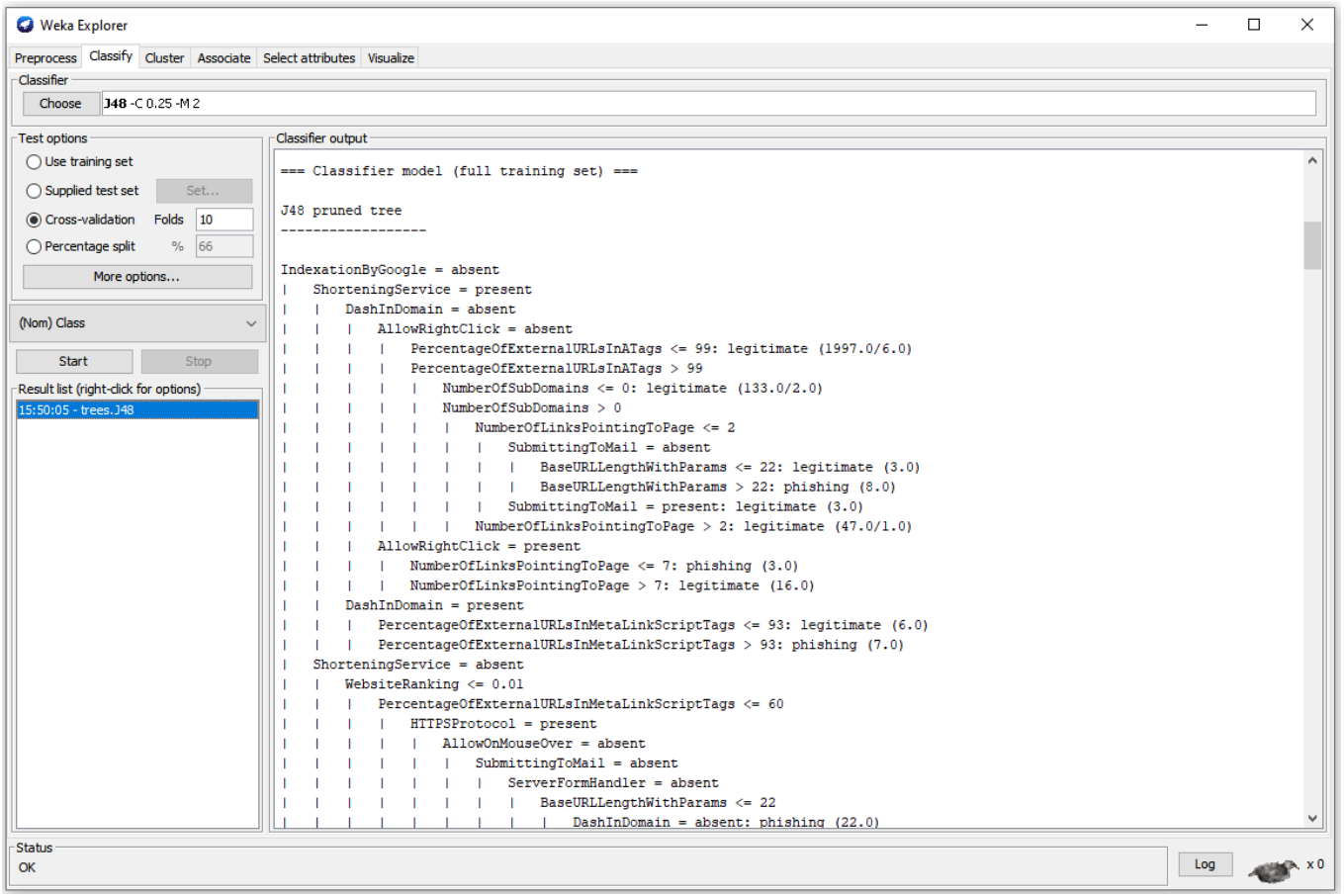


Figure 2. A screenshot of the creation of a data mining model in Weka.

The results of the evaluation are summarized in Table III where the rows correspond to the data mining models supported in the browser extension for monitoring visited webpages and the columns are related to the particular recomputed measures. BayesNet is a Bayesian network model implemented in Weka as a Java class with the same name and displayed in Fig. 1 as Naive Bayes network. J48 is a decision tree model from Weka which is shown in Fig. 1 as j48 decision tree. MultilayerPerception is a neural network model from Weka whose name is MLP in Fig. 1. NNge is a nearest neighbor classifier. The best value of the criterion was achieved by NNge with 0.97350. Its specificity 0.99292 is also very close to one, which means that the use of the extension should not be bothersome. Very similar values were achieved by J48 as well. BayesNet has the worst values.

TABLE III. RESULTS IN 10-FOLD CROSS-VALIDATION

Model	Measure			
	Sensitivity	*Specificity*	*Accuracy*	*Criterion*
BayesNet	0.84266	0.95006	0.92986	0.89636
J48	0.93599	0.99156	0.98111	0.96378
MultilayerPerceptron	0.88712	0.98800	0.96902	0.93757
NNge	0.95407	0.99292	0.98561	0.97350

V. CONCLUSIONS

Webpage data was collected for the development of data mining models classifying webpages into legitimate or phishing. The data has 21755 instances corresponding to webpages described by 29 attributes. In addition, two browser extensions and a backend were developed for the purposes of phishing webpages detection and data collection. Four data mining models such as a Bayesian network, decision tree, nearest neighbor classifier and neural network were created with the collected data and included for monitoring visited webpages. 10-fold cross-validation showed promising results with the best combination of sensitivity and specificity equaling to 0.95407 and 0.99292. Future work may include further development of models for an even more accurate classification of webpages.

ACKNOWLEDGMENT

This publication was realized with support of Operational Program Integrated Infrastructure 2014 - 2020 of the project: Intelligent operating and processing systems for UAVs, code ITMS 313011V422, co-financed by the European Regional Development Fund.

EUROPEAN UNION
European Regional Development Fund
OP Integrated Infrastructure 2014 – 2020

MINISTRY
OF TRANSPORT
AND CONSTRUCTION
OF THE SLOVAK REPUBLIC

The author would like to thank student Filip Glemba for working on software implementations under the supervision of the author as a part of the Master studies of Filip Glemba.

REFERENCES

[1] N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41, no. 13, pp. 5948-5959, 2014.

[2] M. Alwateer, S. W. Loke, N. Fernando, "Enabling drone services: Drone crowdsourcing and drone scripting," IEEE Access, vol. 7, art. no. 110035, 2019.

[3] J. Bohacik, I. Skula, M. Zabovsky, "Data mining-based phishing detection," in Federated Conference on Computer Science and Information Systems, 2020, pp. 27-30.

[4] D. Cvitanic, "Drone applications in transportation," in International Conference on Smart and Sustainable Technologies, 2020, art. no. 20133277.

[5] C. L. Evans, "Clone Zone is an easy tool for building fake websites," available at https://www.vice.com/en/article/3dkyyb/clone-zone-is-an-easy-tool-for-building-fake-websites, 2015.

[6] N. N. Gana; S. M. Abdulhamid, "Machine learning classification algorithms for phishing detection: A comparative appraisal and analysis," in International Conference of the IEEE Nigeria Computer Chapter,

[7] A. Y. Husodo, G. Jati, N. Alfiany, W. Jatmiko, "Intruder drone localization based on 2D image and area expansion principle for supporting military defence system," in IEEE International Conference on Communication, Networks and Satellite, 2019, pp. 35-40.

[8] S. Mahmoud, N. Mohamed, J. Al-Jaroodi, "Integrating UAVs into the cloud using the concept of the Web of Things," Journal of Robotics, vol. 2015, art. no. 631420.

[9] S. Paliath, M. A. Qbeitah, M. Aldwairi, "PhishOut: Effective phishing detection using selected features," in International Conference on Telecommunications, 2020, art. no. 20135977.

[10] S. Shukla, P. Sharma, "Detection of phishing URL using Bayesian optimized SVM classifier," in International Conference on Electronics, Communication and Aerospace Technology, 2020, pp. 1385-1389.

[11] A. J. Van der Merwe, M. Loock, M. Dabrowski, M, "Characteristics and responsibilities involved in a phishing attack," in Winter International Symposium on Information and Communication Technologies, 2005, pp. 249-254.

[12] S. Nair, G. Rodrigues, C. Dsouza, S. Bellary, V. Gonsalves, "Designing of beach rescue drone using GPS And Zigbee technologies," in International Conference on Communication and Electronics Systems, 2019, pp. 1154-1158.

[13] S. Patil, S. Dhage, "A methodical overview on phishing detection along with an organized way to construct an anti-phishing framework," in International Conference on Advanced Computing & Communication Systems, 2019, pp. 588-593

[14] R. Thomazella, J. E. Castanho, F.R.L. Dotto, O.P. Rodrigues Junior; G. H. Rosa; A. N. Marana; J. P. Papa, "Environmental monitoring using drone images and convolutional neural networks," in IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 8941-8944.

[15] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques (4th edition). USA: Morgan Kaufman, 2016.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Rapid bibliometric analysis in deep learning domain

Ravil I. Mukhamediev, Ilyas Assanov, Marina Yelis
Satbayev University (KazNRTU)
Almaty, Kazakhstan
ravil.muhamedyev@gmail.com

Adilkhan Symagulov, Yan Kuchin, Kirill Yakunin
Institute of Information and Computational Technologies
Almaty, Kazakhstan
ykuchin@mail.ru

Aubakirov Margulan
Maharishi International University
Fairfield, Iowa

Laila Tabynbayeva
Kazakh Research Institute of Agriculture and Crop Production
Almaty, Kazakhstan

Peter Sedlacek
University of Zilina
Zilina,Slovakia

***Abstract*—The paper systematizes the Deep Learning domain and calculates the dynamics of changes in the number of scientific articles according to Google Scholar. The method of data acquisition and calculation of dynamic indicators of changes in publication activity is described: speed (D1) and acceleration of growth (D2) of scientific publications. Analysis of publication activity, in particular, showed a high interest in modern transformer models, the development of datasets for some industries, and a sharp increase in interest in methods of explicable machine learning. Relatively small research domains are receiving increasing attention, as evidenced by the negative correlation between the number of articles and D1 and D2 scores. The results show that, despite the limitations of the method, it is possible to identify fast-growing areas of research regardless of the number of articles. The paper presents result for more than 400 search queries related to classified research areas. Calculation results and software can be downloaded https://www.dropbox.com/sh/fkfw3a1hkf0suvc/AACRZ7v9qympen_ht00jeiF6a?dl=0.**

Keywords—machine learning; deep learning; explainable machine learning; transfer learning; convolution neural networks; recurrent neural networks; transformers; explainable machine learning; bibliometric indicators

I. INTRODUCTION

The evolution of each scientific domain is accompanied by an increase or decrease of the interest of researchers, which is reflected in the change of bibliometric indicators. The latter includes the number of publications, the citation index, the number of co-authors, the Hirsch index and others. The identification of "hot" areas in which these indicators are more important allows us to understand better the situation in science and, if possible, to concentrate the efforts on breakthrough areas.

The field of Deep Learning (DL) is characterized by a wide range of methods and tasks, some of which already have acceptable solutions implemented in the form of software, while others require the intensive research.

The number of publications in many areas of DL is growing. Therefore, a simple statement of the increase in the number of publications is not enough. In this connection, bibliometric indicators (BI), such as the number of publications, the citation index, the number of co-authors, etc., are widely used for the evaluation of research domains and organizations [1-4]. BI is used for the assessment of policy making in the field of scientific research [5], the impact of publications databases [6]. Some authors use bibliometric data to build prediction models [7] in some cases in combination with patent analysis [8, 9]. At the same time, bibliometric methods have significant limitations. In particular, numerical indexes are non-linearly dependent on the size of the country and organization [10]. Various authors introduce new or modify former indicators [11, 12].

In order to identify the logic of changes in publication activity the differential indicators are implemented in [13]. Their application allows to estimate the speed and acceleration of changes in bibliometric indicators.

In this paper, the number of publications of articles with selected key terms are considered as analyzed indicators. The differential metrics allow to evaluate the dynamics of changes in the usage of selected key terms by the authors of scientific publications, what indirectly indicates the growth or decrease of the interest of researchers in the scientific field designated by this term.

We have made a brief review and systematization of scientific directions included in machine learning and DL.

Objectives of the study:

1. Systematization of DL domains according to literature data;

2. Development of methods for collecting data from open sources and assessing changes in publication activity using differential indicators;

3. Assessment of changes in publication activity in DL using differential indicators to identify fast-growing and "fading" research domains.

II. LITERATURE REVIEW

According to the definitions given in [14] Deep Learning (DL) is a subset of Machine learning (ML) that provides computation for multilayer Neural networks (NN). Typical DL architectures are deep neural networks (DNN), convolutional neural networks (CNN), recurrent neural networks (RNN), generating adversarial networks (GAN), and more.

An artificial neural network with more than one hidden layer is considered to be a deep neural network. A network with several hidden layers less than two is considered a shallow neural network.

The advantage of deep neural networks is evident when processing large amounts of data. The quality of traditional algorithms, reaching a certain limit, no longer increases with the amount of available data. At the same time, deep neural networks can extract the features that provide the solution to the problem, so that the more data, the more subtle dependencies can be used by the neural network to improve the quality of the solution (Fig. 1 [15]).

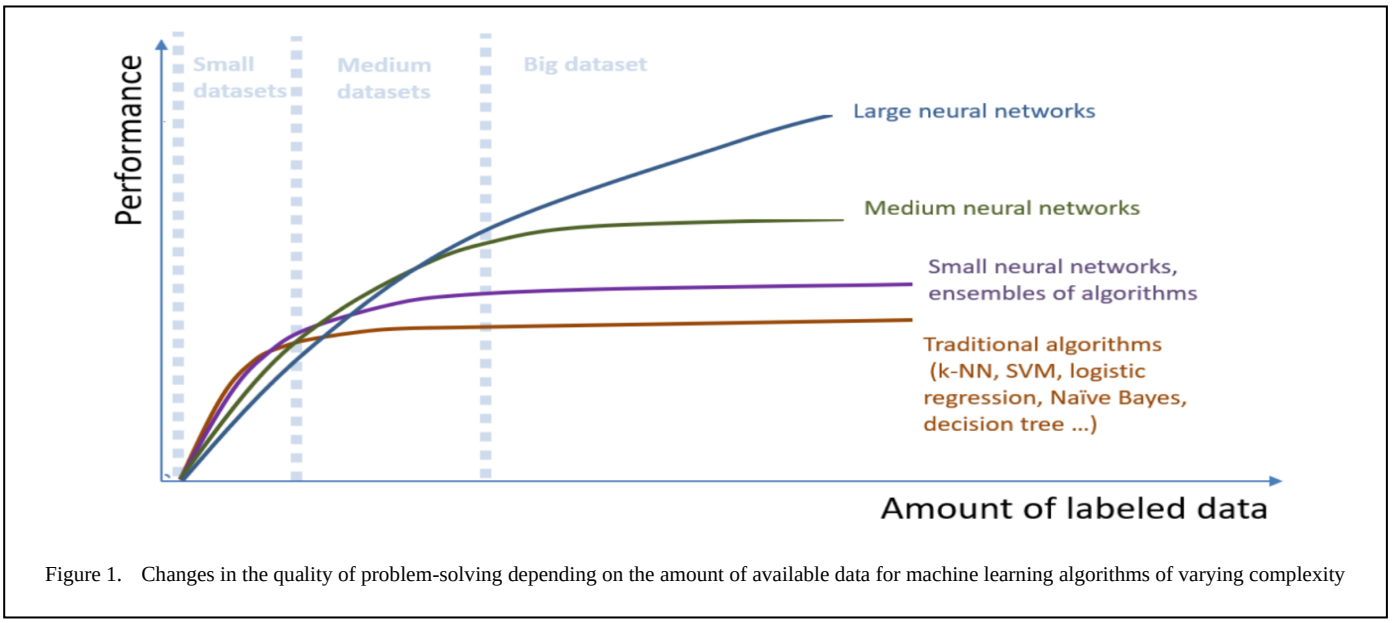


Figure 1. Changes in the quality of problem-solving depending on the amount of available data for machine learning algorithms of varying complexity

The more data, the more accurate the network will be. This phenomenon of deep neural networks predetermined their success in solving the problems of classification and regression.

The variety of neural network architectures can be reduced to four basic architectures (Fig. 2):

1. Standard feed forward neural network - NN;
2. Recurrent neural network – RNN;
3. Convolution neural network- CNN;
4. Hybrid architectures that include elements of 1,2,3 basic architectures, such as Siamese networks and transformers.

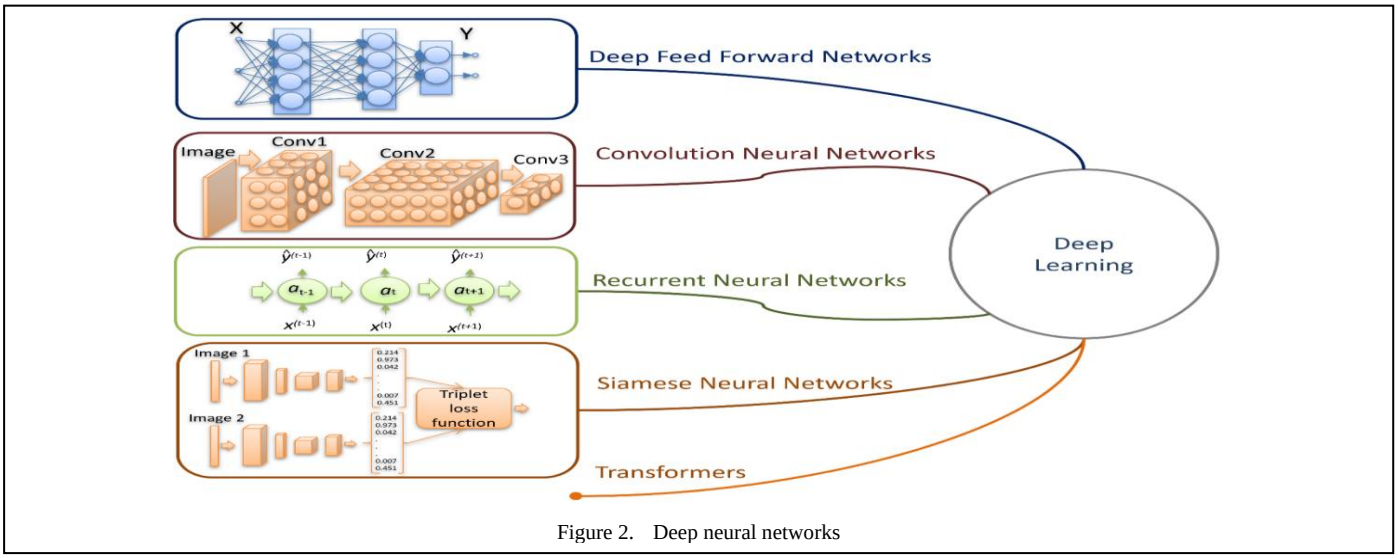


Figure 2. Deep neural networks

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. ИРН AP08856412).

This work was partially supported in part by the Slovak Research and Development Agency under Grant APVV-18-0027 “New methods development for reliability analysis of complex system”.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

A recurrent neural network (RNN) is used in complex classification problems when the result depends on the sequence of input signals or data. And the length of such a sequence is generally not fixed. The data and signals received at previous processing steps are stored in one or another form in the internal state of the network, which allows taking their influence into account in the general result. Examples of tasks with such sequences are Machine translation [16, 17]), when a translated word may depend on the context, i.e., previous or next words of the text:

- Speech recognition [18, 19]), where the values of the phonemes depend on their combination;
- DNA Analysis [20], in which the nucleotide sequence determines the meaning of the gene;
- Classifications of the emotional coloring of the text or tone (sentiment analysis [21]). The tone of the text is determined not only by specific words, but also by their combinations;
- Name entity recognition [22]), that is, proper names, days of the week and months, locations, dates, etc.

Another example of the application of recurrent networks are tasks where a relatively small sequence of input data causes the generation of long sequences of data or signals, for example:

- Music generation [23], when the generated musical work can only be specified in terms of style;
- Text generation [24]), etc.

Convolutional neural network. In the early days of the computer vision development, researchers made efforts to teach the computer to highlight characteristic areas of an image. Kalman, Sobel, Laplace, and other filters were widely used. Manual adjustment of the algorithm for the extraction of the characteristic properties of images allowed to achieve good results in particular cases, for example, when the images of faces were standardized in size and quality of photographs. However, when the foreshortening, illumination, and scale of images were changed, the quality of recognition deteriorated sharply. Convolutional neural networks have largely overcome this problem.

Hybrid architectures. The cv2 problem is often solved using Siamese networks [25] (Fig. 2), where two images are processed by two identical pre-trained networks. The obtained results (image vectors) are compared using a triplet loss function, which can be implemented as a triplet distance embedding [26] or a triplet probabilistic embedding [27].

The triplet loss function "increases" the distance between embeddings of images of different objects and decreases the distance between different embeddings of the same object.

BERT (Bidirectional Encoder Representations from Transformers) [28], ELMO [29], GPT (Generative Pre-Trained Transformer), Generative adversarial networks [30], have recently gained great popularity and are effectively used in natural language processing tasks.

In addition to the systematization of DL sections, their evolution is also of interest. Let us assess the dynamics of changes in the number of scientific publications aimed at the development of individual scientific domains and overcoming the aforementioned technological limitations of DL.

III. METHOD

Publication activity demonstrates the interest of researchers in scientific sections, which are briefly described by some set of terms. Obviously, new and promising in the eyes of the scientific community thematic sections are characterized by an increased publication activity. To identify such sections and their comparative evaluation in the field of DL, we will use the method described in [31]. The paper proposes dynamic indicators that allow to numerically estimate the growth rate of the number of articles and acceleration. The indicators allow us to estimate the scientific field without regard to its volume, which is important for new fast-growing domains that do not yet have a large volume of publications.

The dynamic indicators (D1 - speed and D2- acceleration) of some j-th bibliometric indicator $s_j^{(db,k)}$ time t_n can be calculated as follows:

$$D1_{s_j}^{(db,k)}(t_n) = w1_j \times \frac{ds_j^{(db,k)}(t_n)}{dt}, \quad (1)$$

$$D2_{s_j}^{(db,k)}(t_n) = w2_j \times \frac{d(ds_j^{(db,k)}(t_n)/dt)}{dt}, \quad (2)$$

where k is the search term in database db, $w1_j$ and $w2_j$ are some empirical coefficients that regulate the "weight" of the $s_j^{(db,k)}$.

In our case $s_j^{(db,k)}(t_n)$ - number of articles in t_n - year selected using the search query k in the Google Scholar database. Weights $w1_j$, $w2_j$ are taken as 1. For example, the search query k = "Deep+Learning Bidirectional+Encoder+Representations+from+Transformers" gives the following annual publication volumes 13, 692, 1970 in 2018, 2019 and 2020 respectively. Bibliometric databases often give an estimate of the number of publications at the end of the year. Approximation, the obtained numerical series, is done with the help of polynomial regression model. Increasing the order of the regression n allows us to obtain a high value of the coefficient of determination (r2_score), but usually leads to overtraining of the model. In order to avoid overtraining and to ensure a sufficient degree of generalization the following rules of thumb are used:

- For each search query the regression order is chosen individually, starting from n=2 to ensure r2_score >=0.7. As soon as the specified boundary is reached, n is fixed and the selection process stops;

• Since we are most interested in the values of dynamic indicators for the last year, we used the last value (number of publications for 2020) and the value equal to half of the growth of articles achieved at the end of 2020, which we conventionally associate with the middle of the year, as a test set on which r2_score is determined.

Data processing complex for calculation of D1 and D2 indicators includes scraper, preprocessing, regression calculation and calculation of D1 and D2 indicators (Fig. 3).

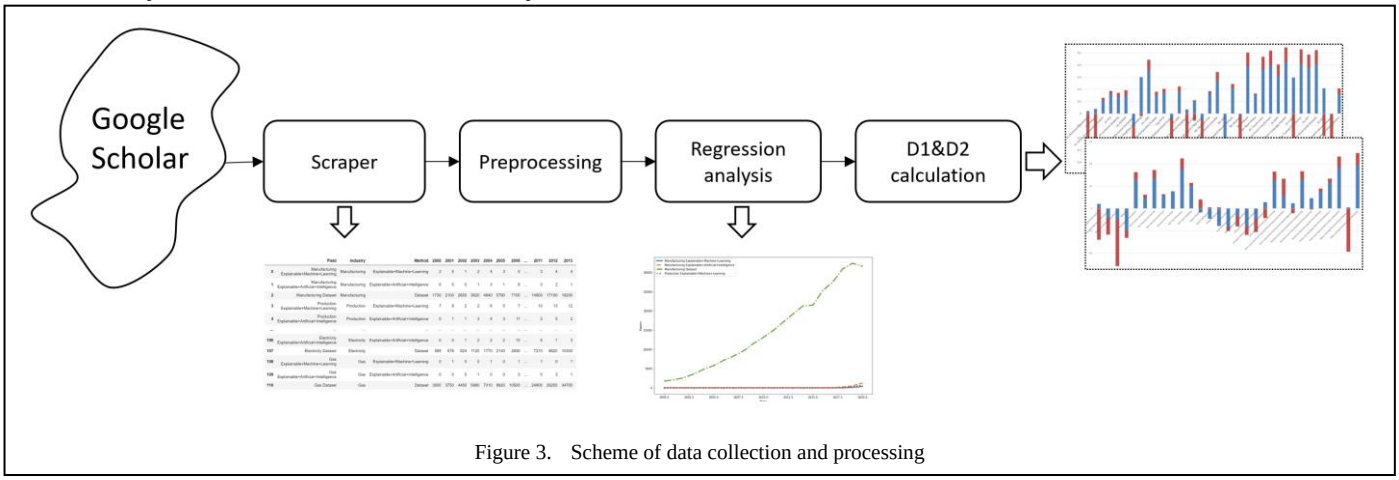


Figure 3. Scheme of data collection and processing

The scraper uses the requests library to retrieve the Html page of each search query and then uses the Beautifulsoup library to select the necessary information from it. Since scholar.google.com has protection against robots, the app.proxiesapi.com service was used to provide a proxy server. Thanks to this it was possible to avoid captcha.

Pre-processing consists in the formation of a dataframe containing only the necessary information (search query and the numeric series of the annual number of publications). No additional processing is required.

As a result, a series of indicators D1 and D2 are formed. The indicators calculated for the last year reflect the dynamics of changes in the interest of the scientific community in the relevant scientific sections at the time of the study.

IV. Results and Discussion

The described set of article counts and calculation of D1 and D2 indicators has been applied to assess the dynamic indicators of the main models of machine learning and deep learning (Fig. 4), publication activity related to explanatory AI (Fig.5).

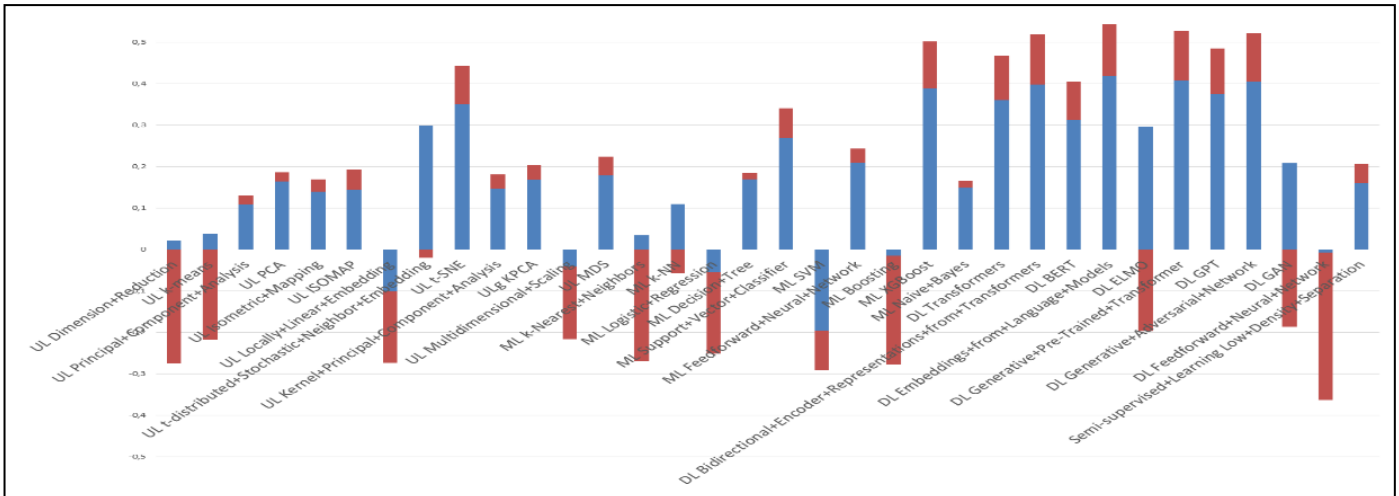


Figure 4. Using popular models of unsupervised learning (UL), machine learning (ML), deep learning (DL). Indicators D1(blue) and D2(red)

The interpretation of the indicator values is as follows:

• When both values (D1, D2) are positive, it indicates an accelerated growth of the number of articles in this domain;

• When D1 is negative and D2 is positive, this indicates a slowdown in the number of articles;

• When D1 is positive and D2 is negative, it indicates a slowdown in the growth of the number of articles;

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

• When both values are negative, it indicates an accelerated decrease in the number of articles.

Figure 4 shows a significant increase in publication activity in the deep learning domain, which is to be expected. Figure 5 shows a significant and accelerating increase in publication activity in the domain of explainable machine learning and transformers applications.

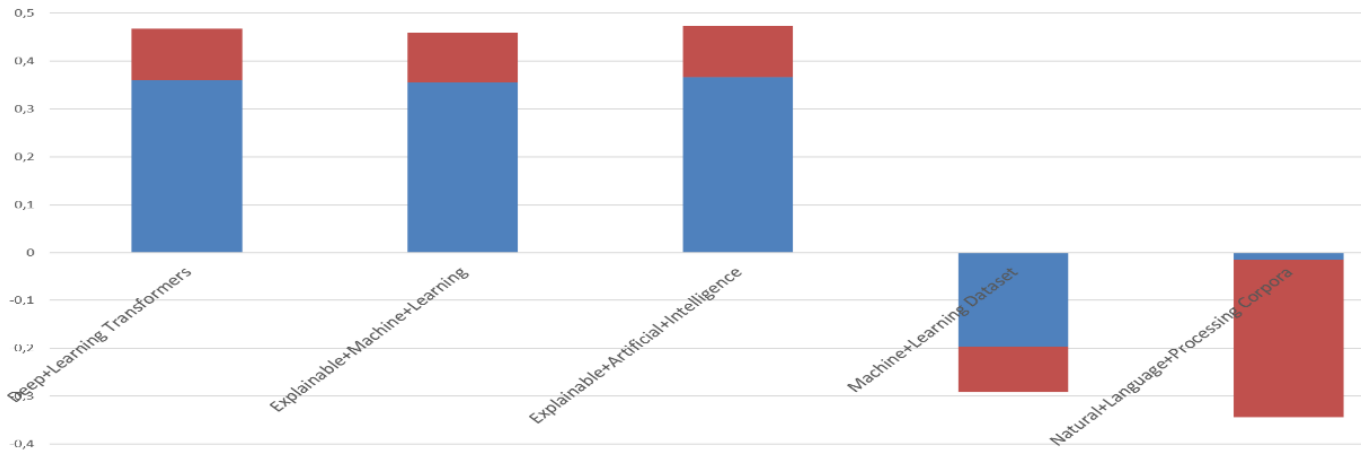


Figure 5. The D1(blue) and D2(red) indicators of researcher interest in explanatory machine learning methods and the development of datasets and corpora of texts

Explaining the results of machine learning models is a serious problem, preventing widespread use of AI in healthcare [32], banking, and many other fields [33]. A complex machine learning model is a "black" box, hiding the mechanism for obtaining results. To turn it into a "white" or "gray" box methods are used to estimate the influence of input parameters on the final result: Treeinterpreter, DeepLIFT, Local Interpretable Model-agnostic (LIME) [34], SHapley Additive exPlanations (SHAP) [35]. However, interpretation of the influence of individual model parameters is possible if they have a clear meaning.

The negative correlation between the number of articles and the indicators D1, D2, r2_score (Fig. 6) allows us to conclude that the domains with a large number of publications are characterized by a decrease in the dynamics of publication activity and model error.

Despite the simplicity of the model, it can also be used for forecasting. For example, by training a regression model of machine learning, we were able to predict the number of publications one year ahead and obtained an average error of about 6% with a standard deviation of 7%.

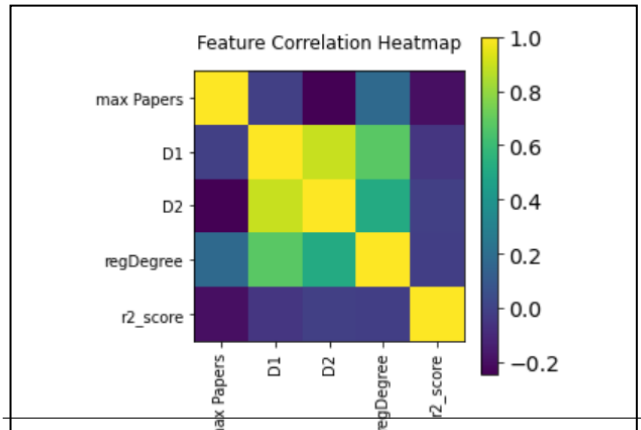


Figure 6. Correlation heatmap of bibliometric indicators

V. Conclusions

In this paper, we systematize the sections of DL and evaluated the dynamics of changes in the number of scientific articles according to Google Scholar. The results show that, firstly, it is possible to identify fast-growing and "fading" research domains for any "reasonable" number of articles (>100), and secondly, the prediction of publication activity is possible in the short term with sufficient accuracy for practice.

The method we used has some limitations, in particular:

1. For all the depth of the informal analysis, the set of terms is still set by the researcher. We also cannot guarantee the exhaustive completeness and consistency of the empirical review performed;

2. This analysis does not take into account the fact that the importance of a particular scientific topic is determined not only by the number of articles, but also by the volume of citations, the "weight" of the individual characteristics of the authors, the quality of the journals, and so on;

3. The method does not evaluate term change processes and semantic proximity of scientific domains.

Nevertheless, the obtained estimates, despite some limitations of the applied approach, correspond to empirical observations related to the growth of applications of deep learning models. The analysis shows that the efforts of the scientific community are aimed at overcoming the technological limitations of ML.

Future research is planned to focus on evaluating the use of deep learning models in economic sectors and also using thematic modeling to cluster research areas.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The program and results of calculations can be downloaded at https://www.dropbox.com/sh/fkfw3a1hkf0suvc/AACRZ7v9qympen_ht00jeiF6a?dl=0.

ACKNOWLEDGMENT

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. ИРН AP08856412).

This work was partially supported in part by the Slovak Research and Development Agency under Grant APVV-18-0027 “New methods development for reliability analysis of complex system”.

REFERENCES

[1] É. Gauthier, "Bibliometric analysis of scientific and technological research: a user's guide to the methodology," ed: Science and Technology Redesign Project, Statistics Canada Canada, 1998.

[2] A. Van Raan, "The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments," TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis, vol. 12, no. 1, pp. 20-29, 2003.

[3] G. Abramo, C. D’Angelo, and F. Pugini, "The measurement of Italian universities’ research productivity by a non parametric-bibliometric methodology," Scientometrics, vol. 76, no. 2, pp. 225-244, 2008.

[4] J. Mokhnacheva and I. Mitroshin, "Nanoscience and nanotechnologies at the Moscow domain: a bibliometric analysis based on Web of Science (Thomson Reuters)," Information resources of Russia, vol. 6, pp. 17-23, 2014.

[5] K. Debackere and W. Glänzel, "Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key," Scientometrics, vol. 59, no. 2, pp. 253-276, 2004.

[6] H. F. Moed, "The effect of “open access” on citation impact: An analysis of ArXiv's condensed matter section," Journal of the American Society for Information Science and Technology, vol. 58, no. 13, pp. 2047-2054, 2007.

[7] T. U. Daim, G. R. Rueda, and H. T. Martin, "Technology forecasting using bibliometric analysis and system dynamics," in A Unifying Discipline for Melting the Boundaries Technology Management:, 2005, pp. 112-122: IEEE.

[8] T. U. Daim, G. Rueda, H. Martin, and P. Gerdsri, "Forecasting emerging technologies: Use of bibliometrics and patent analysis," Technological forecasting and social change, vol. 73, no. 8, pp. 981-1012, 2006.

[9] T. Inaba and M. Squicciarini, "ICT: A new taxonomy based on the international patent classification," 2017.

[10] J. S. Katz, "Scale-independent indicators and research evaluation," Science and Public Policy, vol. 27, no. 1, pp. 23-36, 2000.

[11] L. Egghe, "Dynamic h ‐ index: The Hirsch index in function of time," Journal of the American Society for Information Science and Technology, vol. 58, no. 3, pp. 452-454, 2007.

[12] M. Levene, T. Fenner, and J. Bar-Ilan, "A bibliometric index based on the complete list of cited publications," arXiv preprint arXiv:1304.6945, 2013.

[13] R. I. Muhamedyev, R. M. Aliguliyev, Z. M. Shokishalov, and R. R. Mustakayev, "New bibliometric indicators for prospectivity estimation of research fields," 2018.

[14] G. Nguyen et al., "Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey," Artificial Intelligence Review, vol. 52, no. 1, pp. 77-124, 2019.

[15] A. Ng, K. Katanforoosh, Y. Bensouda. *Neural network and deep learning*. Accessed on: May 16, 2021. [Online]. Available: https://www.coursera.org/learn/neural-networks-deep-learning?specialization=deep-learning.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[17] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.

[18] A. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," arXiv preprint arXiv:1412.5567, 2014.

[19] D. Jurafsky, Speech & language processing. Pearson Education India, 2000.

[20] X. Liu, "Deep recurrent neural network for protein function prediction from sequence," arXiv preprint arXiv:1701.08318, 2017.

[21] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1253, 2018.

[22] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.

[23] A. Nayebi and M. Vitelli, "Gruv: Algorithmic music generation using recurrent neural networks," Course CS224D: Deep Learning for Natural Language Processing (Stanford), 2015.

[24] S. Lu, Y. Zhu, W. Zhang, J. Wang, and Y. Yu, "Neural text generation: Past, present and beyond," arXiv preprint arXiv:1803.07133, 2018.

[25] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701-1708.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815-823.

[27] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in 2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS), 2016, pp. 1-8: IEEE.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[29] M. E. Peters et al., "Deep contextualized word representations," arXiv preprint arXiv:1802.05365, 2018.

[30] I. J. Goodfellow et al., "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.

[31] R. I. Muhamedyev, R. M. Aliguliyev, Z. M. Shokishalov, and R. R. Mustakayev, "New bibliometric indicators for prospectivity estimation of research fields," 2018.

[32] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," Future healthcare journal, vol. 6, no. 2, p. 94, 2019.

[33] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135-1144.

[35] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," arXiv preprint arXiv:1705.07874, 2017.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Reliability analysis of Cognitive Radio Networks with balking and reneging

Mohamed Hedi Zaghouani
Doctoral School of informatics
University of Debrecen
Debrecen,Hungary
zaghouani.hedi@inf.unideb.hu

Hamza Nemouchi
Doctoral School of informatics
University of Debrecen
Debrecen,Hungary
nemouchi.hamza@inf.unideb.hu

János Sztrik
Doctoral School of informatics
University of Debrecen
Debrecen,Hungary
sztrik.janos@inf.unideb.hu

***Abstract*—The principle of balking and reneging on Cognitive Radio Networks taking into consideration servers unreliability, is investigated in this paper. In today's networking environment, the concepts of balking and reneging are very common. The more crowded the system is, the more discouraged potential customers will be, on the other hand, once their total waiting time hits the maximum, impatient users exit the system. To be closer to real-life scenarios, servers unreliability is taken into consideration.**

***Index Terms*—Finite source queuing systems, simulation, Cognitive Radio Networks, performance and reliability measures, non-reliable servers, balking and reneging.**

I. Introduction

Customers' impatience is a crucial factor to consider while modeling call centers. Balking and reneging are two frequent ways for customers to express their impatience. A call-in customer who cannot be helped immediately by a human server might be told how long a wait he/she faces before an operator is available. The customer may then choose to hang up (i.e. balk) or wait. This is known as balking, which occurs when a customer refuses to enter a queue because it is too long. A consumer waiting for an operator, on the other hand, may hang up (i.e. renege) before being served if the line becomes too lengthy. This is the reneging behavior. We will be applying these feature on our system.

The primary aim of our "Cognitive Radio Network" model is to use the free spaces in the primary frequency band to support the secondary one. More details can be found in [**1**], [**2**], [**3**], [**4**], [**5**], and [**6**].

The network is made up of two main components. The first one is designed for Primary Users (PU) with a finite number of sources who produce primary calls after an exponentially distributed time. To be served, all of the created calls will be placed in a FIFO queue. The service time is distributed exponentially. The second subsystem is dedicated to the jobs of Secondary Users (SU), which are created from a limited number of sources and directed to the Secondary Channel Service (SCS) for processing. These calls have an exponential arrival time, but their service time is generally distributed using hypo-exponential, hyper-exponential, and gamma distributions.

The created licensed calls will check the availability of the Primary Channel Service (PCS), if it is available, the service will begin immediately, if it is already in use by a primary call, the later call will be joining the FIFO queue. If the PCS is taken by a secondary customer, service will be interrupted immediately and the customer in question will be led back to the SCS. The aborted call would be restarted from the beginning of its operation or added to the retrial queue (orbit), depending on the current situation of the secondary service unit.

Secondary Channel Service SCS, on the other hand, receives requests that are not licensed. If the intended server is idle, SU is permitted to start the service, if it is occupied, they will attempt to start their service in the PCS in an opportunistic manner. If the last service channel is not occupied, the low priority call will be able to begin the service, otherwise, if it is occupied, the call will be immediately added to the orbit. Based on an exponential distributed time, concerned calls will retry to enter the server.

Several researchers have examined the CRN based on different scenarios. The effect of server unreliability on the CRN, for example, was examined by the authors of [**4**], [**11**] and [**12**], however, balking and reneging were not taken in consideration by anyone of them. In [**7**], the same model was used taking into consideration abandonment, with SUs being required to quit the network when their cumulative patience time reaches a predetermined limit.

Balking and reneging have been applied on different retrial queuing systems, in [**8**], this theory was studied on the M/M/S retrial queue, and in [**9**] was added to an M/M/1/N retrial queue.

After a thorough search of many relevant topics and papers, we were unable to find any investigations that discussed this model in the case of balking and reneging assuming that the secondary model's server is unreliable, which is the novelty of our paper.

II. System model

The queuing cognitive radio system shown in Fig1 is based on the following assumptions. Consider two interconnected subsystems whose primary requests are created from a N_1 finite number of sources and forwarded to the first server after an exponentially dispersed period with a mean value of $1/\lambda_1$. If the unit is available, the service will begin, otherwise, the call will be placed in the FIFO queue. The principal user's

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

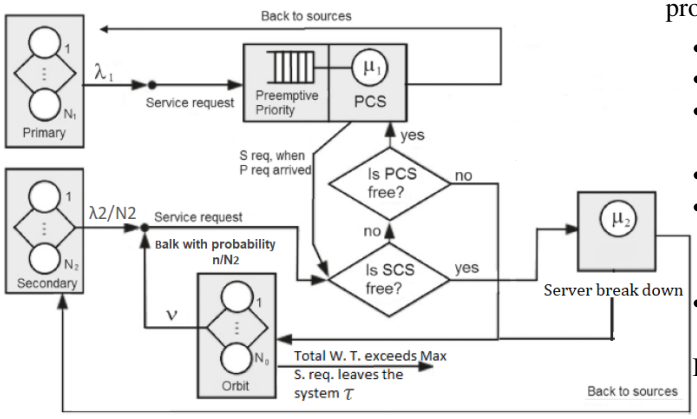


Figure 1. System model.

service time is an exponentially distributed random variable with the parameter μ_1.

N_2 represents the secondary subsystem's number of sources. After an exponentially distributed interval of time, each source produces low priority jobs with parameter λ_2/N_2. The service time of SUs is generally distributed using hypo-exponential, gamma and hyper-exponential distributions, all of which have the same mean and different variances, with a rate μ_2. The retrial time of the secondary customer is considered to be a random variable with an exponential distribution and a parameter of ν.

With a probability of n/N_2, new arriving secondary customers may balk (refuse to join the server), where n is the number of users in the network and N_2 is the number of sources. They may renege from the system (leave the orbit after joining) if service does not start by a given random time, which is exponentially distributed with parameter tau. Some breakdowns and repairs will appear unpredictably at the secondary server level based on exponentially distributed times with parameters γ_1 and γ_2, respectively. All the input parameters are collected in I.

Table I
INPUT PARAMETERS OF THE SIMULATION

Parameters	The maximum	Value at time t
Active licensed sources	N_1	$k_1(t)$
Active unlicensed sources	N_2	$k_2(t)$
High priority arrival parameter		λ_1
Low priority arrival parameter		λ_2/N_2
Jobs in FIFO queue	N_1 - 1	$q(t)$
Jobs in orbit	N_2 - 1	$o(t)$
Primary service parameter		μ_1
Secondary service rate		μ_2
Reneging parameter		τ
Secondary server breakdown parameter		γ_1
Secondary server repair parameter		γ_2
Retrial rate		ν

We suppose the below notations to model a stochastic process defining the functionality of the system:

- $k1(t)$: Number of primary jobs at given time t;
- $k2(t)$: Number of secondary jobs at given time t;
- $q(t)$: The number of high-priority calls in the queue at time t;
- $o(t)$: Number of jobs in the orbit at time t;
- $y(t) = 0$, if the PCS is idle, $Y(t) = 1$, if PCS is occupied by a high-priority request and $Y(t) = 2$, if the PCS is busy with a low-priority request at time t;
- $c(t) = 0$, if SCS is free and $c(t) = 1$, if SCS is busy at time t.

It is easy to see that:

$$k_1(n) = \begin{cases} N_1 - q(t), & y(t) = 0,2 \\ N_1 - q(t) - 1 & y(t) = 1 \end{cases}$$

$$k_2(n) = \begin{cases} N_2 - o(t) - c(t), & y(t) = 0,1 \\ N_2 - o(t) - c(t) - 1 & y(t) = 2 \end{cases}$$

III. SIMULATION RESULTS

The impact of service time distributions and the cognitive technology on our system's main performance measures are investigated in this section. SimPack [**10**] was used to create a stochastic simulation program in the C programming language. Except for the secondary service rate, all the network's random variables are considered to be exponentially distributed. All the numerical results were obtained by the validation of the simulation outputs. II lists the numerical values of the simulation main class input parameters, while III lists the numerical values of the simulation program's statistical class.

Table II
THE INPUT PARAMETERS OF SIMULATION

N_1	N_2	λ_1	λ_2/N_2	μ_1	μ_2	ν	τ	γ_1	γ_2
25	50	0.2	x-axis	2	2	0.2	0.2	0.1	0.2

Table III
PARAMETERS OF THE GENERAL DISTRIBUTIONS

Distribution	Gamma, c_x^2 <1	Hyper	Hypo	Gamma, c_x^2 >1
Parameters	α = 1,7857 β = 1,7857	p = 0,3309 λ_1 = 0,66198 λ_2 = 1,33803	λ_1 = 1,4854 λ_2 = 3,06	α = 0,3906 β = 0,3906
Mean	1	1	1	1
Variance	0.56	2.56	0.56	2.56
c_x^2	0.56	2.56	0.56	2.56

1) Service times are generally distributed: Figure 2 shows how the distribution of primary and secondary service times affects the mean residence time of SUs vs the generation of secondary request times. When service times are gamma distributed with a $C_x^2 > 1$, a significant distributions sensitivity can be observed. The same clear sensitivity is seen in Figure 3, where the effect of primary and secondary service times distribution on the mean reneging time of SUs vs secondary request time generation was displayed, while gamma with a c_x^2 greater than one. Figure 4 confirms the

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

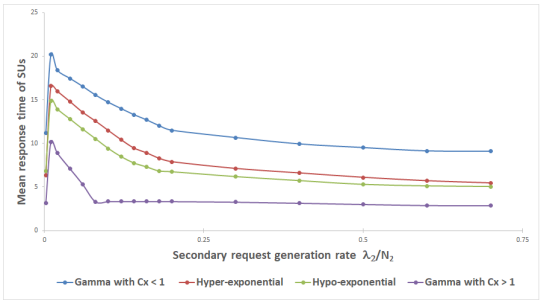


Figure 2. The impact of primary and secondary service times distribution on the mean residence time of SUs vs secondary request time generation

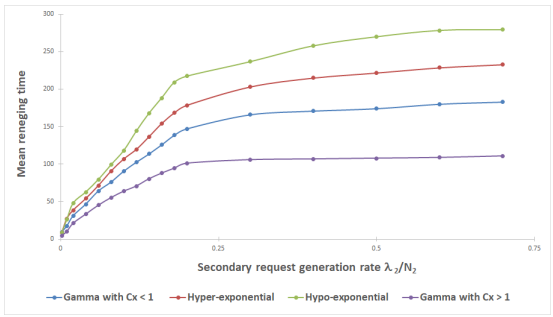


Figure 3. The impact of primary and secondary service times distribution on the mean reneging time of SUs vs secondary request time generation

similar behaviour shown in the previous two figures. Using the gamma distribution, which has a $C_x^2 > 1$. Furthermore, as expected, increasing the SU arrival intensity leads to more balking rate, we note as well a significant number of customers balk from the system.

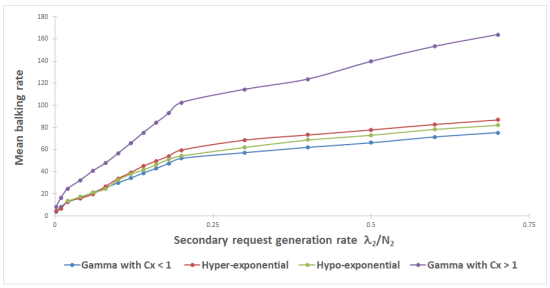


Figure 4. The effect of primary and secondary service times distribution on the mean balking rate vs secondary request time generation

2) All inter-event time are exponentially distributed: In this section, we suppose that all inter-event times are exponentially distributed. Same value of the parameters shown in I are applied with $\lambda_2 = 0.5$. We'd like to look into the effect of conductivity on the system's properties.

Figure 5 depicts the effect of increasing N_2 on the primary arrival rate and the number of sources on the mean response time of cognitive customers. In this graph, we can see that the primary arrival intensity has a significant impact on the average residency time of SUs, as when $\lambda_1 = \lambda_2/2$, the results show a smaller value of the mean than when $\lambda_1 = \lambda_2 * 2$.

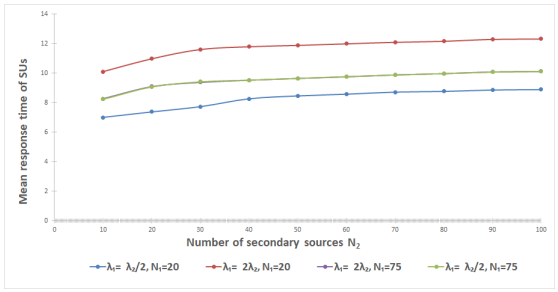


Figure 5. The effect of the primary network parameters on the average response time of SUs vs N_2

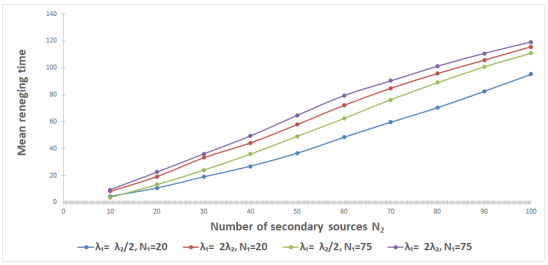


Figure 6. The effect of the primary network parameters on the mean reneging time of SUs vs $N2$

In contrast to the primary number of sources, which has no effect, as when N_2 is high the traffic intensity in the primary sub-subsystem is bigger. However, other effects can be shown in Figure 6 when the primary number of sources is bigger. This is due to more secondary customers are reneging from the system.

Figure 7 illustrates the impact of N_2 on the mean balking rate within different configurations in the primary network. The only effect that can be seen is when the primary inter-arrival parameter is half the secondary arrival intensity. In this case, fewer customers do not enter the system. However, increasing λ_1 or N_2 there is almost no effect on the mean balking rate. As expected, increasing N_2 involves a greater mean balking rate.

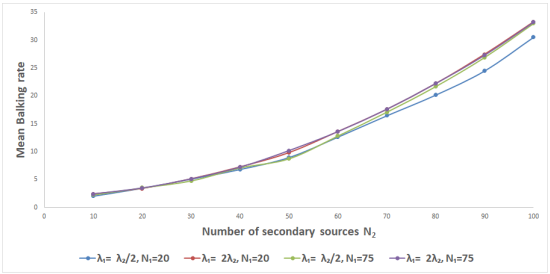


Figure 7. The effect of the primary network parameters on the mean balking rate of SUs vs N_2

Figure 8 depicts the effect of the secondary server's unreliability on the expected response time while the mean failure rate $(1/\gamma_1)$ is increasing. Increasing the secondary mean failure rate, as expected, results in a longer response time for the affected customers. Figure 9 depicts the effect of the server

unit's unreliability on the SUs' mean sojourn time. It shows the difference in the value of the mean response time as the mean repair rate increases ($1/\gamma_2$). As seen in this graph, having a reliable servers in CRN leads to a lower response time.

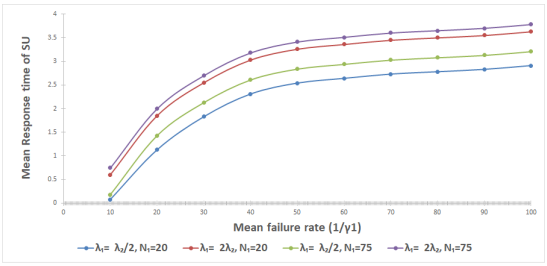


Figure 8. The influence of the non-reliability of the SCS on the secondary mean response time vs mean failure rate $1/\gamma_1$

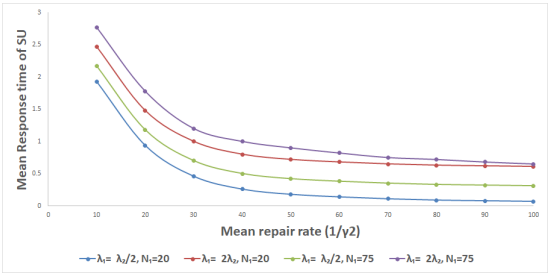


Figure 9. The impact of the non-reliability of the SCS on the mean sojourn time of the SUs vs mean repair rate $1/\gamma_2$

IV. Conclusion

A finite-source retrial queuing system that contains two nonindependent parts was introduced in this paper. Our system was designed to simulate a cognitive radio network with primary and secondary service units with balking and reneging on the second part assuming that the later server is unreliable. A thorough review was carried out using simulation to investigate the effect of the service times distributions and the impact of the cognitive technology on the key performance measures of the system.

Acknowledgment

The research work of János Sztrik is supported by the EFOP-3.6.1-16-2016-00022 project. The project is co-financed by the European Union and the European Social Fund. The research of Mohamed Hedi Zaghouani is supported by the scholarship of Stipendium Hungaricum.

References

[1] Wang L. C. Adachi F. Load-balancing spectrum decision for cognitive radio networks: IEEE Journal on Selected Areas in Communications. 2011. pp. 757–769.

[2] Devroye N., Vu M., Tarokh V. Cognitive radio networks: CIEEE Signal Processing Magazine 25.6. 2008. pp. 12–23.

[3] Gunawardena S., Zhuang W. Modeling and Analysis of Voice and Data in Cognitive Radio Networks: Springer. 2014.

[4] Nemouchi H., Sztrik J. Performance Simulation of Finite-Source Cognitive Radio Networks with Servers Subjects to Breakdowns and Repairs: Journal of Mathematical Sciences. 2019. pp. 702–711.

[5] Akyildiz I. F., Lee W. Y., Vuran M. C., Mohanty S. NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey: Computer networks 50.13. 2006. pp. 2127–2159.

[6] Mitola J., Maguire G. Q. Cognitive radio: making software radios more personal: IEEE personal communications 6.4. 1999. pp. 13–18.

[7] Zaghouani M. H., Sztrik J. Performance simulation of finite-source Cognitive Radio Networks with impatient calls in the orbit: XXXVI International Seminar on Stability Problems for Stochastic Models. 2020.

[8] Wuchner P., Sztrik J., De Meer H. Finite-source M/M/S retrial queue with search for balking and impatient customers from the orbit: Computer Networks. 2009. 53.8. 1264–1273.

[9] Kumar R., Som B. K. An M/M/1/N queuing system with reverse balking and reverse reneging: Advanced Modeling and Optimization. 2014. 339–353.

[10] Fishwick, P. A. Simpack: getting started with simulation programming in C and C++: Proceedings of the 24th conference on Winter simulation. 1992. pp. 154–162.

[11] Zaghouani, M. H., Sztrik, J., and Uka A. Reliability Analysis of Cognitive Radio Networks: Annales Mathematicae et Informaticae. 2020. 52. pp. 255-265.

[12] Zaghouani, M. H., Sztrik, J., and Melikov A. Reliability Analysis of Cognitive Radio Networks: International Conference on Information and Digital Technologies (IDT). IEEE (2019). 557-562.

[13] Zaghouani, M. H., Sztrik, J., and Uka A. Simulation of the performance of Cognitive Radio Networks with unreliable servers. Annales Mathematicae et Informaticae. 2020. 52. pp. 255-265.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Experience with the usage of virtual reality worlds about natural history in Slovenia

Tomaz Amon
Center for scientific visualization
Bioanim Slovenia
Tomaz.amon@bioanim.com

***Abstract*— The virtual reality worlds on the web obviously present in many parameters a better tool for education as classical paper learning software. In its extreme, the electronic paper textbook can be included as a pdf file. The interactivity and ease to transport, copy and adapt such material make it very practical. In addition it is more durable than paper because it can be easily reproduced. We produced a software package "Cell-Tissue-Body" teaching biology in the secondary and primary schools in Slovenia. The Ministry of education supported this project and it is free to obtain. We discuss here first the advantages and disadvantages of the electronical textbooks, then the difficulties when implementing them in our schools and finally our thoughts how to overcome the possible difficulties in future updates of this and other projects.**

Keywords—virtual reality; visualization; natural history; biology

I. INTRODUCTION

With the development of computers and Internet the electronical (web) interactive "textbook" has become increasingly important and popular. However, the educational material in the electronic form is still something new and unexplored for an average biology teacher in our primary and secondary schools. On the other hand, the pupils grow together with the modern informational technology and the web is for them the first place to look for new games and if enough motivated, also for (additional) educational help. Therefore we try to make the educational software attractive enough so that the pupils demand it from a teacher and he sees that this brings an obvious step forward in the learning process.

Ten years ago we started producing animations of biological structures and processes because we thought that the classical paper textbooks alone couldn't teach modern biology effectively enough. We soon proceeded with animations in virtual space and as VRML evolved we started to produce interactive web3D virtual worlds explaining the structure and function of the living cell. The cell is small and cannot be studied without special instruments e.g. optical and electron microscope. It contains a complicated network of membraneous organelles, which are difficult to show and explain only with the help of classical images. So we constructed a simplified cell nucleus, endoplasmic reticulum, Golgi body and the outer cell membrane in order to show the production of proteins, which are finally exported, out of the

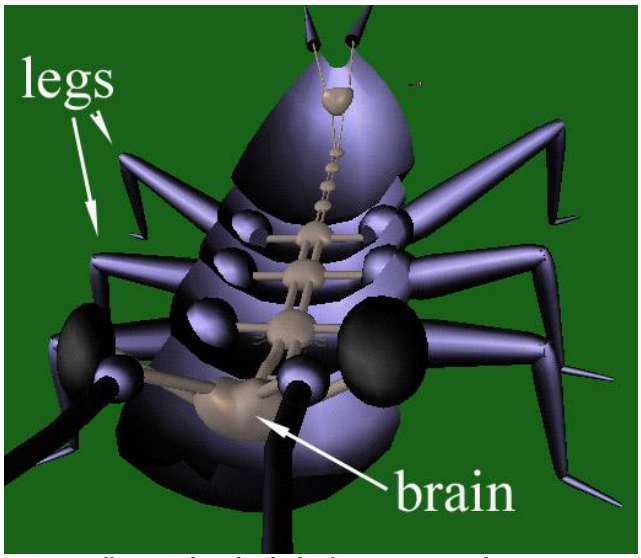


Figure 1. Partially opened cricket body showing its central nervous system. When the user's avatar moves away the animal its body closes again. The cricket can lift its wings and start stridulation

cell like in the gland cells. The story starts so that the messenger RNA, modeled as a VRML extrusion node, emerges out of the nucleus through the nuclear pore. In the cytoplasm the ribosomes start to form on the mRNA creating the so-called polysomes – the units that produce protein. All this happens on the surface of the endoplasmic reticulum and the newly synthesized gets injected into the endoplasmic reticulum. Then the proteins are packed into the membraneous vesicles and transported out of the endoplasmic reticulum into the Golgi body where they are modified and finally transported out of the cell. As you read this story, it would not be strange if it sounded complicated to understand. The story tells about a fundamental process in the living cell and is told in millions of schools worldwide. Typically it is accompanied by illustrations in textbooks in order to make the understanding easier. These illustrations are often painted so that they represent the three dimensional space in which the processes take place. So it is natural to get one step further and to enrich the classical textbook with the virtual reality worlds where the student can travel by himself through the cell and learn about it. Since the web becomes now the first-to-try, if not already the most important source for gathering information. Several years ago there emerged the VRML and we saw it as a great opportunity for us to produce new educational material in 3D. In order to be

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

displayed in real time, the VRML worlds (and now Unity [4] 3D worlds) are not composed in such a detail as photo realistic 3D worlds shown e.g. in the cinema. This does not make any problem, since one has to apply some degree of simplification to educational illustrations in order to point out what is really important to understand.

In such a philosophy the software package "Cell-Tissue-Body" [1] was created and it became one of the official textbooks for the study of biology in our secondary schools. Since the Ministry of education supported it, the software (in Slovene language) is free for all our teachers and students. In principle they can study the biology now more effectively and faster, because they have a new tool which helps them to learn and try exams interactively and to imagine the biological structures in (virtual) space.

II. Virtual Reality and Teaching

The experiments with the living organisms are one of the fundamentals in biology teaching. When the teacher performs such experiments in the class he can show much more if he uses also the advantages modern visualization techniques. For example, an animal introduced to the class is often small and cannot be seen well by all the pupils at the same time. So we bring to the class a video projector, connect it to the video camera and display on a large screen the animal under study – a field cricket in this case. No computer is involved so far and we show the outer structure of the animal so that all can see it. In the next step we put the cricket in the box and connect the computer to the projector. We continue with the web3D virtual reality model of the cricket (fig.1) showing the main features of its hearing organs and the nervous system. With the computer animation and simulation in this interactive VRML world it is easy to show how the hearing organ of the cricket in its front legs is excited by the sound waves and this excitation further processed in the central nervous in order to extract the frequency (4kHz) and the spatial pattern characteristic for the cricket song (fig.2). These parameters are crucial for the survival of the cricket species since the male and female crickets find themselves in the grass by sound communication. Finally we offer printable material (e.g. in pdf format) and the questions to test the knowledge.

III. Students as Authors of the VR Worlds

Another way to promote computer aided education in the class is that the students themselves create web3D worlds. Of course it would be too hard for them to learn instantly advanced modeling and computer animation, but they are ready to insert simple models made by themselves into the already prepared templates. A very attractive way is the technique which one typically uses when making virtual landscapes. A picture is converted into the 3D object simply by elevating those points which appear brighter on the image and on the so created elevation grid one finally superimposes the original picture as the texture. So one can very attractively visualize the images of the cell ultra structure as seen under the electron microscope or even the flowers (fig.3). All what the student needs to do is to apply a filter smoothing the image and so

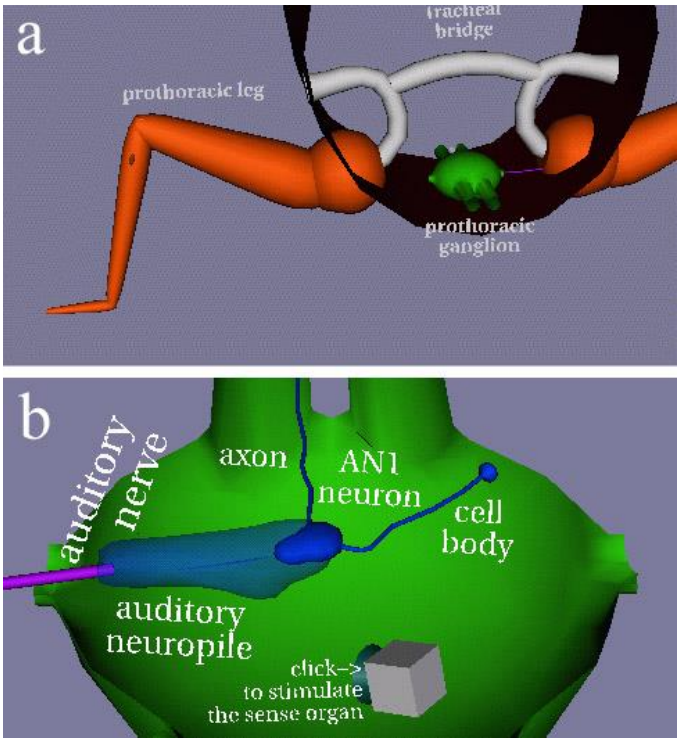


Figure 2. a The VRML model of the first pair of cricket legs containing the hearing organs. The prothoracic ganglion, which plays an important role in the processing of the acoustical information, is also shown. b. The prothoracic ganglion with the functional models of some of the acoustical neurons (here shown AN1) enabling simple simulation of the processing of acoustical information.

eliminate the sharp disturbing peaks. Then he converts the image to the VRML elevation grid with one of the programs that are generating 3D terrain like the Terragen [2].

We also used this technique to produce the geo-VRML models of our vicinity. We produced together with the Museum of Natural History, Ljubljana [3] a 3D representation of the marsh called "Barje" near the city Ljubljana (fig.4). This marsh is the living place of many animal and plant species and is therefore a candidate to become a nature reserve. The terrain modeling has been done so that at first we made the curves that connected the points of the same altitude in 100m steps. We filled these "teases" with different levels of gray and then applied a Gauss filter to smooth the steps between the teases. The product was astonishingly good enough representation of the terrain. In the next step this image was converted to the VRML elevation grid and finally to indexed face set with appropriately reduced number of polygons. The model was small enough to be manipulated when observed on an ordinary web page (e.g. http://www.bioanim.com/arty/) or Unity3D [4] environment. On the model of the terrain was overlaid the texture - simplified map of the Barje with rivers, streets, settlements etc., but no text. Textual explanations were implemented so that we placed over important places on the model transparent cubes. When the user went over such a transparent cube with his mouse, the name of the cube (=the appropriate legend of that geographical location) appeared besides the mouse pointer (see the legend "Bevke" on fig.4). In a separate frame a photographic panorama of an interesting location with the textual explanation is also shown.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

IV. Computer aided education example in physiological experiments on the university

One of the first software products of our lab was the program for evaluating the data acquired in the course of experiments studying the behavior of the nerve cells (electrophysiological data). We decided to write such a program since at that time (1985) there was still no adequate commercially available software for similar purposes and we also felt that the evaluation of nerve recordings is a specific process which needs dedicated software that can be adapted to our needs. The program has been written in C, periodically updated and is still in use - possible only if one has the source code [2]. The original idea was to write software turning the computer into a sophisticated kind of a chart recorder. Up to now, the program has grown to a powerful data evaluation tool enabling effective evaluation of long data sequences. Although the program is menu driven and the data can be edited with the mouse, its main strength lies in a series of macro commands, which enable effective evaluation of long data sequences. So typically first the right data evaluation path is interactively found on a short data sample. The "recipe" for the evaluation is stored as a sequence of macro commands that drive the batch procedure that processes all the remaining data. The new functions can be easily added and the program can be adapted to a new machine. In spite of the fact that there exists a rich palette of commercially available software for the data processing in electrophysiology one sometimes still misses just the effective software routine for the solution of a specific problem in data evaluation or in representing the data in a new way like the three dimensional VRML "terrain" imaging of nervous activity. Such difficulties can be easily overcome if one can accordingly change the source code.

The key task in the analysis of electrophysiological data evaluation is to find the characteristic oscillations of amplitude in time. They are mainly composed of the so called spikes and postsynaptic potentials, both conveying the information to and from the nerve cells. Our program offers several functions for finding them, the most used is the so called transient finding function (fig.5.). This function detects neuronal spikes, postsynaptic potentials and other transients. Let's say for example that the transients we are interested in are the neuronal spikes. The spike rises from the noise level upwards with a certain slope, has a peak and decreases to the base line in a certain time. The transient function uses all these characteristics as parameters in order to determine the oscillations we are interested in. The search process starts so that it makes use of a line of given length and steepness (fig.5), which "surfs" on the data trace. When the right end of this surfing line intersects with the data curve, this is the sign for a potentially interesting transient on the data curve. Now all additional required parameters are checked like the amplitude and time duration of the transient and if they meet the given criteria, the transient gets accepted.

When evaluating the nervous activity we routinely find the different types of transients with the above-described transient finding function and the identified transients are "cut" out of the data curve and laid on a rectangular data matrix. Here every

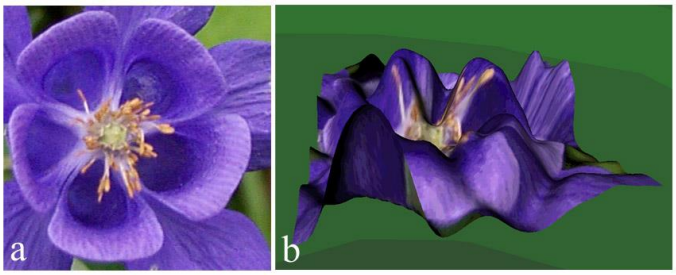


Figure 3. a. The picture of the flower Aquilegia sp. b. The 3D "terrain" produced from the flower slide shown on the left only by removing color, applying a smoothing filter and converting the image to the VRML elevation grid on which the original image is placed as a texture. Such "flower terrains" are easy to produce and often enhance learning motivation in students.

Figure 4. Screen shot of the introductory web page (in slovene language) about the marsh called "Barje" that lies in the vicinity of the city Ljubljana is an important biotope that should be preserved. We created a web3D presentation of the terrain with "hot spots" like the one of the village called "Bevke", shown in the central frame of the picture. The 3D representation is enhanced with the panoramic presentations of interesting topics (shown in the right frame).

matrix cell contains numbers telling how many times a data point ran over that matrix cell. The cells of the matrix so contain the cumulative numbers of the data points superimposed with their coordinates (fig.6). The result is the grayscale picture (fig.6a), which has black pixels where no data point passed over and white pixels where most data points were superimposed. So one obtains something like a 3D histogram: x axis (or the width of the picture) is time, y axis (height) is the amplitude (e.g. in milivolts) and the z-axis is the count number of the superimposed data points shown as the optical density of the picture. So one sees the distributions of the transient classes and can check how well the transient finding function (described above) did its task. Such a picture can be associated with a color palette to produce false colors and it can be edited with an image editing program where the student marks e.g. with a black paintbrush those spikes which do not belong to the neuronal response, but instead to the noise which has not been filtered out by the computer. This manually modified picture is processed again by the data evaluating program which eliminates from the original data trace the transients that have been marked on the image. Alternatively, the image can be transformed into a VRML elevation grid and observed as a 3D object (fig.6b). This makes easier distinguishing different types of nervous activity.

V. How we design our content

We found that the best way to go is to start with nicely looking web pages containing dynamic presentations of the content. The page shows at first only the most important

features of the matter under and is dynamically updated as the user interacts (e.g. with the mouse) with it. So a classical textbook illustration is typically split into several structural subunits (“layers”), dynamically shown and combined as necessary. Such an image still reminds us to the classical textbook illustration, but the information given to the student this way is clearer and easier to understand. The virtual reality world is added when necessary or available. As the speed of rendering of the VR world slows with the increased size of its window, we keep the VR window initially relatively small (about 500 pixels wide). In this first step we restrict the VR world manipulation just to the examination of the object under study. It is good to eliminate the need for a browser plugin here like displaying with Ar-Ty tool [5] or in the Unity3D environment [4]. They can be safely included into the ordinary web pages and it often motivates the user to get further interested in the VR world.

As the user enters a web 3D world, he observes objects from different perspectives, moves around as he wishes and investigates it by changing the parameters of the processes happening in the virtual world. In the first glance this seems to be the ideal solution, but the computers in a classroom are often slow and the VR world does not instantly respond to the pupil's actions. It is also necessary to restrict the movements in the VR world to a certain number of "guided paths" which lead the student in an optimal way through the points of interest. Otherwise one can easily get lost in the VR world. In praxis this leads to the situation where pupils in the class wildly spin around the objects what is by itself very interesting and stimulative way of edutainment, it only needs to be understood also by the teacher in this way.

VI. Conclusion

Internet is the library of the future. Some years ago we went to a classical library to find the printed material we needed, now we start a web search engine, type the keywords and get many hits which we have to study like we studied books some time ago. As the web is larger than any classical library, we are confronted with far more of information to digest as it was the case before. Therefore good web sites offer an effective way of getting information like dynamic illustrations and web3D worlds telling with relatively little text much about the structure and function of living organisms. The duty of educators is therefore also to prepare pupils for electronical ways of acquiring knowledge.The primary and secondary schools in Slovenia have enough platforms and software to be able to run smoothly web or offline education. But many biology teachers don’t go this way because they are not familiar enough with the computer-aided education or they are confronted with logistical problems in the school like the availability of computers for biology teaching etc. As we were discussing with our local publisher the creation of the new textbook for biology for the primary school, I mentioned the importance of the interactive, mainly web based educational material as an addition to the classical paper textbook. The author and the editor said that this was very nice, very important, but the electronical way of education

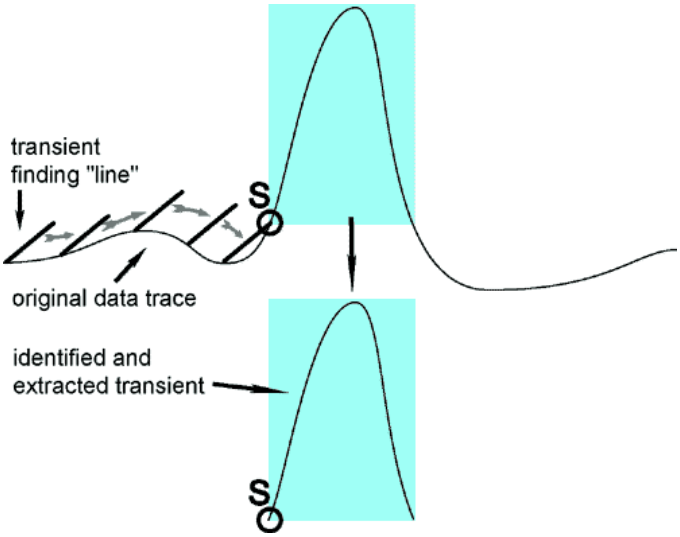


Figure 5. Description of the transient (spike) searching strategy used by the transient finding function. Every data point of the original data trace is scanned with the transient finding “line”. When the data trace gets steeper than the transient finding line and also other criteria like the amplitude of the transient are met, the transient has been found and extracted from the original data curve.

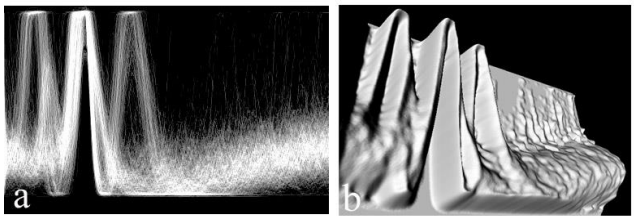


Figure 6. a. The distributions of transient classes as determined by the transient finding function and superimposed on the data matrix as described above. b. The data from a have been converted to the VRML elevation grid and observed within a VRML virtual worl.

is in slovene schools still so undeveloped that it does not make any sense to produce in the electronic format more as just the introductory information. The experience of publishers is also that the CD ROMs do not sell well, internet is “free” and they so are not keen to invest much into the educational software. Therefore the main supporter of e-learning remains the state or the international community which has been doing great steps on this field lately.

References

[1] Cell-Tissue-Human Body. http://www.bioanim.com

[2] Terragen: http://www.planetside.co.uk/terragen/

[3] Andrej Gogala: Narava Slovenije, Ljubljansko barje in Iška. Prirodoslovni muzej Slovenije, 2001

[4] Unity 3D https://unity.com/

[5] Amon, T. A new computer program for neuronal spike data evaluation. Biol Vestn 1992; 40(2):pp 1-8

[6] 3D Model Viewer JavaScript Ar-Ty tool: https://artystore.com/product/3d-model-viewer-store-master/

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Usage profile in physical systems modelized with stochastic hybrid automata

Gaël Hequet, Nicolae Brînzei, Jean-François Pétin
Université de Lorraine, CNRS, CRAN
F-54000 Nancy, France
{gael.hequet, nicolae.brinzei, jean-francois.petin }@univ-lorraine.fr

***Abstract*— This paper focuses on a way to represent complex systems by using a modification in the Stochastic Hybrid Automata (SHA) first designed by the CRAN by expending finite-state automaton. SHA are used to represent continuous physical phenomenon (aging for example) that can induce discrete event (failures). With this tool, it is possible to represent the lifecycle of a component, its aging, and failures. To have a more precise model, a new addition to the SHA can be done. A system can have many usage profiles and each of these profiles can change the aging and probability of failures of the system. This work aims to add this view to the SHA by creating a virtual racing car. The driver of this racing car can drive in different ways. Depending on his driving profile, the components of the car (body, engine, tyres, electronics, clutch, transmission, brakes, and gearbox) can suffer from different types of degradations. The level of degradation of a component has an impact on its probability of failure. In addition to this impact given by the degradation, another impact is considered, the profile of driving itself. Depending on the driving profile, the failure rate can vary. In this idea, different failure rates are linked to each driving profile. The impacts of these profile variation will be shown in this paper.**

Keywords— Usage profile; Stochastic Hybrid Automaton; Reliability; hybrid system.

I. INTRODUCTION

In reliability engineering, the needs to have precise and close to reality models are important [1-2]. More critical a system is, more we need to have a good knowledge of it. In this domain, the goal is to be able to predict the degradation and the time before system failure. With those data, it is easier to manage maintenance strategies and improve the availability. Unreliability can lead to great risks in human and economical ways.

Because of it, reliability studies are becoming vitals in systems design. To this extend, reliability engineers need models to represent their systems. At component level, many models already exist like exponential, Rayleigh, or Weibull probabilities laws to describe the lifetime distribution of system components. Those probability laws give tools to help represent the occurrence of failure in a system. They are commonly used in reliability engineering to prevent failure and create maintenance strategies.

At system level, modelling tools such as Petri net or finite-state automaton can be used to represent the occurrence of components failure and its impact on the global behaviour of the whole system. To be more precise, a reliability engineer needs to consider continuous phenomena and discrete event. Aging is a phenomenon depending on the continuous physical process carried out by system, but a discrete failure can occur when the aging reaches a critical threshold. In addition to this problematic of having precise models, todays' systems are becoming more and more complex. A system is composed of many subsystems and each have its own discrete and continuous evolution.

To answer to these difficulties, stochastic hybrid automaton was defined at CRAN [3-4]. Thus, an automaton can be described by a set of states, continuous equations, and discrete events.

In addition, a system can evolve differently depending on its surrounding or its usage profile. An airplane won't degrade the same way if it is on the ground or flying. A car will degrade faster if it is used more aggressively. In this idea, this paper will focus on an addition to the SHA to consider the usage profile of a system.

To present this add, this paper will firstly present the SHA more precisely before presenting the new model. In a third part, a scenario with a virtual racing car and a driver who can change his driving profile during a race will be created. The car and its components will have different failures and degradation rates according to the driving profile. The results of this model will be presented and discussed in the last part.

II. STOCHASTIC HYBRID AUTOMATON

In the industry, systems behaviour can change depending on continuous physical process, like aging, and also discrete event, like starting or stopping the system. To represent those behaviours, stochastic hybrid automaton can be used.

A. Concept of stochastic hybrid automaton

Stochastic hybrid automaton is a concept used to represent a hybrid system. A finite-state automaton is used to list each state of the system.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

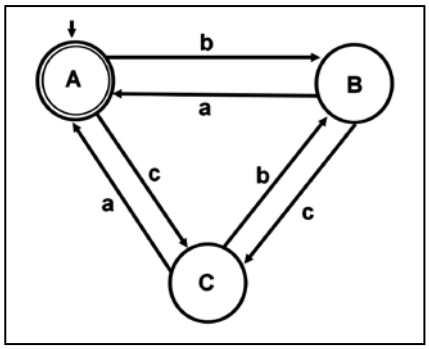


Figure 1. Example of a finite-state automaton with 3 states (χ = {A,B,C}) and 3 types of transitions ($\sum$ = {a,b,c})

A finite-state automaton (FSA) can be described as a 5-tuple [5]:

$$FSA = (\chi, \sum, f, x_0, \chi_f) \qquad (1)$$

where χ is the finite set of states, $\sum$ is a finite alphabet, f is the transition function used to change states, x_0 is the initial state and χ_f is the set of final states.

The goal of the hybrid stochastic automaton (SHA) is to link states to continuous equation and transitions to stochastics or determinist phenomenon. Continuous equations are used to describe the physical processes of the system. Thresholds on these continuous variables are added to initiate the transitions between states in addition to the stochastic variables. A SHA can be described as a 11-tuple [3]:

$$SHA = (\chi, E, Ar, X, A, H, F, p, \chi_0, x_0, p_0) \qquad (2)$$

where χ represent a finite set of states, E a finite set of events, Ar the edges, X is a finite set of continuous variables. A is a function of "activities" that connects a state (χ) to a variable (X). H is a finite state of clocks. F is an application that links each clock to a probability distribution function. Here, p is a probability distribution of state transition. Finally, χ_0, x_0 and p_0 are used to give the initial state of the automata with, respectively, the initial state, initial values in the initial state and initial probability distribution of state transitions.

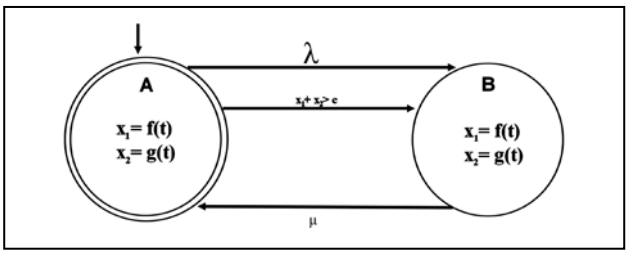


Figure 2. Example of a Stochastic Hybrid Automaton with 2 states, 2 variables governed by 2 activity functions and 3 transitions between states.

A simple stochastic hybrid automaton is shown above with two states, A and B. This automaton has two continuous variables x_1 and x_2 each assigned to a time dependent function (f(t) and g(t)). There are two transitions going from A to B. The first transition represents the occurrence of a failure described by a failure rate λ. The second transition is a condition, it is fired if x_1 plus x_2 is upper than a constant "c". Finally, the last transition is a governed by a repair rate "mu".

B. Use cases

SHA were used to study and model hybrid system [3-4, 6]. For example, in [3-4], an oven and its controlled temperature system were modeled using SHA. In this example, the temperature needs to be followed. If a failure in the control system occurs, it can still heat up the system until a temperature that could be harmful to the system and cause new failures. A stochastic failure can create a continuous phenomenon and induce a new failure.

SHA were used also for determining event sequences that occur in the case of controlled steam generator of a nuclear power plant and the assessment of occurrence probability of such event sequences [7], and for the taking into account aging of a data cluster installation [8].

With more and more complex hybrid systems, we need to find tools that can adapt and consider the various factors that can lead to feared events. Sometimes, simplifications allow us to have similar results. But in the case of critical systems, it is sometimes necessary to accept a greater complexity of the models to prepare for cases which would not necessarily be visible after simplifications. For this purpose, SHA may be of interest to study the evolution of hybrid systems. From this perspective, it may be interesting to seek to further improve the precision of the studied models. To do this, a proposition will be given in the next part.

III. AN IMPROVEMENT TO STOCHASTIC HYBRID AUTOMATON

To improve precision and pertinence of SHA, a modification can be done to these automata. Real hybrid systems can have different degradation depending on their uses. This part will explain the purpose and how they can be achieved.

A. Goals and models

A system won't degrade the same in a nominal use or in a over-revving use. Over-revving will not systemically create a failure, but it can degrade the component. To consider it, the model needs to keep a track of it. Depending on the passed time in this usage regime, it won't have the same impact. More time the system passes in this over-use, more it will be degraded.

For this reason, the SHA will be extended by adding new attributes to the definition tuple: a finite set of usage profile "U_p"; "D_l", a variable that represents the degradation level of the automaton; "D_{lf}", a set of finite functions linked to "D_l" and "U_p"that represents the impact of the usage profile on the degradation level. "f_λ" if a set of functions especially linked to failures rates to change them depending on time, variables, or conditions. "U_{p0}", "D_{l0}" being the initials usage profile and

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

degradation level. Consequently, the SHA will be defined by the following 17-tuple:

$$PBSHA = (\chi, E, Ar, X, A, H, F, p, U_p, D_l, D_{lf}, f_\lambda, \chi_0, x_0, p_0, U_{p0}, D_{l0}) \quad (3)$$

With these new attributes, the automaton can consider usage profile and keeps a memory of previous phenomenon and events. Those memories can then be used to refresh degradation levels and failures rates. This modified automaton will be called Profile-Based Stochastic Hybrid Automaton (PBSHA) in the rest of this paper.

Another purpose of the PBSHA is to bring together the effects of the functional use and the dysfunctional behaviour of the system. Here, it can be done through the usage profiles. Another feature is added to follow the degradation level of a component.

Here is a simple example to show the general behaviour of a PBSHA:

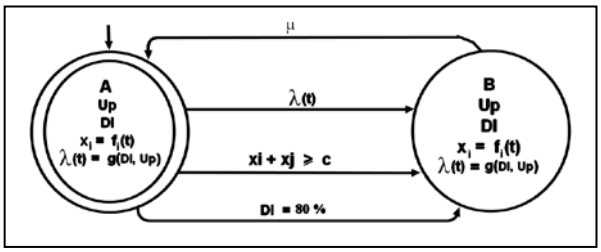


Figure 3. Example of a Profile-Based Stochastic Automaton with 2 states and 4 transitions with A, the initial state.

The automaton presented previously has 2 discrete states (A and B). These states could be OK and KO states. The controller is described by the system use profile (Up). Its level of degradation (Dl), depending on the time spent in the different use profiles. System-specific variables (x_i could be, for example, a temperature, a volume or other). Finally, the failure rate (λ) depends on the degradation level of the system as well as its usage profile.

The two states are linked by 4 types of possible transitions. The first is drawn according to a repair rate μ. The second by a failure rate $\lambda(t)$. This failure rate is variable in this example, and it considers the time spent in the various possible use profiles as well as the level of degradation of the system. The third transition works in the same way as for SHA, a condition on continuous quantities of the system. Finally, the last transition is fired when the degradation level reaches a threshold (here, 80%).

B. Use case

The hybrid stochastic automaton will aim to increase the precision on hybrid models where usage profiles will be observed.

For example, by taking the previous oven system in [3], it could be interesting to see if fast or slow heating profiles are present. How the reliability of the system could fluctuate. To have relevant models it would be necessary to carry out preliminary studies on the system to observe the impact of different physical quantities on the system. It is therefore necessary to know the system well as well as its interactions to model it faithfully.

In the remainder of this paper, an example in the form of a thought exercise will be used as a model. The case of a racing car allows aspects of a complex hybrid system to be considered. The pilot has his own behaviour and his manner of driving. In addition, he can choose to be aggressive or not in his driving and in the use of the car components. These choices can have a significant impact on the degradation of the car's subsystems. The car will be considered as a system made up of sub-systems which will each have their associated PBSHA.

IV. EXPERIMENT

A. Experimental design

a) Context

The studied system is a racing car with a race lasting 1 hour and 45 minutes (1.75 hours). The car is made up of different subsystems. The components considered are engine, brakes, gearbox, tires, electronics, transmission, clutch and the body car.

This car will be driven by a driver who will follow a predefined type of strategy. The pilot will have an impact on handling, the fuel mix used and the use of Kinetic Energy Recovering System (KERS) [9]. In this study case, not all possible strategies will be covered. In the studied strategies, this will have four possible profiles. A profile called "resting" in which the car is not racing. It is chosen not to degrade the subsystems in this profile. Another “conservation” profile where the car is racing, and where the goal is to degrade the car as little as possible. A “standard” profile where the car undergoes normal degradation. Finally, an “aggressive” profile in which the driver pushes the car to its limits with a strong degradation.

Here, four scenarios are played to study the degradations. The first is the case where the pilot remains in “conservation” mode throughout the race. The second corresponds to the case where the pilot remains in “standard” mode. The third concerns the "aggressive" regime. Finally, the last scenario is mixed, i.e., it will be composed of different regimes with firstly 19 minutes in “conservation” profile, then 24 minutes in the "aggressive" profile, 42 minutes in the "standard" profile and lastly, 20 minutes in the "conservative" profile. These times spent in the different profiles are considered here completely arbitrary and in no way represent real strategies. They are used to see the difference between the beginning of the race and the end in the “conservative” profile.

It is considered that when a component fails, the driver puts the system to rest, and the component is considered as totally degraded. The other components therefore no longer degrade. It is also considered that there is no repair or replacement during the race. Likewise, common cause failures will not be taken into account (accident with another car, loss of control or other).

The performance character will not be studied. The impact of the environment will not be considered either.

Indeed, depending on the environment and the conditions of the race, the subsystems will not degrade in the same way. Finally, for the sake of simplification, the different subsystems will all be subjected to the same profile and will have the same degradation and failure rates (as will be shown in the next section), but different profiles can be taken into account without any restriction.

b) Developed models

To model the system, it was decided to represent it as follows:

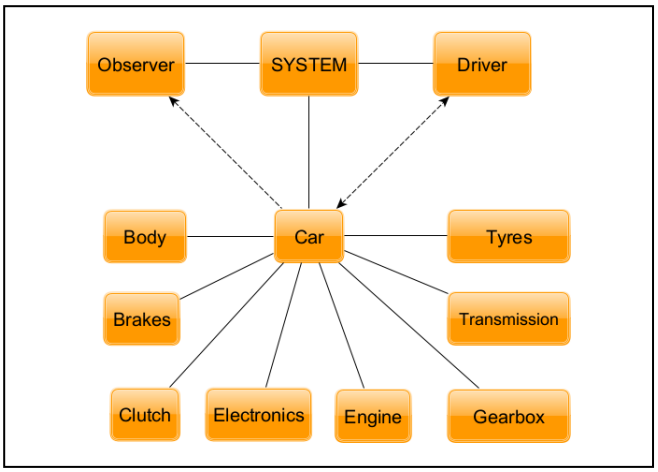


Figure 4. Global representation of the system

A first class named “system” will initialize the driver, an observer as well as the car.

The role of the driver is to drive the car. Its strategies and profiles will be sent to the car. The pilot can modify his profile on the handling, fuel mixture and use of the KERS. Here is the automaton governing the Driver class:

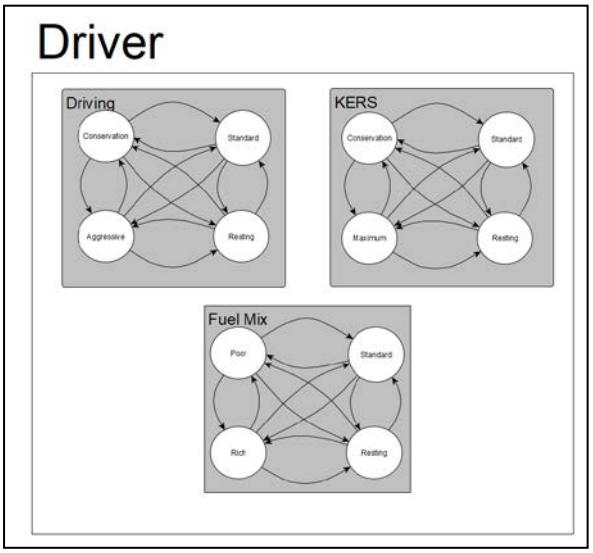


Figure 5. Automaton governing Driving, ERS and Fuel mix behaviour.

Despite these possibilities, it will be considered for this example that the pilot will only have access to four types of profiles from the four strategies stated in section IV. A.a). The strategies will therefore be as follows:

TABLE 1. STATES OF AUTOMATON ACCORDING TO THE CHOOSEN STRATEGY

Strategy	**Driving**	**KERS**	**Fuel Mix**
Resting	Resting	Resting	Resting
Conservation	Conservation	Conservation	Poor
Standard	Standard	Standard	Standard
Aggressive	Aggressive	Maximum	Rich

These behaviours will be transmitted to the car and its components, which will thus be subject to degradation specific to the use profile.

The components behavior is modeled by the following automaton:

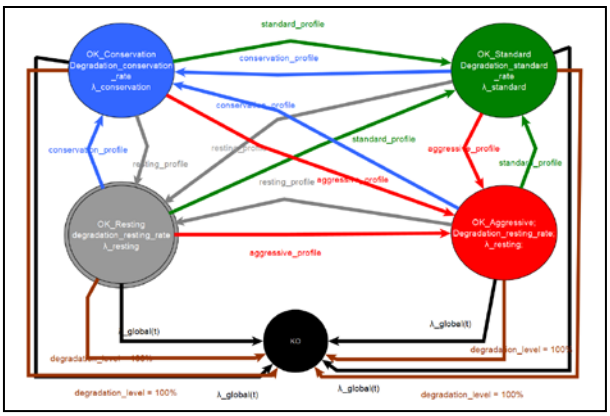


Figure 6. “aBasicComponent“, a profile-based stochastic hybrid automaton

This PBSHA consists of five discrete states to observe the possible cases according to the strategies. Four of these states represent component operating profiles. These operating profiles make it possible to modify the failure rate associated with the system as well as the rate of degradation undergone by the component. For example, in an aggressive strategy, the components will be in the OK_aggressive state and will be subject to the magnitudes of this usage profile. That is, the failure rate will be that associated with this profile added to the failure rate resulting from the level of degradation of the system. Level of degradation that will increase at the rate associated with the aggressive state.

This level of degradation is updated according to a time step and it is given by following formula:

$$D_l(t_n) = (t_n - t_{n-1}) * D_{RateProfile} + D_l(t_{n-1}) \quad (4)$$

where D_l is the degradation level of the component, t is the time and $D_{RateProfile}$ is the degradation rate related to the profile used. If the component fails, its degradation level will be equal to 100%.

This degradation has an impact on the component failure rate λ. There is therefore for each component a profile failure rate ($\lambda_{profile}$), a degradation failure rate ($\lambda_{degradation}$) and these two rates form the overall failure rate (λ_{global}).

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

$$\lambda_{degradation} = e^{(Dl/10 - 10)} \quad (5)$$

$$\lambda_{global} = \lambda_{profile} + \lambda_{degradation} \quad (6)$$

with the level of degradation $\{D_l \in R_+ \mid 0 \leq D_l \leq 100\}$. The value "10" is here to have $\lambda_{degradation} = 1$ when the degradation level 100%. It has no real meaning; it only gives an upper limit.

These models and variables will be implemented under PyCATSHOO [10], a software created within EDF and which has been designed to integrate hybrid stochastic automata with discrete and continuous components. Simulations will be done using the Monte-Carlo method.

c) Values

In this section, the initial values for the different scenarios are presented.

The first scenario is that of "conservation". That is to say that during the 1 hour and 45 minutes of racing, the driver will not change the use profile. The second scenario is the "standard" scenario. The third scenario is the "aggressive" scenario. Finally, the last scenario is the "mix" scenario. In this scenario, the pilot will go through the different user profiles during the race. The initials values are the same for each scenario. Only the variable "Strategy" is changing to "conservation", "standard", "aggressive" or "mix" according to the corresponding scenario.

The number of sequences simulated is to obtain trends but not precision. 100,000 sequences will be simulated by scenarios. The lowest failure rate is 10^{-3} for an exponential law. This means that for a confidence of 95%, we will have an interval of plus or minus 6.2% of error on the results of the 100,000 simulations.

The method used to generate random numbers is the "yet another random number generator" (yarn5) method [11].

The simulations are calculated in a computer with an AMD Ryzen 5 2600 Six-Core processor on 11 threads at 3.8 MHz. The required time for 100,000 simulations was between 180 and 240 seconds.

TABLE 2. INITIALS VALUES

Variable	**Values**	**Reason**
Number of sequences	100,000	The goal is to see trends. Precision is not the primary goal of this study.
RNG seed	50	Seed number used for yarn5
V_clock_step	0.005 hour	Decrease time complexity while maintaining sufficient precision
V_degradation_level	0.0	No degradation at the beginning
V_resting_degradation	0.0 %/h	No degradation in resting profile
V_conservation_degradation	5 %/h	Small degradation, not a realistic value
V_standard_degradation	15 %/h	Standard
		degradation, not a realistic value
V_aggressive_degradation	30 %/h	High degradation, not a realistic value
P_lambda_degradation	$e^{(Dl/10 - 10)}$	Formula to represent the fact that the more degraded a component, the more chances it has to fail
T_resting	0.0	No time in profile at initialisation
T_conservation	0.0	No time in profile at initialisation
T_standard	0.0	No time in profile at initialisation
T_aggressive	0.0	No time in profile at initialisation
P_lambda_resting	1/876	Ten failures per year
P_lambda_conservation	2/876	Twenty failures per year
P_lambda_standard	10/876	One hundred failures per year
P_lambda_aggressive	15/876	One hundred and fifty failures per year
P_lambda_global	1/876	Ten failures per year. It takes the resting failure rate at initialisation (before being updated during the simulation)
V_A_profile	False	The driver will give the profile after reading the strategy
V_B_profile	False	The driver will give the profile after reading the strategy
V_C_profile	False	The driver will give the profile after reading the strategy
V_Z_profile	False	The driver will give the profile after reading the strategy
Tmax	1.75 hour	The duration of a race
Strategy	"conservation", "standard", "aggressive" or "mix"	This variable is used to give orders to the driver on how he will manage the race

B. Results and discussion

During a scenario, the data collected and monitored are the availability of the car, the availability of its components, the level of degradation of the components, the associated degradation failure rates, and the overall failure rates of the components.

These data are shown below:

“Conservation” scenario:

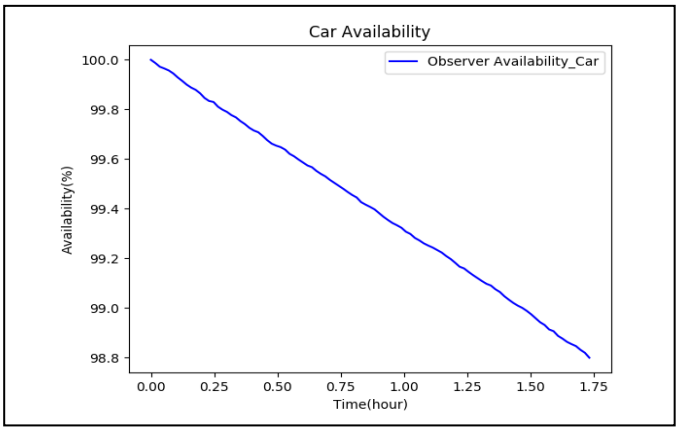


Figure 7. Car availability (%) over time (hour) in the “conservation” scenario.

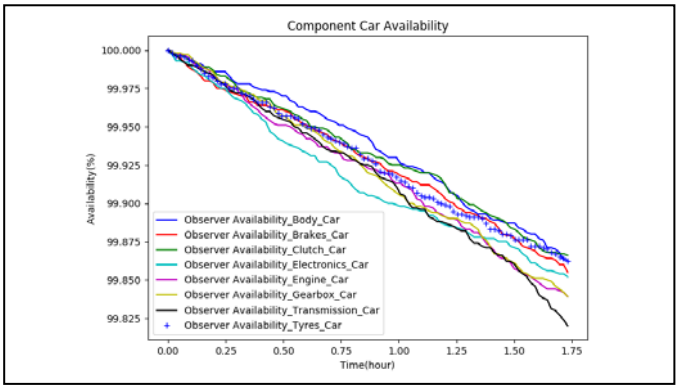


Figure 8. Component availability (%) over time (hour) in the “conservation” scenario

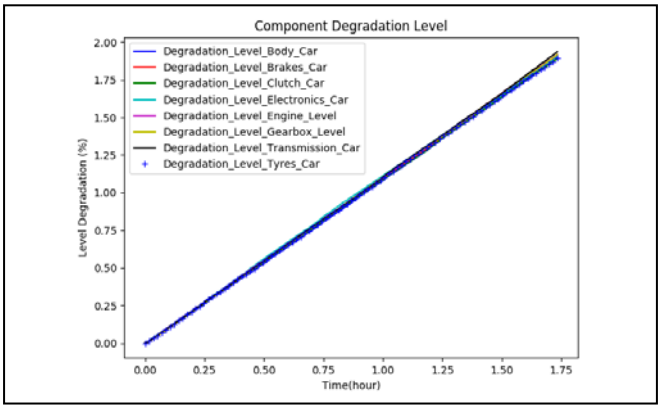


Figure 9. Degradation of the components (%) over time (hour) in the “conservation” scenario

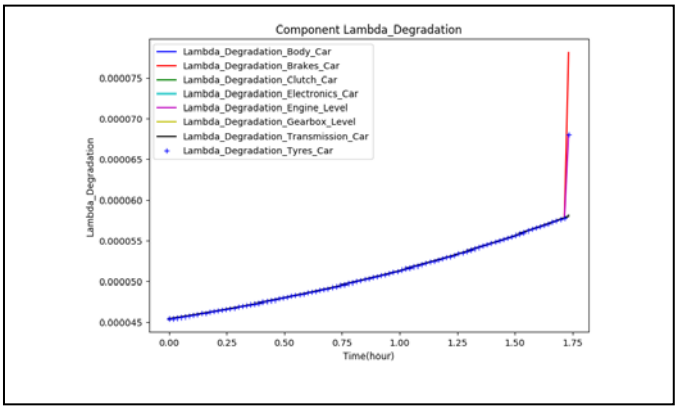


Figure 10. Evolution of the failure rate $\lambda_{degradation}$ over time in the “conservation” scenario

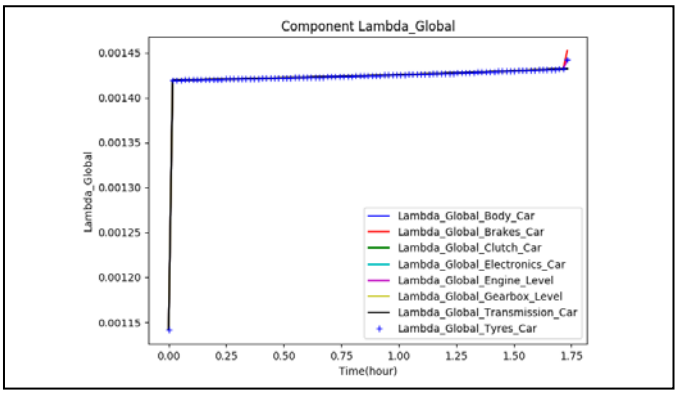


Figure 11. Evolution of the global failure rate λ_{global} over time in the “conservation” scenario

In this scenario, in the figure 11, the first jump at the start of the race is explained by taking into account of the first degradations and the launch of the car on the track. Indeed, this jump represents the considering of $\lambda_{degradation}$ in the λ_{global}. This behaviour is seen in the other scenarios as well.

Another special case is seen at the end of the race. This can be explained by an increase in the failure rate which was very low initially. An increasing number of failures significantly increases the mean level of degradation D_l which increases $\lambda_{degradation}$ and which in turn increases λ_{global}.

“Standard” scenario:

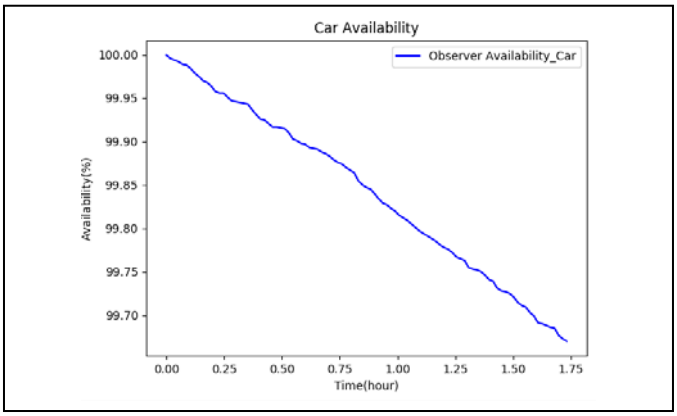


Figure 12. Car availability (%) over time (hour) in the “standard” scenario.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

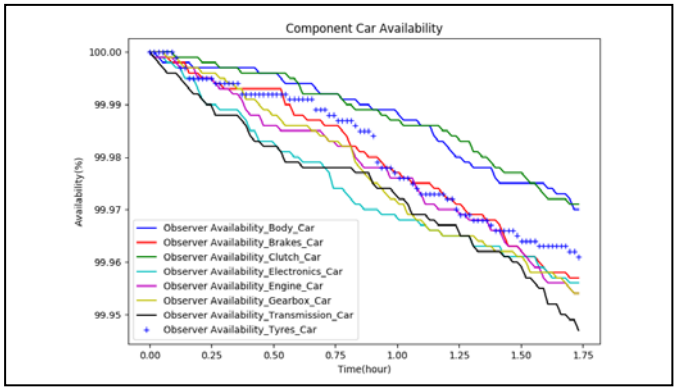


Figure 13. Component availability (%) over time (hour) in the “standard” scenario

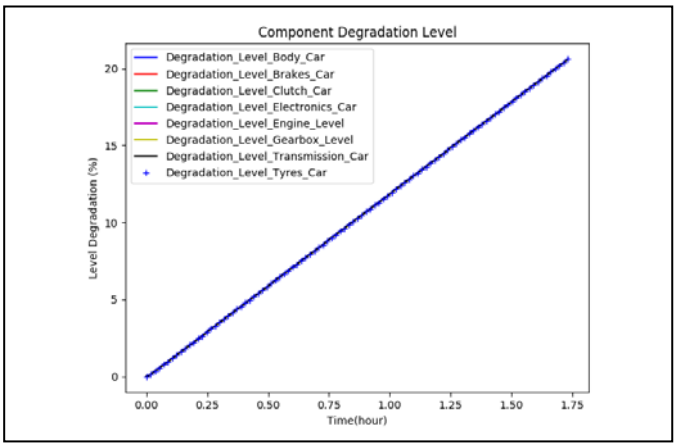


Figure 14. Degradation of the components (%) over time (hour) in the “standard” scenario

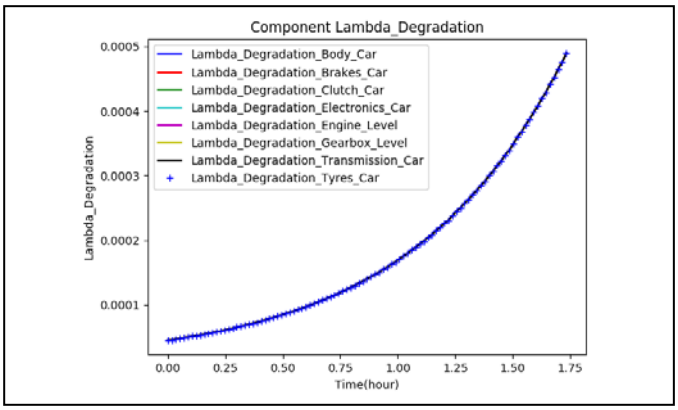


Figure 15. Evolution of the failure rate $\lambda_{degradation}$ over time (hour) in the “standard” scenario

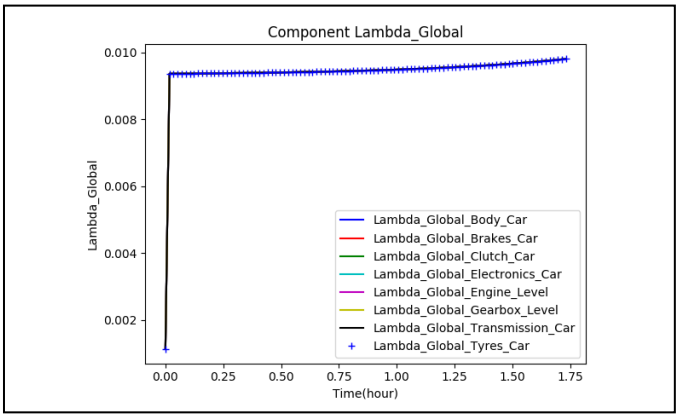


Figure 16. Evolution of the global failure rate λ_{global} over time (hour) in the “standard” scenario “

Aggressive” scenario:

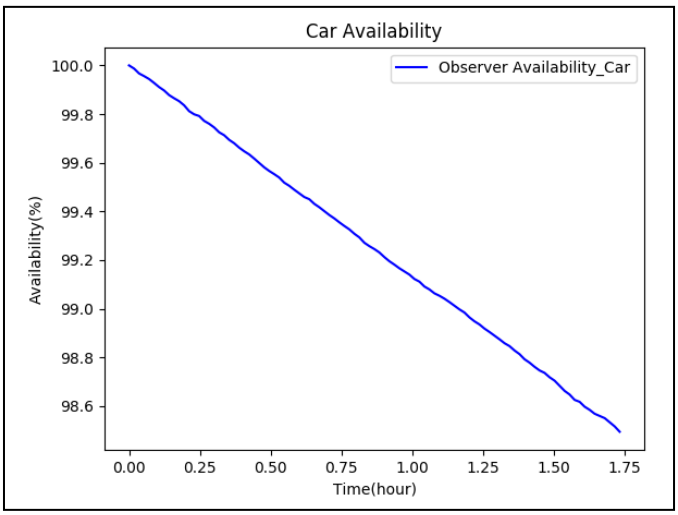


Figure 17. Car availability (%) over time (hour) in the “aggressive” scenario.

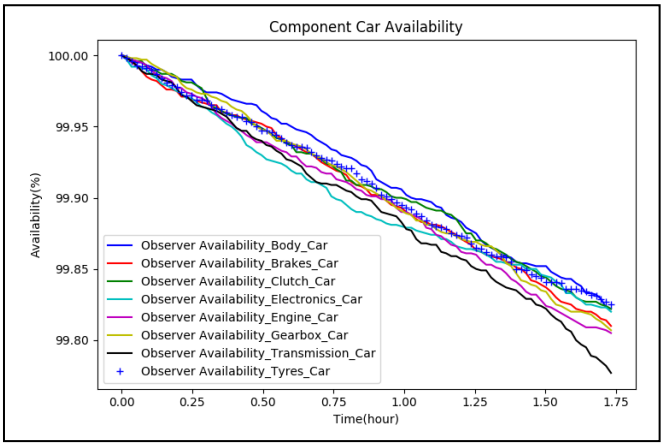


Figure 18. Component availability (%) over time (hour) in the “aggressive” scenario

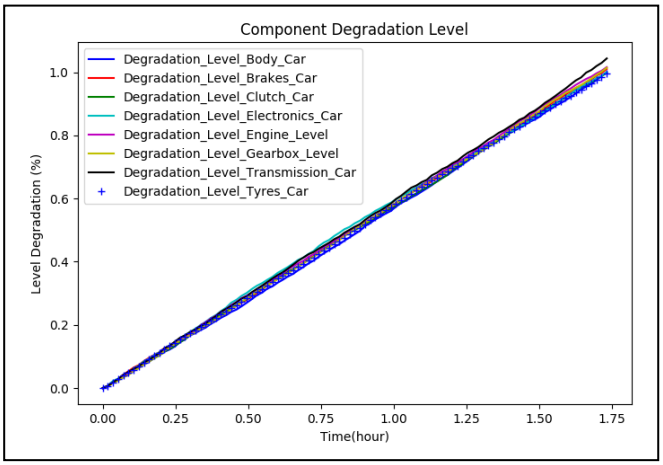


Figure 19. Degradation of the components (%) over time (hour) in the "aggressive" scenario

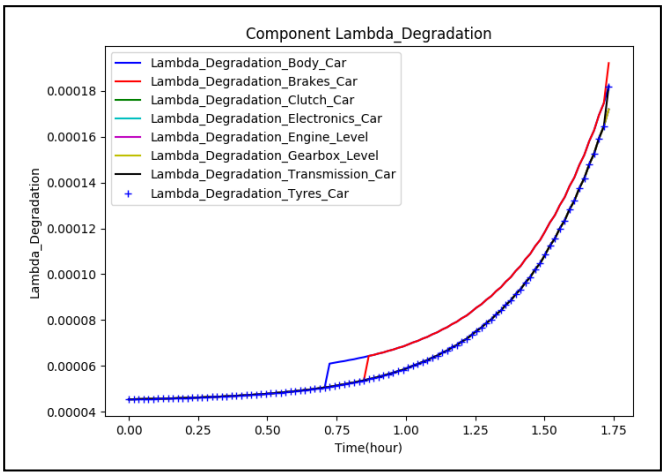


Figure 20. Evolution of the failure rate $\lambda_{degradation}$ over time (hour) in the "aggressive" scenario

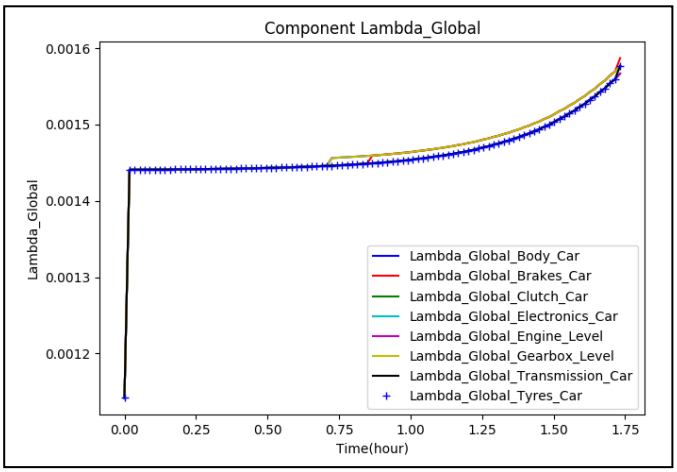


Figure 21. Evolution of the global failure rate λ_{global} over time (hour) in the "aggressive" scenario

All the availability curves (figure 7, 8, 12, 13, 17, 18) are given to have an overall idea of the evolution. The differences between the scenarios are small because the time of a race is too small and the car is new. If the car is used for more races, differences will be more relevant because they will be influenced by degradations from previous races.

In addition, the values displayed are averages. For example, one of the components fails in the first minutes, then the other components no longer degrade. Their degradation level will therefore be low, inducing mean values which may also be low. This can explain the low degradation level for the "aggressive" scenario (figure 19) and the low failure rates (figure 20 and 21). Those low values shows that more components fail at the beginning of the race. Which induce more scenarios where components have lower degradation levels and lower degradation failure rate.

It is still interesting to observe the evolution of failure rates of each scenario. It shows that more aggressive a usage profile is, its failure rate increases faster. With a more important failure rate, there is bigger chances to observe failures. It is easier to see this phenomenon with the "aggressive" scenario. When the global failure rate increases in figure 21, at 1.25 hours and after, of race, a greater interval of degradation level between components is seen in figure 19. This phenomenon can be seen in other scenarios but more hidden.

In these cases, using an exponential function to determine the degradation failure rate considers the fact that at the start of the component's life, the degradation does not significantly increase the chance of failure. But over time the slightest degradation can greatly weaken the system and make it more prone to failures. Although this function is not inspired by real cases, this behaviour is interesting to remember the previous life of the component.

The last "mix" scenario being the scenario allowing to see all the profiles, its data is presented below.

"Mix" Scenario:

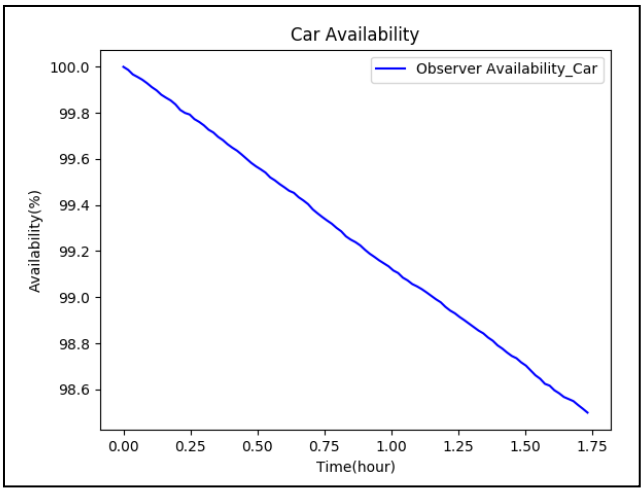


Figure 22. Car availability (%) over time (hour) in the "mix" scenario

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

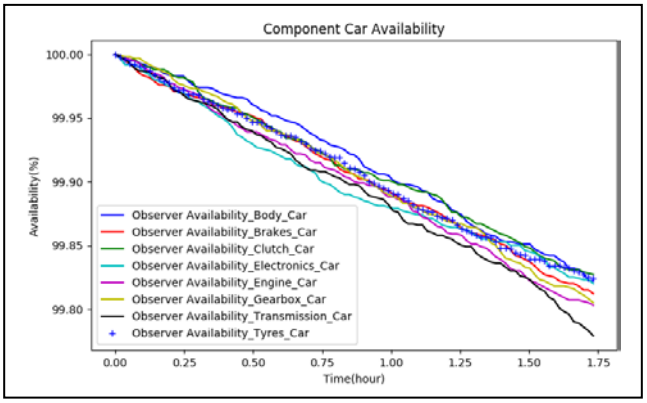


Figure 23. Component availability (%) over time (hour) in the “mix” scenario

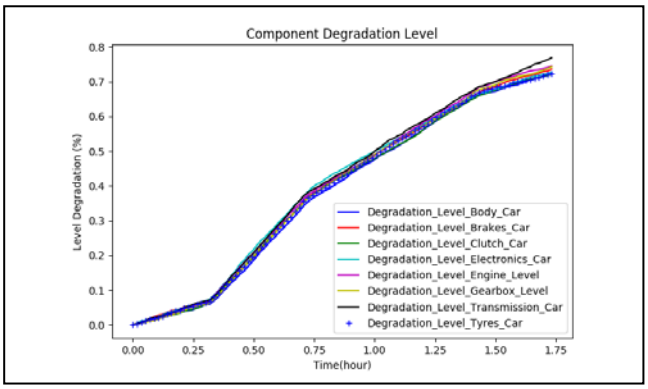


Figure 24. Degradation of the components (%) over time (hour) in the “mix” scenario

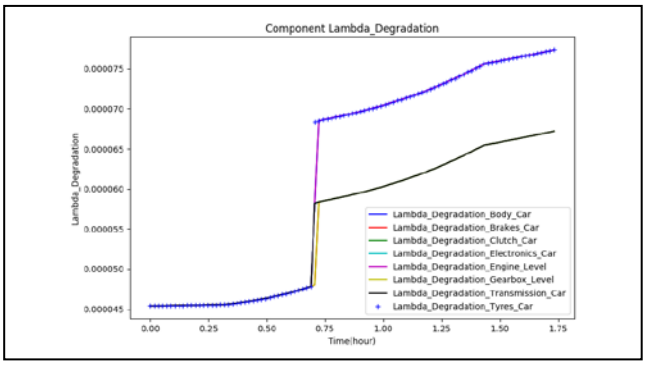


Figure 25. Evolution of the failure rate $\lambda_{degradation}$ over time (hour) in the “mix” scenario

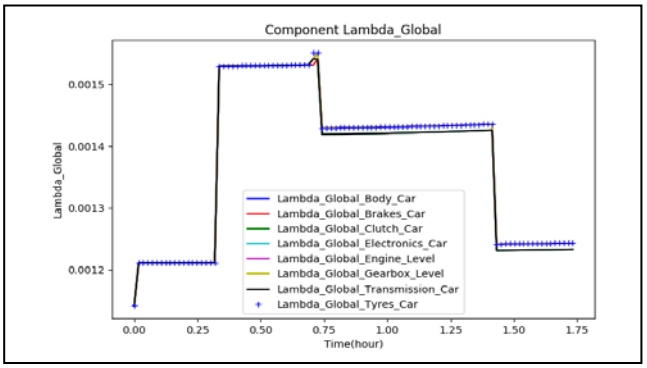


Figure 26. Evolution of the global failure rate λ_{global} over time (hour) in the “mix” scenario

It should be noted in the “mix” scenario that the degradation is very different between the different profiles as it can be seen in figure 24. This difference is clearly seen in the overall failure rate (figures 25, 26). Likewise, a difference in value can be seen between the two parts passed in the “conservation” profile. This is explained by the consideration of the degradation of the components. Degradation resulting from the use of the system in "aggressive" and "standard" profile severely degraded the system as can be seen in fig. 24. The value jump of the degradation failure rate in fig. 25 can be explained in the same way as the jump observed in fig. 10. Failures increasing the degradations to 100% and causing a jump in the average values. This burst resulting from failures then has an impact on the overall failure rate for the rest of the race. Thus, depending on the systems studied, it is possible to observe a variable failure rate and above all which can be very different depending on the system's operating strategy.

V. Conclusion and Perspectives

It is therefore possible to modify hybrid stochastic automata to make them take into account use profiles. This can be useful for monitoring the dynamics of systems while considering their usage profiles. This type of automaton can be interesting to use on complex systems for which it is needed to consider (in their reliability and availability assessment) the degradation of their components over time; it becomes relevant to study systems keeping in memory the past life of the system. As well as considering its failure rate as variable over time.

Furthermore, some perspectives could be interesting to be developed in the future. For the future, it would be relevant to add common cause failures but also to add an impact between the degradation and the system performances. Indeed, a degraded system could see its performances being modified with time. Finally, in the construction of the automaton, it is easily visible that the number of states of the system can grow extremely, if the profiles are considered as states and the use of profiles can avoid such large increase of state numbers. Another perspective could be to transform these discrete profiles into a level of use ranging from 0 to 100% (or more to represent overspeed) and consider the system as a multi-state system (MSS). It would thus be possible to make comparisons between the approach by PBSHA and the approach using the theory of MSS.

Acknowledgements

This research work was supported jointly by the “Digital Reactor” project funded by Public Investment Bank (BPI France) under the call for "Structuring Projects for Competitiveness" (PSPC) of the Future Investments Program (PIA) and by the ACeSYRI (Advanced Centre for PhD Students and Young Researchers in Informatics) project funded by the Erasmus+ Program of the European Union under the contract number 610166-EPP-1-2019-1-SK-EPPKA2-CBHE-JP.

REFERENCES

[1] O'Connor, Patrick D. T., et Andre Kleyner. *Practical Reliability Engineering*. 5th ed. Chicester : Wiley, 2012.

[2] Zio, E. « Reliability Engineering: Old Problems and New Challenges ». *Reliability Engineering & System Safety* 94, nº 2 (février 2009): 125-41. https://doi.org/10.1016/j.ress.2008.06.002.

[3] Perez Castaneda Gabriel Antonio, Aubry Jean-François, Brinzei Nicolae. "Stochastic hybrid automata model for dynamic reliability assessment", *Proceedings of the Institution of Mechanical Engineers Part O Journal of Risk and Reliability*, 225, 1 (2011) 28-41.

[4] Aubry J.F. and Brînzei N. (2015). Systems Dependability Assessment: Modeling with Graphs and Finite State Automata. Wiley-ISTE.

[5] Arnold, André. "Finite transition systems and semantics of communicating processes", « Systèmes de transitions finis et sémantique des processus communicants ». Paris : Masson, 1992.

[6] Broy, Perrine. "Safety assessment of complex hybrid dynamic systems. Application to hydraulic systems", « Evaluation de la sûreté de systèmes dynamiques hybrides complexes. Application aux systèmes hydrauliques ». Phdthesis, Université de Technologie de Troyes, 2014. https ://tel.archives-ouvertes.fr/tel-01006308

[7] Babykina G., Brînzei N., Aubry J.F., Deleuze G., "Modeling and simulation of a controlled steam generator in the context of dynamic reliability using a Stochastic Hybrid Automaton", *Reliability Engineering and System Safety*, 152, (2016) 115-136.

[8] Chiacchio F., D'Uso D., Manno G., Compagno L. "Stochastic hybrid automaton model of a multi-state system with aging: Reliability assessment and design consequences", *Reliability Engineering and System Safety*, 149, (2016) 1-13

[9] « Kinetic Energy Recovery Systems in Formula 1 ». Accessed 27 april 2021. http://large.stanford.edu/courses/2015/ph240/sarkar1/.

[10] Chraibi, Hassane. « Getting Started with PyCATSHOO V1.2.2.8 Document Version V1. », s. d., 86. http://pycatshoo.org/

[11] *intel/yarpgen*. C++. 2016. Reprint, Intel Corporation, 2021. https://github.com/intel/yarpgen.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Reliability Analysis of Finite-Source Retrial Queuing System with Collisions and Impatient Customers in the Orbit Using Simulation

Ádám Tóth, János Sztrik, Ákos Pintér and Zoltán Bács.

***Abstract*—In this paper, we have developed a simulation program to investigate M/M/1//N and M/G/1//N retrial queuing systems with collisions and impatient customers in the orbit. In our model, for lack of waiting queues, the service of an arriving customer begins immediately. Otherwise, it has the ability to bring about a collision in which both the arrived and request under service are forwarded to the orbit where spending some exponentially distributed random time they try to get their service demand to be executed. All requests possess an impatience property resulting in an earlier departure from the system through the orbit if they spend too much time waiting for being served properly. The phenomenon of blocking is applied not allowing the customers into the system while the service unit resides in a faulty condition. The server is supposed to break down according to several distributions and this work concentrates on examining the effect of these distributions on several performance measures like the distribution of the number of collisions and failures of customers. The results are graphically illustrated to experience the difference among the used parameter settings of the various distributions.**

***Index Terms*—retrial queues, finite-source queuing system, server breakdowns and repairs, simulation, impatience.**

I. Introduction

Currently Internet usage increases due to mainly the rapid development of cloud service providers providing highly reliable, scalable, low-cost infrastructure platforms and the great number of devices having the ability to transfer data over a network. Network activity initiated by the users from home and companies escalates in such a way that models of communication systems are needed to fully understand and optimize the operation of network traffic. In real-life scenarios, the usage of retrial queues is an effective tool to cope with issues in telecommunication systems like telephone switching systems, call centers, CSMA-based wireless mesh networks in frame level and computer systems. The following articles express its importance in different fields of informatics [1], [2], [3], [4], [5].

It is not an uncommon feature of retrial queues possessing a virtual waiting area the so-called orbit in the case of modeling such systems. During managing administration duties a customer may wait for its turn to receive its service need when all agents are occupied in a call center. Numerous papers carry out investigation on models having orbit for example in [3],[4],[5].

Á. Tóth, J. Sztrik, Á. Pintér and Z. Bács are with University of Debrecen, University Square 1, Debrecen H-4032, e-mail: toth.adam,sztrik.janos@inf.unideb.hu, apinter@science.unideb.hu, bacs.zoltan@econ.unideb.hu

It can be observed that in many cases customers may abandon or renege from waiting in a queue or a system that is presented in many situations in our everyday life. For instance, after waiting a specified amount of time customers can hang up waiting for an agent in a call center or customers may decide not to enter a shopping mall if it is so crowded. Every customer has a patience threshold value which depends on many factors, in this model, this relies on the amount of time spent in the system so customers are characterized by reneging features. Some results in connection with impatience can be viewed in the following articles [6],[7].

In connection to the available channels or facilities, users and sources are fighting for them to fulfill their service needs as soon as possible. Because of the possibility of launching attempts at the same time collisions may take place resulting in the loss of transmission and an unfortunate delay during the provision of services. Steps are needed to be done having an efficient procedure to prevent conflict and corresponding message delay and to ensure retransmission if it is necessary. Many papers have published results including models with collisions in [8],[9],[10],[11].

Unfortunately, service units are subjected to breakdowns which play quite an important role to alter the performance of the system and performance measures. In the literature, some articles neglect failures in the model construction which is quite unrealistic but hardware could fail and networks could have intermittent timeout periods. It is important to have plans for such events experiencing disruptions so the examination of retrial queueing systems with random server breakdowns and repairs is essential, the impact would be crucial for the system if the downtime is too high. Transmission failures, interruptions throughout transferring packets may have severe consequences that can cause the end of a company or big financial losses. Reliability analysis has been carried out by several works like [12],[13],[14],[15],[16],[17],[18],[19].

We examined the operation of type $M/M/1//N$ a retrial queueing system where the service unit is unreliable and customers may depart from the orbit if they spend too much time waiting to be served. The novelty of our work is to carry out sensitivity analysis using different distributions of failure and service time on performance measures including the distribution of collisions and the average number of failures of an arbitrary customer. Results were obtained by a self-developed simulation program based on SimPack [20] toolkit that constitutes the base of our model and provides utilities to create a working simulation from a model description.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

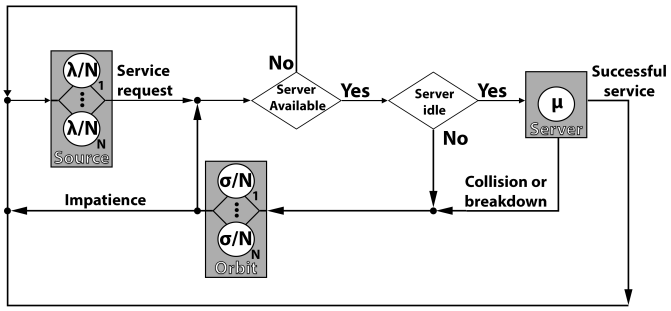


Fig. 1. System model

Graphical illustrations are proven to highlight the gathered results to exhibit the effect of the utilized distributions.

II. System model

Figure 1 presents the considered model as a type of $M/M/1//N$ finite-source retrial queueing system with an unreliable server, impatient customers in the orbit, and blocking. The finite-source possesses N customers in which each of them generates calls towards the system and this inter-request time follows exponential distribution with parameter λ/N. If the service unit is idle the service of an incoming request begins instantly because there is no waiting queue. This variable is also exponentially distributed with parameter μ. Alternatively, when the server is busy with a call, an arriving customer brings about a collision which results in moving both customers to the orbit, which is a virtual waiting room. Here, the requests spend an exponentially distributed time before launching another attempt to arrive at the service facility. Every customer is characterized by the property of impatience meaning that after a certain amount of exponentially distributed random time with parameter τ it may decide to leave the system earlier without obtaining the appropriate service. The service unit is supposed to break down according to gamma, hypo-exponential, hyper-exponential, Pareto and lognormal distribution selecting the parameters having the same mean value. Promptly, the restoration period initiates upon the breakdown that is also an exponential random variable with parameter γ_2. In the case of failure occurrence, the service of a customer terminates and it is delivered to the orbit. Throughout this period incoming customers are not capable of entering the system because they are rejected and get back to the source. This is the so-called blocking. Every mentioned variable consisting of this model is assumed to be independent of each other.

III. Simulation environment

We used the simulation approach to acquire all the desired performance measures. As we wanted to depict special measures and some of them are difficult to give exact formulas we chose to develop our own simulation package based on SimPack as it is mentioned earlier. Sometimes the state space of the Markov chain is so huge that makes it impossible to solve the derived steady-state equations which are also valid for the available software packages. These are capable of performing an analytical evaluation of complex systems if the variables in the model construction follow exponential distribution. In the case of simulation, we can utilize other distributions as well which gives us the opportunity to perform a sensitivity analysis. The estimation is carried out by applying a statistical package in which the method of batch means is used. In short, n observations occur in every batch and the useful run contains enough batches to accomplish a valid estimation. The batches should be long enough and approximately independent of each other to increase the accuracy of the achieved results. Among the confidence interval techniques for a steady-state mean of a process, this is the most common technique and the following works [21],[22] include more detailed information about the process. The simulations are performed with a confidence level of 99.9%. The relative half-width of the confidence interval required to stop the simulation run is 0.00001.

IV. Simulation results

A. Different distributions of service time of the customers

In this section, we worked with different distributions of service time of the customers with an unreliable service unit to investigate the effect on some performance measures. Observing various simulation runs with different parameter settings, we selected the scenario with the most interesting results and Table I shows that one. γ_0 and γ_1 are the parameters of failure time when the server is busy and idle, accordingly. In [23],[24] similar systems were treated by an asymptotic method where N tends to infinity that is why we use rates λ/N and σ/N. Table II demonstrates the values of parameters in the case of every distribution including hyper-exponential, gamma, lognormal, and Pareto. These distributions are suitable to perform a valid comparison because we can select parameters in order the squared coefficient of variation would be greater than one meaning that both the mean value and variance are the same. About the fitting process, which is needed to be done, can be found in [25].

TABLE I
Numerical values of model parameters

N	γ_0	γ_1	σ/N	τ
100	0.05	1	0.05	0.0005

TABLE II
Parameters of service time of the customers

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.0625$ $\beta = 0.0625$	$p = 0.4697$ $\lambda_1 = 0.939$ $\lambda_2 = 1.061$	$\alpha = 2.031$ $k = 0.508$	$m = -1.417$ $\sigma = 1.683$
Mean	1			
Variance	16			
Squared coefficient of variation	16			

The steady-state probability of a customer in the orbit is presented in Figure 2 when the arrival intensity equals 0.01, which shows the probability that exactly i customers reside in the orbit in a random time. Having the same mean and variance, the obtained results surprisingly vary significantly from each other which is especially true in the case of the gamma and the Pareto distribution. This figure ensures us that

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

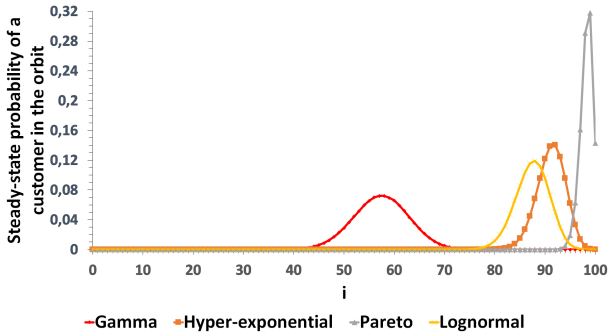


Fig. 2. Distribution of number of customers in the orbit, $\lambda/N = 0.01$

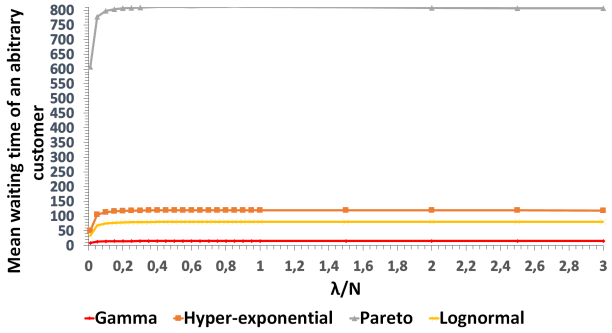


Fig. 3. Mean waiting time vs. arrival intensity

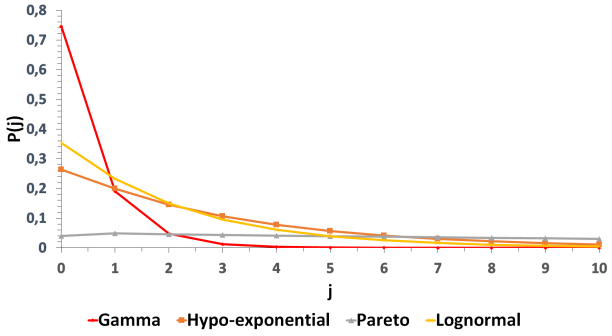
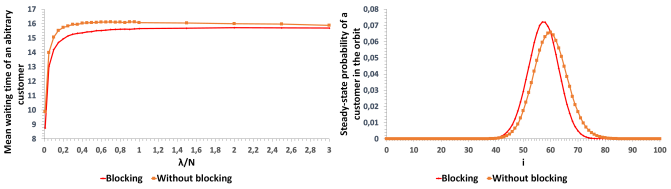


Fig. 4. The distribution of the number of collisions of an arbitrary customer

the selected distribution has a great impact on the operation of the system. The shape of all curves seems to correspond to Gaussian distribution.

Figure 3 shows the mean waiting of an arbitrary customer in the function of the arrival intensity. The same tendency occurs here regarding the results and a huge gap can be observed among the curves which are quite interesting. When Pareto distribution is applied customers spend averagely much more time in the orbit than in the case of other distributions. Because the source is finite the interesting maximum property characteristics appear despite the increasing arrival intensity which is true for every graph.

Figure 4 illustrates the number of collisions an arbitrary customer suffers under service during its residence in the system. Taking a closer look at the graphs the service of the customers is not interrupted by an arriving request most of the time when gamma distribution takes place. This can not be said to the others especially for Pareto distribution where customers undoubtedly experience more collision comparing to the others. Near resemblance is noticeable to geometric distribution scrutinizing closely the results because the division of the consecutive probabilities is almost the same.



(a) Mean waiting time vs arrival intensity (b) Distribution of number of customers in the orbit

Fig. 5. The effect of blocking

Two graphs are presented in Figure 5, which depicts the effect of blocking on two performance measures. Examining closely the curves the expected behaviour is seen, lower values of mean waiting time, and less number of customers are experienced during running the code. However, the same characteristics remain in the case of without blocking as the mean waiting time has maximum value despite the increasing arrival intensity and the number of customers seems to follow normal distribution.

B. Different distributions of failure time of the server

In this scenario, after performing a sensitivity analysis on the service time of a customer we decided to carry out another one on the failure time of the service unit. Basically, we use the same parameter setting that Table II,Table I exhibit but these are the parameters of failure time. The service time of the customers is exponentially distributed with parameter μ. We analyze the same performance measures as in the previous section to see how the different distribution has an influence on the operation of our model.

Figure 6 displays what the probability is if exactly i customers are located in the orbit. The figure is completed with the reliable case when server failures do not occur to see the impact of this feature as well. The difference is not as huge as in the previous section but it is still presented among the utilized distributions and surprisingly we found the lowest mean value at gamma distribution. But similarly to Figure 2 the obtained results indicate that the distribution of this variable probably is normally distributed.

How the mean waiting time starts to increase and get stagnant after a while can be seen in Figure 7 in the function of arriving intensity. Of course, the mean waiting time is the lowest by far when there is no breakdown but interestingly the disparity among the other graphs is not relatively high except gamma distribution the gathered results are almost the same. Using this parameter setting still results in the phenomenon of maximum property characteristics, after a certain arrival intensity value the mean waiting of an arbitrary customer decreases.

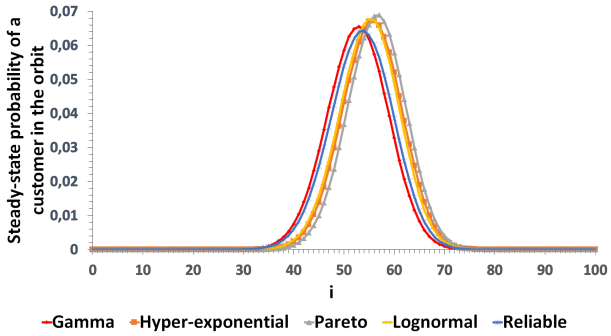


Fig. 6. Distribution of number of customers in the orbit, $\lambda/N = 0.1$

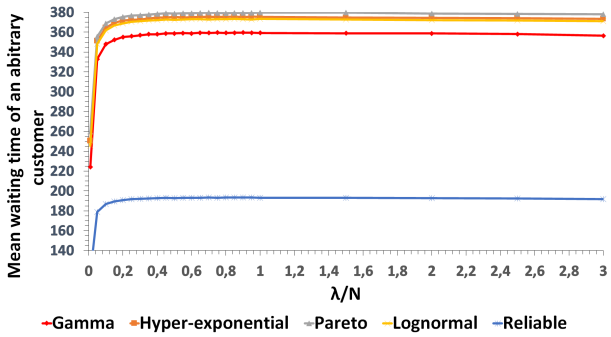


Fig. 7. Mean waiting time vs. arrival intensity

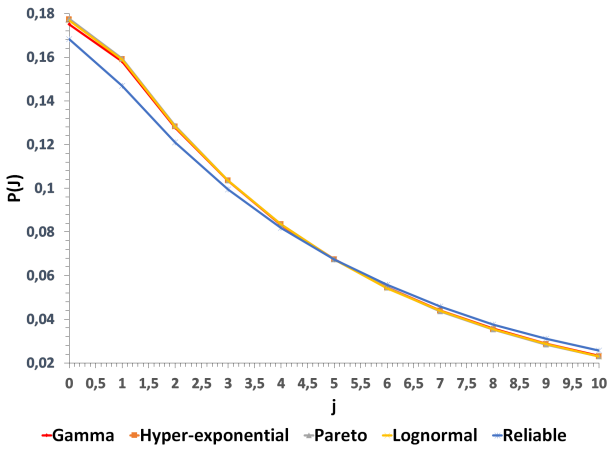
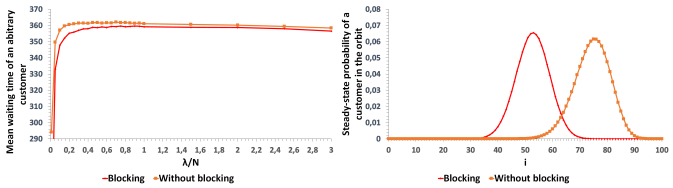


Fig. 8. The distribution of the number of collisions of an arbitrary customer

The next figure (Figure 8) represents the distribution of the number of collisions of an arbitrary customer showing that what is the probability of a customer suffering exactly j collision under service. The lines nearly overlap each other illustrating that concerning this aspect there is no distinction among the cases. When the service unit is reliable customers are less exposed to go through collision but other than the curves are almost identical and reveal signs of decreasing tendency as the number of collisions increases.

Lastly, the comparison of the effect of blocking is shown in Figure 9 displaying both the mean waiting time of an arbitrary customer and the number of customers in the orbit.



(a) Mean waiting time vs arrival intensity (b) Distribution of number of customers in the orbit

Fig. 9. The effect of blocking

Indubitably, in the case of blocking higher values can be found and in this scenario, the difference is noticeable higher as in the previous section, but naturally, fewer customers could stay in the system if frequent outages take place.

V. Conclusion

A finite-source retrial queueing system is included with a non-reliable server, impatient customers and collisions in this work. In two scenarios we carry out a sensitivity analysis on two different random variables using numerous distributions to investigate the effect on some performance measures. Running our simulation program, graphical illustrations represented that in both scenarios the obtained values significantly differ from each other showing the importance of choosing a distribution having the same mean and variance when the squared coefficient of variation is greater than one. We also examined the effect of reliable operation demonstrating how remarkably changes the system behaviour compared to cases including random breakdowns. We also studied that applying the feature of blocking will lower the mean waiting time and the mean number of customers in the orbit. In the future, the authors want to continue their research work including more components in the model like catastrophic feature, more service units, or including other distributions.

Acknowledgment

The research was supported by the Thematic Excellence Programme (TKP2020-IKA-04) of the Ministry for Innovation and Technology in Hungary.

References

[1] J. Artalejo and A. G. Corral, *Retrial Queueing Systems: A Computational Approach*. Springer, 2008.
[2] G. Falin and J. Artalejo, "A finite source retrial queue," *European Journal of Operational Research*, vol. 108, pp. 409–424, 1998.
[3] D. Fiems and T. Phung-Duc, "Light-traffic analysis of random access systems without collisions," *Annals of Operations Research*, pp. 1–17, 2017.
[4] A. Gómez-Corral and T. Phung-Duc, "Retrial queues and related models," *Annals of Operations Research*, vol. 247, no. 1, pp. 1–2, 2016.
[5] J. Kim and B. Kim, "A survey of retrial queueing systems," *Annals of Operations Research*, vol. 247, no. 1, pp. 3–36, 2016.
[6] K. Rakesh and S. Sapana, "Transient performance analysis of a single server queuing model with retention of reneging customers," *Yugoslav Journal of Operations Research*, vol. 28, no. 3, pp. 315–331, 2018.
[7] C. Kim, S. Dudin, A. Dudin, and K. Samouylov, "Analysis of a semi-open queuing network with a state dependent marked markovian arrival process, customers retrials and impatience," *Mathematics*, vol. 7, no. 8, pp. 715–734, 2019.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

[8] A. Kvach and A. Nazarov, "Sojourn time analysis of finite source markov retrial queuing system with collision," in *Information Technologies and Mathematical Modelling - Queueing Theory and Applications*, A. Dudin, A. Nazarov, and R. Yakupov, Eds. Cham: Springer International Publishing, 2015, pp. 64–72.

[9] A. Kvach, "Numerical research of a Markov closed retrial queueing system without collisions and with the collision of the customers," in *Proceedings of Tomsk State University. A series of physics and mathematics. Tomsk*, ser. Materials of the II All-Russian Scientific Conference, vol. 295. TSU Publishing House, 2014, pp. 105–112, (In Russian).

[10] A. Kvach and A. Nazarov, "Numerical research of a closed retrial queueing system M/GI/1//N with collision of the customers ," in *Proceedings of Tomsk State University. A series of physics and mathematics. Tomsk*, ser. Materials of the III All-Russian Scientific Conference, vol. 297. TSU Publishing House, 2015, pp. 65–70, (In Russian).

[11] A. Nazarov, A. Kvach, and V. Yampolsky, *Asymptotic Analysis of Closed Markov Retrial Queuing System with Collision*. Cham: Springer International Publishing, 2014, ch. 1, pp. 334–341.

[12] V. I. Dragieva, "Number of retrials in a finite source retrial queue with unreliable server." *Asia-Pac. J. Oper. Res.*, vol. 31, no. 2, p. 23, 2014.

[13] N. Gharbi, B. Nemmouchi, L. Mokdad, and J. Ben-Othman, "The impact of breakdowns disciplines and repeated attempts on performances of small cell networks," *Journal of Computational Science*, vol. 5, no. 4, pp. 633–644, 2014.

[14] A. Krishnamoorthy, P. K. Pramod, and S. R. Chakravarthy, "Queues with interruptions: a survey," *TOP*, vol. 22, no. 1, pp. 290–320, 2014.

[15] J. Roszik, "Homogeneous finite-source retrial queues with server and sources subject to breakdowns and repairs." *Ann. Univ. Sci. Budap. Rolando Eötvös, Sect. Comput.*, vol. 23, pp. 213–227, 2004.

[16] J. Sztrik, B. Almási, and J. Roszik, "Heterogeneous finite-source retrial queues with server subject to breakdowns and repairs," *Journal of Mathematical Sciences*, vol. 132, pp. 677–685, 2006.

[17] J. Wang, L. Zhao, and F. Zhang, "Performance analysis of the finite source retrial queue with server breakdowns and repairs," in *Proceedings of the 5th International Conference on Queueing Theory and Network Applications*. ACM, 2010, pp. 169–176.

[18] F. Zhang and J. Wang, "Performance analysis of the retrial queues with finite number of sources and service interruptions," *Journal of the Korean Statistical Society*, vol. 42, no. 1, pp. 117–131, 2013.

[19] Á. Tóth, T. Bérczes, J. Sztrik, and A. Kvach, "Simulation of finite-source retrial queueing systems with collisions and a non-reliable server," in *International Conference on Distributed Computer and Communication Networks*. Springer, 2017, pp. 146–158.

[20] P. A. Fishwick, "Simpack: Getting started with simulation programming in c and c++," in *In 1992 Winter Simulation Conference*, 1992, pp. 154–162.

[21] E. J. Chen and W. D. Kelton, "A procedure for generating batch-means confidence intervals for simulation: Checking independence and normality," *SIMULATION*, vol. 83, no. 10, pp. 683–694, 2007.

[22] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*. McGraw-Hill Education, 1991.

[23] A. Nazarov, J. Sztrik, and A. Kvach, "A survey of recent results in finite-source retrial queues with collisions," in *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*. Springer, 2018, pp. 1–15.

[24] A. Nazarov, J. Sztrik, A. Kvach, and T. Bérczes, "Asymptotic analysis of finite-source M/M/1 retrial queueing system with collisions and server subject to breakdowns and repairs," *Annals of Operations Research*, vol. 277, no. 2, pp. 213–229, Jun 2019.

[25] J. Sztrik, Á. Tóth, Á. Pintér, and Z. Bács, "Simulation of finite-source retrial queues with two-way communications to the orbit," in *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*, A. Dudin, A. Nazarov, and A. Moiseev, Eds. Cham: Springer International Publishing, 2019, pp. 270–284.

Conceptual analysis of single and multiple path routing in MANET network

Ibrahim Alameri
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
Jabir ibn Hayyan Medical University
Najaf, Iraq
st61833@upce.cz

Jitka Komarkova
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
jitka.komarkova@upce.cz

Mustafa K. Ramadhan
Faculty of Computer - Techniques Engineering
Al-Safwa college university
Karbala, Iraq
mustafa.ramadhan@safwa.edu.iq

***Abstract*—Mobile ad-hoc network (MANET) has attracted the attention of networking industries owning to their desirable characteristics such as multi-hop routing, self-configuration, self-healing, self-managing, reliability, and scalability. Routing over wireless mobile networks is a critical problem due to the dynamic nature of the link qualities, even when nodes are static. A key challenge in MANETs is the need for an efficient routing protocol that establishes a route according to certain performance metrics related to the link quality. The routing issue in MANETs is generally concerned with finding a good path between the source and the destination pairs. Based on that, there is a demand for the development of a high throughput routing protocol. The impact of a single-path routing protocol and a multi path routing protocol on the performance of MANETs is required to be investigated. In this work, a performance comparison in terms of throughput, packet delivery, routing overhead, and end to end delay of well-known routing protocols such as AODV, AOMDV, and OLSR using network simulator version 2 (NS-2) has been introduced. The simulation results of this work show that the single-path AODV protocol out performance the multi path OLSR and AOMDV protocols in terms of throughput and packet delivery ratio. In addition to that, the single-path routing protocol presents less routing overhead in comparison to the AOMDV and OLSR. While the OLSR and AOMDV demonstrate a relatively better end to end delay in comparison to the AODV protocol.**

***Keywords*—MANET, Routing protocols, AODV, AOMDV, OLSR, routing metrics**

I. Introduction

Information Communication Technologies (ICT) has laid down the foundation of the New World Order (NWO). Since its birth to 5th Generation (5G) [1], wireless networks have shown instrumental growth and development to solve real-world problems. We encounter many different types of wireless networks on day to day basis, for example, Bluetooth, Wireless Local Area Networks (WLAN), 4th Generation (4G) mobile networks, etc. One prime reason for the availability of several wireless technologies is the research and development (R D) which has taken place in this domain [2].

Wireless communication networks are divided into two types (i) infrastructure-based and (ii) infrastructure-less wireless networks [3]. Infrastructure-less wireless networks are also referred to as ad-hoc networks. A literature study shows that several ad-hoc networks are presented, like Vehicular ad-hoc Network (VANET), Mobile ad-hoc Network (MANET), etc. Although most ad-hoc networks share common characteristics and challenges, they also differ in few aspects [4].

In this paper, our core objective is the study of topology-based route selection interventions for mobile ad-hoc networks. This study focuses on single-path and multi-path protocols. A mobile ad-hoc network (MANET) is a brand of no infrastructure-based wireless networks that comprises nodes that can freely change their location and operate in a decentralised manner [5]. The decentralised nature primarily relies on the self-configuration and autonomously of nodes [6]. The self-configuration (also known as auto-configuration) feature of MANET eliminates the need for an administrator to configure them. These properties of MANET differentiate it from the traditional infrastructure-based wireless network. A comparison of the infrastructure-based and infrastructure-less wireless network is presented in Table I [7]. Table 1, presents several striking features of infrastructure-less mobile networks. However, such networks have to deal with different challenges also, such as:

- Limited Range Transmission: Nodes in mobile ad-hoc networks can only communicate with nearby nodes (up to only a few meters). Because mobile nodes cannot carry advanced high-end radios (transceiver) and large antennas.
- Limited Energy: Since mostly mobile nodes are battery operated (do not have a direct electric supply). Due to this reason, power is a significant challenge in MANET. In literature, many researchers have proposed energy-efficient routing protocols for MANET [8]–[10].
- Mobility: In MANET, nodes/devices can move freely anywhere, which leads to network topology changes more frequently. Due to network topology changes, efficient routing becomes a significant challenge.
- Routing: In MANET network can change dynamically at run time, due to which efficient routing interventions are needed to operate in the network.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE I: Comparison of infrastructure-based VS infrastructure-less wireless networks.

Feature	Infrastructure-based	Infrastructure-less
Cost	It needs heavy investment to lay down the required infrastructure such as base station or mobile towers etc.	It is infrastructure-less networks
Deployment	Complex deployment sometimes require experts to design the network	No deployment cost or planning required.
Decentralized Control	Infrastructure-based networks require central server / manager (generally known as Access-Point (AP) to control the network e.g. nodes admission etc.	Adhoc networks do not require any central server for the their management.
Dynamic Nature	Generally less dynamic in nature as nodes communication remain confines to AP	In MANET nodes can become part of the network or leave any time.

- Quality of Service (QoS) or Service Quality: Due to the number of reasons presented above, providing a satisfactory extent of service quality becomes a challenge to deliver. Mobile ad-hoc Networks (MANETs) are being used in many situations where infrastructure-based wireless networks generally fail to deliver or operate efficiently.

Due to the unique features of the MANET, this type of the network has been employed in several services such as video streaming, online shopping, and mobile surgery. On the other hand, MANET could be employed in emergency relief environments owning to their low cost of implementation. In literature review, it has been realised that the employed routing protocol play a crucial role to improve the QoS and scalability of MANET. Several ad-hoc routing protocols have been introduced to transmit information using a single-path in the last few years. In contrast, another type of transmitted data builds on the concept of multi-path. These technologies enable a multi-path to generated among a source and a destination used to transfer information. Anyway, the optimal route is chosen based on the various performance measures known as metrics such as hop count, distance from the source to the destination, remaining energy, etc [11].

The objective of this work is to evaluate the impact of a single-path routing protocol and a multi-path routing protocol on the performance of the ad-hoc networks. In this study, three well known protocols which are AODV, AOMDV, and OLSR have been employed for different network scenarios of different network densities. The AODV represents a single-path routing protocol while the other two protocols represent the multi-path routing protocol.

The rest of this paper is organised as follows: The second section of this paper will present the routing protocols overview. Section III presents the literature review to summaries the key findings of previous research work published in this domain and their limitations in the context of MANET. The simulation Methodology and simulation approach assesses single-path and multi-path routing protocols are discussed in Section IV. Last but not least, simulation findings are rendered and discussed in Section V, and followed by the conclusion in section VI.

II. MANET ROUTING PROTOCOLS OVERVIEW

Routing is when nodes determine the optimally best route/path to forward the packets toward the destination. If a device receives or nodes receives a packet, and the target destination is not an actual destination for which packet was sent, it must route it. All intermediate nodes in ad-hoc networks need to take routing decision for each packet by routing table lookup. Routing protocols populate the routing table. Routing protocols play a significant role in MANET, especially when nodes are mobile and network dynamics change at run-time. In literature, routing interventions are grouped into different classes based on how they operate, build and maintain the routing table. The taxonomy of routing protocols is presented in Fig. 1. Fig. 1, shows an overview of a few set of protocols sharing standard functionality.

The following subsection below critically shows each category of network routing briefly.

A. Position-based Routing Protocols

In this type of routing, each node knows its current position in the grid. This position information is made available to the node using the Global Positioning System (GPS) sensor embedded on the node [12]. In position-based routing protocols, senders are also aware of the position of the prime destination. Senders use location services advertised previously by entire nodes that exist in the network. These routing protocols are also classified into three groups such as reactive, predictive and hybrid.

B. Energy-based Routing Protocols

In MANET and other similar arrangements, nodes are generally battery operated, due to which nodes energy consumption have a significant effect on network lifetime [13]. In literature, researchers have produced different energy-aware routing protocols that make routing decisions based on energy consumption, remaining energy left, etc. In such scenarios, devices are aware of other nodes' utilized energy, which is part of the network and network routing decision is supported with this. For example, nodes having less residual energy will be avoided to become part of forwarding.

C. Heterogeneity-based Routing Protocols

Different types of wireless mobile networks need to collaborate in some scenarios, such as MANET, VANET, etc., as the dynamics of these networks are different. Therefore

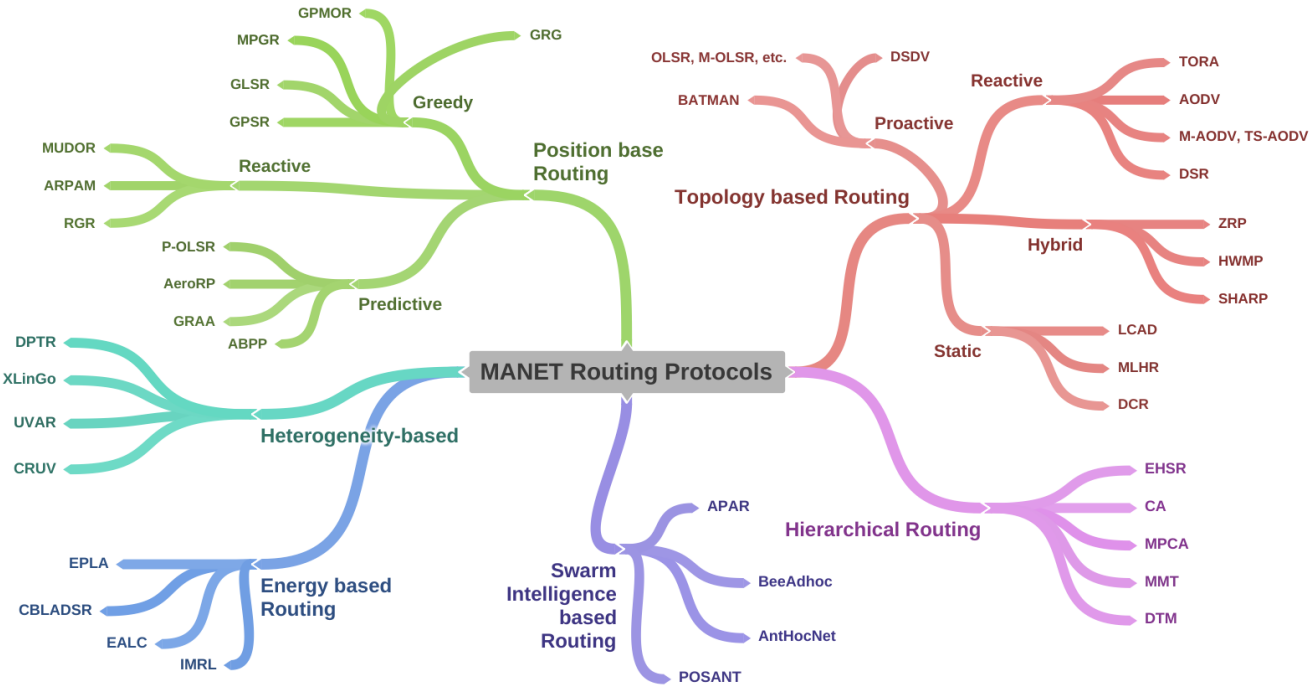


Figure 1: Moile ad-hoc networks - Protocol hierarchy.

researchers have proposed different routing approaches that can be operated in similar conditions as shown in Fig. 1.

D. Swarm Intelligence-based Routing Protocols

Swarm intelligence-based routing protocols are generally inspired by the biological behaviour of different animals, birds, insects, etc [14]. Therefore, they are also known as the bioinspired routing strategy. Several bio-inspired or swarm-intelligence based routing strategies have been proposed in the past research work, e.g. Ant Colony Optimization (ACO) algorithm, BeeAdhoc routing protocol etc [15]. These routing protocols have produced outstanding results in some scenarios.

E. Hierarchical Routing Protocols

Fall under this group usually forms clusters. A cluster can be defined as 'nodes reside in closed proximity' build a cluster [16]. A cluster is formed following a process. Each nearby node participates in the cluster formation process, which is known as the setup phase. During the setup phase, a cluster head is elected by a process known as election. The elected cluster head will remain the leader in that cluster for some fixed number of rounds. When a certain number of rounds have passed, nodes in the proximity go again in the setup phase to elect a new cluster head. In this paper, we will focus on and discuss in detail topology-based protocols. All protocols which build routing table based on sharing topology information are included in this group.

F. Topology-based Routing Protocols

Topology-based routing techniques make use of network topology data in order to build a routing table. This class of routing approach is most widely utilized in ad-hoc networks. Topology-based protocols are divided into the following groups: proactive, reactive, hybrid routing, and static [17]. In mobile ad-hoc networks, a static technique to build routing table is not often used and recommended due to its static nature.

1) Proactive Routing Protocols: A complete routing table is built by all proactive routing protocols in advance at the start of operation. This table aims to establish and maintain a path to every destination node alive in the network/topology [18]. All devices/nodes that speak any proactive routing protocol start to exchange network information initially, which is also known as the setup phase like the Lower Energy Adaptive Clustering Hierarchy (LEACH) protocol [19]–[21]. In the setup, nodes exchange routing messages and calculate the best next hop for every destination. Nodes exchange complete routing table periodically and at each change occurred in the topology due to which it consumes heavy network bandwidth and computational resources of the nodes. However, each node in the network carries a complete 'map' for the network; therefore, the best route to any target is readily available, which reduces route discovery for each packet. It is advised that these protocols should not be operated in large networks or where topology changes more frequently.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Optimized Link State Routing (OLSR)
Link state routing approaches are entirely different from traditional distance vector-based routing options. Link state routing protocol shares the status of their link (cost) with their neighbours. OLSR is also an LS routing strategy that follows the same school of thought [22]. The OLSR routing protocol is a multi-path [23], [24]. OLSR do not broadcast anything. All hello messages are shared only with the neighbours periodically. In case of topology changes, they exchange topology change notification (TCN) only with their neighbours. Furthermore, to reduce routing communication, OLSR selects Multipoint Relay (MPR) nodes from its neighbours. MPR is also responsible for forwarding such messages to other neighbours/peers. A significant amount of network bandwidth is reduced by using this approach.

2) Reactive Routing Protocols: As opposed to proactive methods, reactive routing techniques start route lookup or discovery process when a node receives a request. That is why they are known as reactive. There is no setup phase like in proactive routing protocols, nor they maintain routes for each target. Due to this nature, they are referred to as 'on-demand routing protocols. When a route for a particular target is discovered, then the node only keeps that route for a limited time in the routing table. When no more packet is received for the target node for a specific duration, route entry is removed from the table. The benefits of the reactive protocol are that they generate less routing load, and network size is scalable, meaning they can be used in significant typologies. Reactive protocols may have unpredictable delays.

Adhoc On-Demand Distance Vector (AODV)
AODV is a reactive routing protocol that incorporates DSDV and DSR protocols [25], [26]. It shows adaptive behaviour at each hop which is known as Hop-by-Hop nature. Hop by hop feature in AODV is adapted from the DSR protocol. It utilizes the periodic exchange of messages technique from the DSDV protocol. In the beginning, it initiates route discovery to create a routing path. It tries minimum hop count path to be selected. This feature significantly reduces the overhead and minimizes network congestion. To maintain established links up to date, it exchanges update messages.

Ad-hoc On-Demand Multipath Distance Vector (AOMDV)
AOMDV is a multi-path routing protocol [23], [24], [27]. Multi-path routing protocols discover and maintain multiple paths for the target node [28], [29]. The purpose of keeping multiple paths is to avoid or reduce frequent discovery of routes. AOMDV is based on AODV reactive routing protocol.

3) Hybrid Routing Protocols: Hybrid methods choose the best attributes of both groups, i.e. proactive and reactive. These interventions curtail the limitations and overheads of reactive proactive strategies. The Zone Routing Protocol (ZRP) is a hybrid routing strategy that partitions the area into different segments known as 'zone' [30]. Intra-zone path selection is performed with the help proactive routing approach, and inter-zone is achieved by utilizing reactive interventions.

III. LITERATURE REVIEW

This chapter began by describing a previous research study on routing limitations in mobile ad-hoc networks. Relay routed-DSR is officially implemented to manage data packets effectively. To collect information from neighbour nodes, this novel routing strategy employs a broadcasting mechanism. During the flooding process, redundant paths are discovered, increasing overhead in the network [31]. Preemptive-DSR (PDSR) protocol predicts connection failures, but the mechanism is slow and costly. Due to low signal strength, P-DSR sets a threshold, and warning signals are sent to source nodes [32]. The new variant of AODV, ad-hoc On-Demand Multipath Distance Vector (AOMDV) routing protocol, is the most widely used multi-path routing strategy. The new routing method avoids connection loss and relies on a minimal hop count [33]. Fibonacci multi-path load balancing and multiple AODV are considered to deliver data packets to conserve energy in nodes [34]. However, AODV routing vulnerabilities enable many typical network attacks such as a black hole, grey hole, wormhole, etc., which can easily access data packets and set up malicious nodes within the network [35]. Fig 2, shows a routing study of the AOMDV protocol. Attackers send data packets in a continuous stream to increase the number of false information in the network, directly affecting the system's dynamics [36]. Context-aware routing improves node energy levels, introducing a novel solution that will help to secure channel links. Adaptive routing decision helps to monitor routes [37].

AOMDV Routing Protocol
Final Destination
Sequence Number
Broadcast Hop count
Route List
Selected Path
Expiration Timeout

Figure 2: AOMDV protocol.

The idea of meta-heuristics, which improve local monitoring in mobile ad-hoc networks, is computed by mobility aware-Termite [38]. Therefore, extended ad-hoc networks, finding the exact position using GPS-based knowledge predictive-OLSR, show exponential improvement [39], [40]. In a recent study, authors developed an ad-hoc routing protocol to conserve energy at every node [41]. Single-path or multi-path routing strategies can be used in mobile ad-hoc networks. Single-path routing is recommended for forwarding all data packets over the route. However, some significant problems with single-path routing have been found, including an increase in end-to-end delay and slower route discovery time. As a result of

these reasons, the single-path routing fails to perform tasks in all environments. Multi-path routing protocols select several routes from source to target. Compared to a single-path, specific metrics such as delay, bandwidth, and throughput have improved [42], [43]. Response surface optimization finalizes the optimal response time during data analysis in AODV and AOMDV [44]. The working concept of AODV routing is visualized in fig 3. While sending an AODV message from one node to another, two steps are usually followed: (i) path exploration and (ii) route repair. For route discovery and maintenance, message data comes in four forms: reply, request, error, and hello packet [44]. Multi-path routing protocols have several advantages. Multi-path routing protocol tends to reduce end to end delay. These protocols utilize network bandwidth more efficiently as compared to single-path routing protocols.

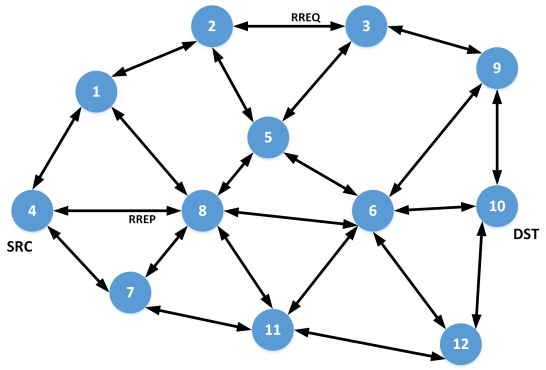


Figure 3: AODV routing protocol.

Since authors maintain multiple paths to the same destination, the traffic will be forward via an alternate path/route without adding a delay factor if a path becomes unavailable. They can perform better network load balancing, which causes homogenized energy consumption in the network. Multiple path/route selection methods also aid in minimizing the packet drop rate of the network. Besides many advantages of multi-path protocols, they also introduce few limitations. Generally, they use the broadcast mode of communication. At times they suffer from good end to end multiple path discovery issues. Since from source to target node, they maintain more than one path; therefore, the data duplication problem also arises in few cases. Limited control information distribution and longer routes. The summary of the pros and cons of multi-path routing interventions are presented in table II.

TABLE II: Multi-path routing protocols advantages and disadvantages.

Advantages	***Limitations***
Reduction in End-to-End Delay	Broadcasting Issues
Efficient bandwidth utilization	Insufficient path discovery
Taking Backup of the system	Data Packet Duplication
Load distribution in entire network	Longer Routes
Less Packet Loss	No Control Information distribution

IV. Materials and Methods

The current section is concerned with the methodology used for this study. The presented paper uses a simulation approach and parameters configured for the successful execution of the work. Overall, the study highlights the need for network simulator 2 (NS-2) to perform the simulation. NS-2 is an open-source discrete event simulator that is widely used in research [45]. The core engine of NS-2 is built in C++ programming language, and the front system, which is used to create simulation topology, uses Object-Oriented Tool Command Language (OTCL). We used the latest version of NS-2, which is 2.35. It generates two types of trace files which are (i) simulation trace and (ii) nam trace. The simulation trace file is further used for data analysis. In contrast, the Nam trace file can be fed into the network animator (Nam) utility to view how the simulation is carried out. Fig. 5, shows how simulation is carried out using NS-2. The MATLAB programing language has been used to generate the graphics and analyse the trace file by formulas and MATLAB. We created three different testbeds or scenarios for our study. In all three scenarios, the basic simulation parameters are the same. However, the main difference in all three testbeds is the number of nodes each.

Scenario 1: In this scenario, we configured our simulation topology as mentioned in Table III. The number of nodes in this scenario was configured to 100.

Scenario 2: In this scenario, we configured our simulation topology as mentioned in Table III. The number of nodes in this scenario was configured to 150.

Scenario 3: In this scenario, we configured our simulation topology as mentioned in Table III. The number of nodes in this scenario was configured to 200.

For each simulation scenarios, performed ten iterations of Simulation. The purpose of this repetitive exercise is to reduce the statistical anomalies/discrepancies in the result. Therefore, 30 total rounds/iterations of the Simulation are carried during this study. The time duration of the Simulation contributes towards an essential role in studying the behaviour of any phenomenon. In Literature, found that most researchers used a low time window. Therefore, the presented paper performed the Simulation for up to 4 minutes or 240 seconds within various network load. The coverage area, also known as the simulation area, also plays a significant role in the study. In order to accommodate hundreds of nodes, built a large coverage area such that nodes can also move freely and easily. The network area for each scenario was configured as 1500m x 1500m in our study. The mobility of nodes affects the performance of a network drastically. Presented paper used Random Way Point (RWP) mobility model in our study. RWP is the most used movement pattern. The velocity (speed) of nodes also play a crucial role. In our topology, all nodes can move with up to 20 $m.s^{-1}$ velocities. As discussed in Section I, we included single and multiple path routing protocols.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

This paper suggested to use AODV, AOMDV and OLSR routing protocols. Further, as mentioned above, the Simulation was performed ten times by altering the number of nodes in each iteration. However, the relevant parameters and topology design are summarized in Table III.

TABLE III: Network simulation scenario.

Parameters	***Value***
Simulator software	NS-2
Simulation Time	240 seconds
Area of work	1500 m^2
Nodes Speed	20 m/s
Transport Layer Protocol	UDP
Application Type	Constant Bit Rate (CBR)
Number of Nodes	100, 150 , 200
Movement Scenario	RWP
Routing Protocols	AODV, AOMDV, OLSR

A. Simulation trace analysis

During any simulation process, the simulator generates a vast amount of data, as per the configurations in the OTCL TCL scripts. NS-2 can generate data in different trace files for the type of analysis. The two standard trace formats used in NS-2 are (i) simulation trace and (ii) nam trace. The simulation trace file is commonly used for further processing to extract required information that can be used to generate graphical notations and tables. At the same time, the nam trace file is fed into Nam utility which can execute all events chronologically to show the animation type video. The purpose of the nam trace file is to view the simulation in real as shown in Fig. 4.

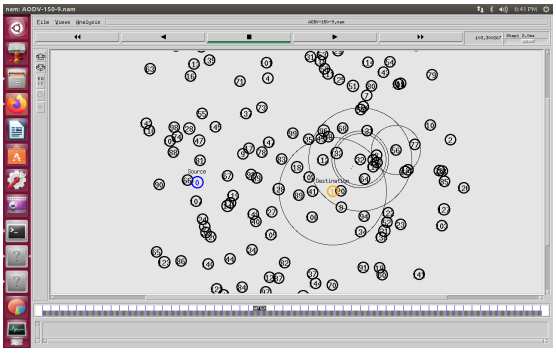


Figure 4: Network topology.

Older NS-2 versions used the old trace format with a different set of issues. One essential weakness in the old wireless trace format is the absence of a type field in the trace file.
Fig. 5, provides an overview of NS-2 structure and code process.
Due to this limitation, programmers have to memorize the field's name as per its position in the trace file. To overcome this problem, NS-2 now comes we a new wireless trace format. The current work used the new wireless trace format. The new wireless trace format has several advantages over the old trace format. Every field in a new format has a type field associated

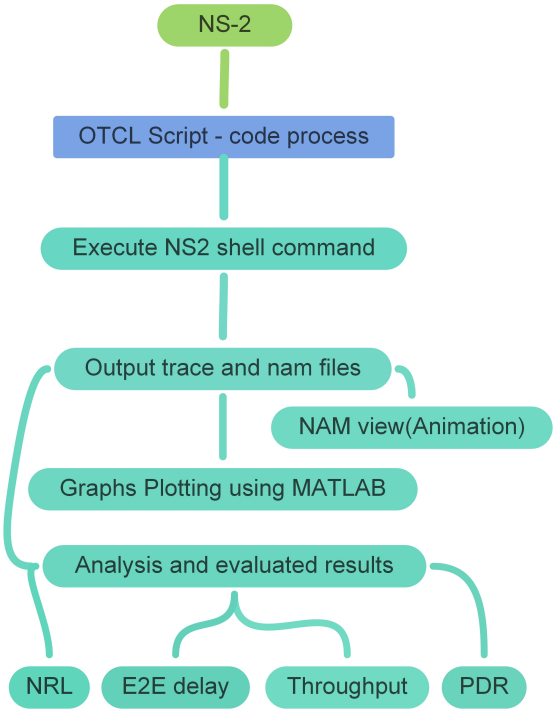


Figure 5: Code process and NS-2 structure.

with it. The benefit of adding a type field in front of a data field is to enhance the effectiveness of data extraction utilities.

V. SIMULATION RESULTS AND DISCUSSION

These rather interesting evaluation results could be due to their relates to critical criteria in MANET networks; as described in the next text. As described in Section I, we studied four different parameters to gauge the effectiveness and performance of each routing protocol.

- Network Throughput.
- End to End Delay (E2E delay).
- Normalized Routing Load (NRL) / Routing Overhead (RO).
- Packet Delivery Ratio (PDR).

A. Network parameters

Throughput: The efficiency of a protocol is measured by its throughput. Higher performance rates indicate optimal results, while low throughput indicates restricted activity in the network [46], [47]. Technically, throughput is described as follows: frames, packets, or bytes efficiently transmitted per unit time are referred to as throughput. Throughput is calculated by using Eq 1.

$$\text{Throughput} = \frac{\Sigma \text{ received packets size}}{\text{time}} \quad (1)$$

E2E delay: Another important metric for network assessment is end to end delay. The time taken by a packet to arrive at its final destination is known as end to end delay [48], [49]. In practice, this means subtracting the time obtained by a data

packet when it arrives at its destination from the initial time and calculated by Eq 2.

$$D_{avg} = Tr_{avg} - Ts_{avg} \tag{2}$$

NRL: Normalized routing load or routing overhead is considered as the overhead. It is defined as "Ratio of network control packets to all delivered packets" [50]. NRL / RO for our simulation is presented Fig. 9, and can be measured by using Eq 3.

$$NRL = \frac{\sum \text{ Routing packets}}{\sum \text{ Packets received}} \tag{3}$$

PDR: Packet delivery ratio is the ratio between packets successfully delivered at the target nodes to the total number of packets sent [51]. The Equation 4, is used to calculate PDR.

$$PDR = \frac{\sum \text{ Number of packets received at destination}}{\sum \text{ Number of packets send by node}} \tag{4}$$

B. RESULT ANALYSIS AND DISCUSSION

In the previous subsection V-A, we presented simulation results. In this section, we will critically analyze them. These are significant results that will describe and discuss sequentially.

1) Network throughput: From Fig. 6, it is evident that the network throughput of the OLSR routing protocol started to drop as the number of nodes decreases. This might be because the OLSR routing protocol selects Multi-Point Relays (MPR) to forward control messages. As node density increases, it had to select several relays in the topology that become the cause of the decreased throughput. The exact figures also show that AODV and AOMDV are less sensitive to changing the number of nodes [52].

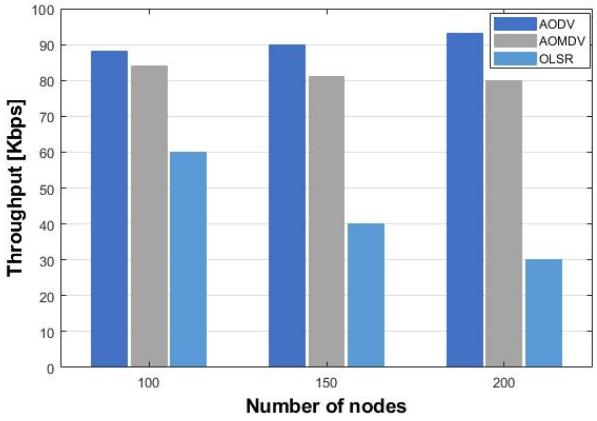


Figure 6: Network throughput based variety nodes number.

Due to the significant importance of the throughput, the presented paper suggested another method known as a standard deviation to double-check the accuracy of results related to the throughput of routing protocols. The table IV, shows the standard deviation study of single and multi path routing protocols for throughput. In the table IV, we calculated the cumulated standard deviation throughput for all three testbeds. Standard deviation measures the amount of deviation from the average figures.

TABLE IV: Standard deviation analysis of routing protocols.

NO. of Nodes	*AODV*	*AOMDV*	*OLSR*
100	83.089	76.752	60.4
150	84.788	78.318	40.1
200	83.911	79.81	30.2
Average	83.92933	78.29333	43.56667
Standard Deviation	0.693735	1.248545	12.57042

These results provide further support for the throughput of protocols. Where the AODV standard deviation for all three cases is far below AOMDV and OLSR protocols. Results suggest that if we further increase or decrease the number of nodes, AODV throughput will be marginally affected. In addition, AOMDV, which is a multi-path routing protocol, also rendered relatively acceptable numbers as compared to OLSR. It suggests that AOMDV protocol throughput will be affected if we change the number of nodes compared to AODV. In contrast, the OLSR has the worst standard deviation value. It means its throughput will be highly affected if we change the number of nodes. Therefore it is not recommended to use in large complex typologies. The Fig. 7, shows the standard deviation of routing protocols.

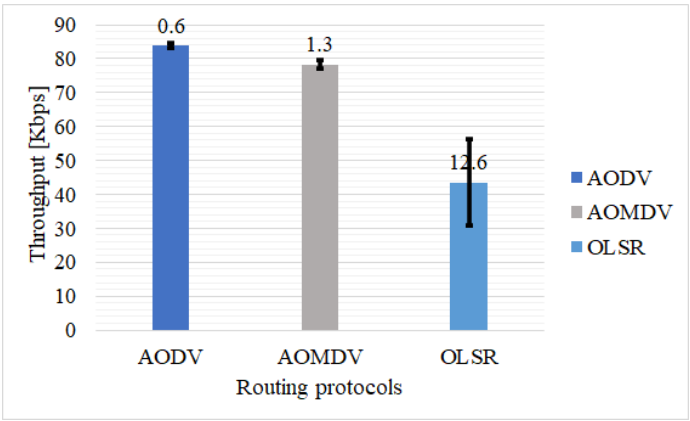


Figure 7: Network throughput based on the standard deviation.

2) E2E delay: Fig. 8 shows a significant increase in the end to end delay for OLSR protocol as node density increases, making it less useful for large and complex topologies. In addition to that, notable improvement in the delay is observed for the AOMDV routing protocol. We observed that initially, AODV and AOMDV delay are higher than OLSR. As the number of nodes increases, they show remarkable improvement.

3) NRL: Fig. 9 proves that reactive routing protocol like AODV has extraordinarily low routing overhead as compared to other two techniques studied in the research, i.e. AOMDV

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

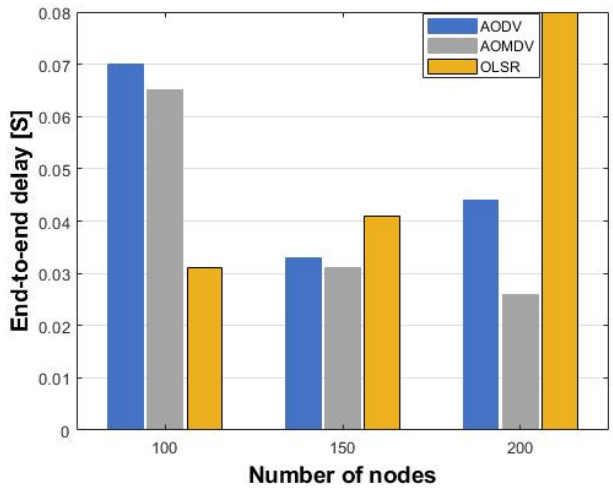


Figure 8: E2E delay of routing protocols.

and OLSR. Rather AOMDV and OLSR have almost similar routing overhead.

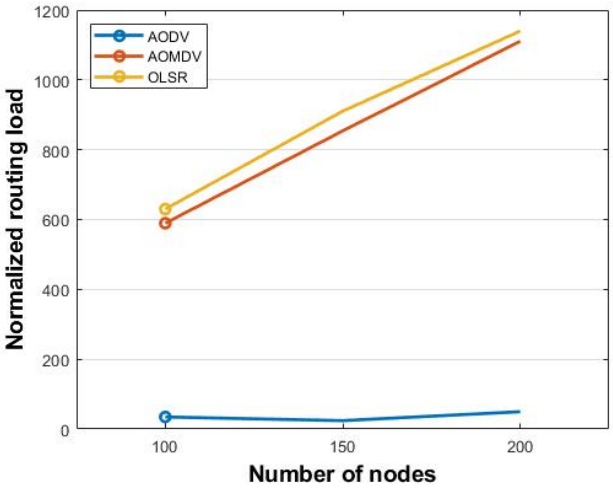


Figure 9: NRL of routing protocols.

4) PDR: The packet delivery ratio graph shows that OLSR has the worst delivery ratio compared to AOMDV and AODV. In addition, another observable pattern that we can analyze, is the effect of an increasing number of nodes. The Fig. 10 shows that AODV is not affected as the number of nodes increases in the network. However, AOMDV and OLSR are affected. The PDR is shown in Fig. 10.

VI. Conclusion

The main goal of this study is to compare and analyse the network performance of different routing protocols in terms of different parameters such as throughput, packet delivery, routing overhead, and end to end delay. The simulation results

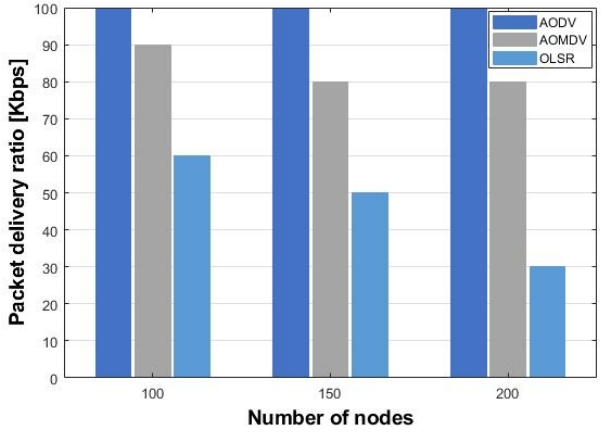


Figure 10: PDR of routing protocols.

reveal that the throughput of the OLSR protocol is highly affected by the varies of the node density in comparison to the AODV and AOMDV protocols. In relation to the throughput, the single-path AODV routing protocol exhibits better network throughput in comparison to the multi path protocols employed in this study. The influence of node density on the performance of the networks has also been investigated in this work. It has been proven in this study as the node density increases, the OLSR protocol reveals a significant increase in the end to end delay. Based on that, it can be stated that the OLSR protocol is not a suitable protocol for high density networks. While the AOMDV shows less end to end delay in comparison to the AODV and OLSR protocols for the examination of the same scenarios. In respect to the routing overhead, the AODV protocol shows less routing overhead comparing to the OLSR and AOMDV protocols.

The future work of this study suggests further investigations of the dynamic behavior of the AODV routing protocol which can lead to a modification to the routing mechanism of the protocol to handle the instability of the link quality.

Acknowledgment

This paper was supported by SGS University of Pardubice project No. SGS_2021_011.

References

[1] X. Ge, H. Cheng, M. Guizani, and T. Han, "5g wireless backhaul networks: challenges and research advances," *IEEE Network*, vol. 28, no. 6, pp. 6–11, 2014.

[2] F. Hu, *Opportunities in 5G networks: A research and development perspective*. CRC press, 2016.

[3] I. A. Alameri and J. Komarkova, "A multi-parameter comparative study of manet routing protocols," in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2020, pp. 1–6.

[4] A. Guillen-Perez and M.-D. Cano, "Flying ad hoc networks: A new domain for network communications," *Sensors*, vol. 18, no. 10, p. 3571, 2018.

[5] T. K. Saini and S. C. Sharma, "Recent advancements, review analysis, and extensions of the aodv with the illustration of the applied concept," *Ad Hoc Networks*, vol. 103, p. 102148, 2020.

[6] R. R. Roy, *Handbook of mobile ad hoc networks for mobility models*. Springer, 2011, vol. 170.

[7] N. Raza, M. U. Aftab, M. Q. Akbar, O. Ashraf, and M. Irfan, "Mobile ad-hoc networks applications and its challenges," *Communications and Network*, vol. 8, no. 3, pp. 131–136, 2016.

[8] H. Riasudheen, K. Selvamani, S. Mukherjee, and I. Divyasree, "An efficient energy-aware routing scheme for cloud-assisted manets in 5g," *Ad Hoc Networks*, vol. 97, p. 102021, 2020.

[9] K. Thangaramya, K. Kulothungan, R. Logambigai, M. Selvi, S. Ganapathy, and A. Kannan, "Energy aware cluster and neuro-fuzzy based routing algorithm for wireless sensor networks in iot," *Computer Networks*, vol. 151, pp. 211–223, 2019.

[10] J. Li, X. Li, Y. Gao, Y. Gao, and R. Zhang, "Dynamic cloudlet-assisted energy-saving routing mechanism for mobile ad hoc networks," *IEEE Access*, vol. 5, pp. 20 908–20 920, 2017.

[11] X. He and F. Y. Li, "Metric-based cooperative routing in multihop ad hoc networks," *Journal of Computer Networks and Communications*, vol. 2012, 2012.

[12] R. K. Jaiswal, "Position-based routing protocol using kalman filter as a prediction module for vehicular ad hoc networks," *Computers & Electrical Engineering*, vol. 83, p. 106599, 2020.

[13] R. A. Rehman, M. Sher, and M. K. Afzal, "Efficient delay and energy based routing in cognitive radio ad hoc networks," in *2012 International Conference on Emerging Technologies*. IEEE, 2012, pp. 1–5.

[14] A. M. Zungeru, L.-M. Ang, and K. P. Seng, "Classical and swarm intelligence based routing protocols for wireless sensor networks: A survey and comparison," *Journal of Network and Computer Applications*, vol. 35, no. 5, pp. 1508–1536, 2012.

[15] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," *IEEE computational intelligence magazine*, vol. 1, no. 4, pp. 28–39, 2006.

[16] N. Sabor, S. Sasaki, M. Abo-Zahhad, and S. M. Ahmed, "A comprehensive survey on hierarchical-based routing protocols for mobile wireless sensor networks: Review, taxonomy, and future directions," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[17] I. A. Alameri and J. Komarkova, "Network routing issues in global geographic information system," in *SHS Web of Conferences*, vol. 92. EDP Sciences, 2021.

[18] I. A. S. Alameri and J. Komárková, "Comparative study and analysis of wireless mobile adhoc networks routing protocols," in *Recenzovanỳ sborník příspěvk mezinárodní vědecké konference Mezinárodní Masarykova konference pro doktorandy a mladé vědecké pracovníky 2019*. MAGNANIMITAS, 2019.

[19] F. Xiangning and S. Yulin, "Improvement on leach protocol of wireless sensor network," in *2007 international conference on sensor technologies and applications (SENSORCOMM 2007)*. IEEE, 2007, pp. 260–264.

[20] S. Limin, L. Jianzhong, C. Yu, and Z. Hongsong, "Wireless sensor networks," *Beijing: Tsinghua University Press*, vol. 5, pp. 7–8, 2005.

[21] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless sensor networks," in *Proceeding of the Hawaii international conference on system sciences, Hawaii*, 2000.

[22] D. Kang, H.-S. Kim, C. Joo, and S. Bahk, "Orgma: Reliable opportunistic routing with gradient forwarding for manets," *Computer Networks*, vol. 131, pp. 52–64, 2018.

[23] J. Dongyao, Z. Shengxiong, L. Meng, and Z. Huaihua, "Adaptive multipath routing based on an improved leapfrog algorithm," *Information Sciences*, vol. 367, pp. 615–629, 2016.

[24] S. Xuekang, G. Wanyi, X. Xingquan, X. Baocheng, and G. Zhigang, "Node discovery algorithm based multipath olsr routing protocol," in *2009 WASE International Conference on Information Engineering*, vol. 2. IEEE, 2009, pp. 139–142.

[25] E. M. Royer, "Ad hoc on-demand distance vector routing," in *Second IEEE Workshop on Mobile Computer Systems and Applications, Feb. 1999*, 1999, pp. 25–26.

[26] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Proceedings WMCSA'99. Second IEEE Workshop on Mobile Computing Systems and Applications*. IEEE, 1999, pp. 90–100.

[27] Y. Yuan, H. Chen, and M. Jia, "An optimized ad-hoc on-demand multipath distance vector (aomdv) routing protocol," in *2005 Asia-Pacific Conference on Communications*. IEEE, 2005, pp. 569–573.

[28] M. A. Jubair, M. H. Hassan, S. A. Mostafa, H. Mahdin, A. Mustapha, L. H. Audah, F. S. Shaqwi, and A. H. Abbas, "Competitive analysis of single and multi-path routing protocols in mobile ad-hoc network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 2, 2019.

[29] P. Sarao, "Ad hoc on-demand multipath distance vector based routing in ad-hoc networks," *Wireless Personal Communications*, vol. 114, pp. 2933–2953, 2020.

[30] Z. J. Haas and M. R. Pearlman, "The performance of query control schemes for the zone routing protocol," *IEEE/ACM Transactions on networking*, vol. 9, no. 4, pp. 427–438, 2001.

[31] K. Shobha and K. Rajanikanth, "Efficient flooding using relay routing in on-demand routing protocol for mobile adhoc networks," in *2009 IEEE 9th Malaysia International Conference on Communications (MICC)*. IEEE, 2009, pp. 316–321.

[32] V. Ramesh, P. Subbaiah, and M. K. S. Supriya, "Modified dsr (preemptive) to reduce link breakage and routing overhead for manet using proactive route maintenance (prm)," *Global Journal of Computer Science and Technology*, 2010.

[33] M. K. Marina and S. R. Das, "On-demand multipath distance vector routing in ad hoc networks," in *Proceedings Ninth International Conference on Network Protocols. ICNP 2001*. IEEE, 2001, pp. 14–23.

[34] A. Bhattacharya and K. Sinha, "An efficient protocol for load-balanced multipath routing in mobile ad hoc networks," *Ad Hoc Networks*, vol. 63, pp. 104–114, 2017.

[35] M. Soni and B. K. Joshi, "Security assessment of routing protocols in mobile adhoc networks," in *2016 International Conference on ICT in Business Industry & Government (ICTBIG)*. IEEE, 2016, pp. 1–5.

[36] A. Abdollahi and M. Fathi, "An intrusion detection system on ping of death attacks in iot networks," *Wireless Personal Communications*, pp. 1–14, 2020.

[37] B. K. Tripathy, S. K. Jena, P. Bera, and S. Das, "An adaptive secure and efficient routing protocol for mobile ad hoc networks," *Wireless Personal Communications*, vol. 114, pp. 1339–1370, 2020.

[38] K. Manjappa and R. M. R. Guddeti, "Mobility aware-termite: a novel bio inspired routing protocol for mobile ad-hoc networks," *IET networks*, vol. 2, no. 4, pp. 188–195, 2013.

[39] S. Rosati, K. Krużelecki, L. Traynard, and B. R. Mobile, "Speed-aware routing for uav ad-hoc networks," in *2013 IEEE globecom workshops (GC Wkshps)*. IEEE, 2013, pp. 1367–1373.

[40] H. Wang, Z. Wang, G. Shen, F. Li, S. Han, and F. Zhao, "Wheelloc: Enabling continuous location service on mobile phone for outdoor scenarios," in *2013 Proceedings IEEE INFOCOM*. IEEE, 2013, pp. 2733–2741.

[41] I. U. Khan, I. M. Qureshi, M. A. Aziz, T. A. Cheema, and S. B. H. Shah, "Smart iot control-based nature inspired energy efficient routing protocol for flying ad hoc network (fanet)," *IEEE Access*, vol. 8, pp. 56 371–56 378, 2020.

[42] M. Z. Oo and M. Othman, "Analysis of single-path and multi-path aodvs over manhattan grid mobility model for mobile ad hoc networks," in *2010 International Conference on Electronics and Information Engineering*, vol. 1. IEEE, 2010, pp. V1–214.

[43] P. R. Satav and P. M. Jawandhiya, "Review on single-path multi-path routing protocol in manet: A study," in *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*. IEEE, 2016, pp. 1–7.

[44] A. M. El-Semary and H. Diab, "Bp-aodv: Blackhole protected aodv routing protocol for manets based on chaotic map," *IEEE Access*, vol. 7, pp. 95 197–95 211, 2019.

[45] T. Issariyakul and E. Hossain, "Introduction to network simulator 2 (ns2)," in *Introduction to network simulator NS2*. Springer, 2009, pp. 1–18.

[46] P. Li, Y. Fang, and J. Li, "Throughput, delay, and mobility in wireless ad hoc networks," in *2010 Proceedings IEEE INFOCOM*. IEEE, 2010, pp. 1–9.

[47] U. C. Kozat and L. Tassiulas, "Throughput capacity of random ad hoc networks with infrastructure support," in *Proceedings of the 9th annual international conference on Mobile computing and networking*, 2003, pp. 55–65.

[48] X. Lin, G. Sharma, R. R. Mazumdar, and N. B. Shroff, "Degenerate delay-capacity tradeoffs in ad-hoc networks with brownian mobility," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2777–2784, 2006.

[49] R. M. de Moraes, H. R. Sadjadpour, and J. J. Garcia-Luna-Aceves, "Mobility–capacity–delay trade-off in wireless ad hoc networks," *Ad Hoc Networks*, vol. 4, no. 5, pp. 607–620, 2006.

[50] S. Taneja and A. Kush, "Evaluation of normalized routing load for manet," in *International Conference on High Performance Architecture and Grid Computing*. Springer, 2011, pp. 442–448.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

[51] L. Qin and T. Kunz, "Increasing packet delivery ratio in dsr by link prediction," in *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 2003, pp. 10–pp.
[52] S. Malini, E. Kannan, and A. Valarmathi, "Performance optimization of single-path and multi-path aodv using response surface method (rsm)," in *2012 Proceedings of IEEE Southeastcon*. IEEE, 2012, pp. 1–4.

Evaluation of impact of mobility, network size and time on performance of adaptive routing protocols

Ibrahim Alameri
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
Jabir ibn Hayyan Medical University
st61833@upce.cz

Štěpán Hubálovský
Department of Informatics
University of Hradec Králové
Hradec Králové, Czech Republic
stepan.hubalovsky@uhk.cz

Jitka Komarkova
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
jitka.komarkova@upce.cz

***Abstract*—A Mobile ad-hoc Network (MANET) protocol must be configured correctly to ensure efficient data transfer. To achieve this aim a suitable routing protocol must be selected. Therefore, selecting the correct routing protocol is a critical condition, and it presents a classic problem in MANET. Also, using the proper values of the parameter in routing protocols plays a crucial role in MANET. MANET comprises several node devices run by battery as a power source. The primary function of MANET nodes is transmitting data based on routing protocols; thus, routing protocols play an essential role in MANET. Simultaneously, all routing protocols serve the same function in the network, but they differ in their performance. The current paper investigates four routing protocols performance by using the network simulator (NS-2) with various nodes speed, time simulations, network load, and network size. The current project evaluated the protocol performance based on metrics parameters such as throughput, end-to-end delay and packet delivery ratio. The simulation results showed that the ad-hoc On-demand Distance Vector (AODV) protocol was the best in all previous metrics parameters. In contrast, Zone Routing Protocol (ZRP) has the lowest performance. More details of the parameters have been presented in the current paper.**

***Keywords*—MANET, Routing protocols, Random Waypoint Model, AODV, routing metrics**

I. Introduction

A mobile ad-hoc network (MANET) involves a mobile collection of nodes that communicate directly with each other without a fixed infrastructure [1] [2]. Nodes (or hosts) of MANETs self-organize, keeps moving in any path and at any velocity [3]. Recently developed MANET showed a particular interest in having characteristic features of fast adaptation, reconfiguration, economic viability, and adaptation to flash flood scenarios such as emergency deployment for military service and population health monitoring [4] [5] [6]. The nature of MANET nodes is a dynamic and, wireless links frequently changes, to communication completion. Consequently, the main cons of MANET's is to keep all nodes fully updated with necessary information for routing [7]. The location of MANET nodes and their capability of transmission power (tp) present a significant role in determining network topology. MANET's nodes are run by batteries and have short communication ranges, therefore all unnecessary communications should be avoided to enhance the network lifespan and throughput. The development, management, organization, and overall administration of a MANET are all carried out by the network [8]. To address dynamic topology variations and achieve a quality reliable connection, the choosing of routing protocols is essential in MANET configuration. Therefore, an effective routing protocol is essential to improve MANET connectivity. Selecting the proper routing protocol acts as an essential key in network efficiency, where the current paper primarily aims to investigate and analyze the effect of various protocols (proactive, reactive, and hybrid) on network performance by testing multiple parameters. Very few papers studied scalability and mobility of routing protocols and network parameters. The parameters investigated in the current study includes various number of nodes and various speeds evaluated in the experimental results using network simulator 2 (NS2). Eventually, the optimal protocol is operating more nodes number and changeable speed. We chose the Ad hoc On-Demand Distance Vector (AODV) and Dynamic Source Routing (DSR) as reactive protocols were through research, and results for several papers proved the efficiency of these protocols. Destination Sequenced Distance Vector (DSDV) has been chosen as a proactive protocol, which is a predecessor of the AODV protocol. The rest structure of the current article is arranged as follows. An overview of the MANET routing protocols is introduced in the second section. While the third section presents the methods of the current project. The simulation Technique outline is explained in the fourth section. Results and discussion presented in the fifth section. Eventually, in the sixth section, the conclusions are presented.

II. MANET Protocols

As mentioned in the literature review, the Wireless ad-hoc networks have various designs and categorizing protocols. There are several different MANET routing protocols classified summarized in the Fig. 1 [9]. The various routing protocols are mainly based on the distance-vector or link-state, sometimes mix between the distance-vector and link-state but may use different methods and mechanisms in an adaptation context of particular purposes.

MANET routing protocols is classified as proactive, reactive, and hybrid. Proactive protocols actively refresh their routing

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

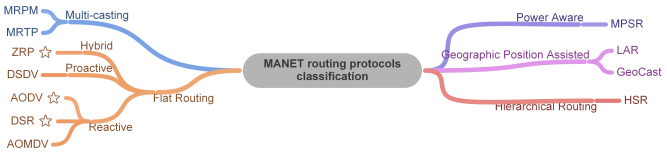


Figure 1. MANET routing protocols classification.

tables on a regular schedule [10]. Reactive routing protocols adaptively maintain a route and keep it while the path has long been used. Finally, the most common hybrid protocols combine the strengths of both approaches (proactive and reactive) to solving a problem.

1) Proactive Routing Protocols: Proactive protocols use a strategy similar to that of the protocol used in wired protocols. One goal for Proactive routing protocols is to provide the most updated networked path information and discover new paths [9].This enables them to transfer packets optimally because the route is calculated when the packet is forwarded to the node. An example of this type of routing protocol is Destination Sequenced Distance Vector (DSDV).

DSDV: is a MANET routing method that utilizes tables. Following the Bellman-Ford algorithm, several other changes were made and applied to the DSDV algorithm. The routing table has three components: hops number, access nodes, and sequence number commissioned to the destination node. Sequence numbers differentiate routes that have already been declared stable from those in the process of being established and avoid loops [11]. Routing tables are regularly broadcast to all the interconnected neighboring nodes to keep them up to date, or in case there is an essential update done in the table [12]. It is preferable to send updates as a small batches rather than constantly to maintain network stability. The routing table entry also contains a number generated by the transmitter and called a sequence number. The path selects the highest sequence number. In situations where two or more paths have the same sequence number, the one with the better metric (i.e., the path with the shortest length) is chosen [13] [14].

2) Reactive Routing Protocols: The reactive routing protocols are also known as on-demand protocols. If the on-demand protocols are used, the routes are checked only when required [15], this meaning that the paths are only checked when any nodes need to connect—Discovery process of the path, when a path is discovered, or when no path is found at all. In this context, these characteristics make it a reactive protocol. Several MANET routing protocols implement the reactive technique include dynamic source routing (DSR) and ad-hoc on-demand distance vector (AODV) [16].

AODV:AODV uses three routing messages for three types of requests: Route Request (RREQ), Route Reply (RREP), and Route Error (RERR). When AODV tries to obtain a path to the destination, it flooding the network with request routing information. If the request is sent to the intended destination successfully, the destination will reply by gives an RREP to the source of this RREQ. However, if it was not, the last node would respond by provides a RERR with to the source of RREQ. Like the DSDV routing protocol, AODV uses the sequence numbers in the interchange information route process []. Every RREQ will be addressed only once, thus minimizing routing overhead. It only tracks the next hop among other features in the route table information [17] [18]. In AODV routing process uses the intermediate nodes. This process is also known as a hop by hop. The intermediate nodes use to transmitting packets between source and destination.

DSR: Path exploration in DSR uses RREQ/RREP packets, the same method as in AODV. In contrast to AODV, the paths are kept in a path cache. Additionally, the DSR is a path-based protocol that keeps data about the entire paths between the source and destination. Instead of forwarding the packet hop by hop like AODV, the packet conveys the whole route from source to destination in DSR.

A. Hybrid Routing Protocols

Hybrid routing combines close reactive routing protocols and proactive routing protocols to mitigate overhead routing and delays in the network resulting from discovery routing operations [19]. Higher reliability and scalability provide the contributions of hybrid routing protocols. The drawback of hybrid protocols is that new routes are being found connectivity issues presented within a network's latency. Zone Routing Protocol (ZRP) is one of the significant protocol type [20] [8].

ZRP: Utilizes reactive and proactive protocols in a hybrid system by using the proactive exploration of nearby nodes and using reactive communication routing protocol features between nodes [21]. A single config factor defines how the ZRP is designed for a given network. ZRP is a combination of two sub-routing protocols called Inter-zone Routing Protocol (IERP) and Intra zone Routing Protocol (IARP) [22].

Source table can be used to recognize a path to the destination zone's entry through a constructive cache table lookup. IARP allows the path to be found by looking in the source zone using the cash routing table when it has already been sent a certain amount of response time. When the source and destination are both in the same area, IARP determines the path and instantly sends the packets. According to these advantage features, IARP is being used in the algorithm of ZRP routing protocols [23].

III. Materials and Methods

A. Mobility Patterns

Random mobility does not impose limits on the nodes' movements in the MANET. In other words, in a random manner, the speed, destination, and orientation are chosen for each node, where each one of these factors is determined on its own and separately for the nodes [24]. MANET has several mobility types: Random waypoint model (RWM), Random walk model, Random direction model, Street random waypoint, Reference point group model (RPGM), Manhattan

mobility model and Freeway mobility model. The random Waypoint model (RWM) is a widely employed mobility model. The RWM has two different models the Random Waypoint model and the Random Walk model. Due to the efficiency, plainness, unsophistication, and availability, the RWM has became the MANET standard mobility model. The setdest tool is used to create the node trace of the RWM. The commonly used network simulator NS-2 includes this function. It is advantageous to use mobility models as these imitate the way mobile nodes react to network efficiency. There's a significant correlation between mobility type and network performance [25].

B. Work proposed

Although several articles provided a foundation for further analysis based on essential suppositions, they did not address the fundamental analysis because critical conclusions were not taken to study the results with the different types of routing protocols such as proactive, reactive, and hybrid protocol. The current study investigates the various MANET routing protocols comprised of five aspect speed of nodes, number of nodes (Network overload), mobility, simulation time and network area. Besides, the current study takes into count applied all these aspects with different routing protocol types. MANET routing protocols' efficiency is measured by four parameters: packet drop rate (PDR), end-to-end delay (E2E Delay), throughput, and normalized routing load (NRL). The current study conducted with three different scenarios summarizes in the following section. The presented study examined the MANET routing protocols thirty-five times for each phase to substantiate and prove routing protocol efficiency. The current project used a variant number of nodes between 40 till 100 nodes. The number of nodes between (40 - 80) refers to the small network, while the number between (81 -100) refers to the large network. Fig. 2 refers to network typology. In

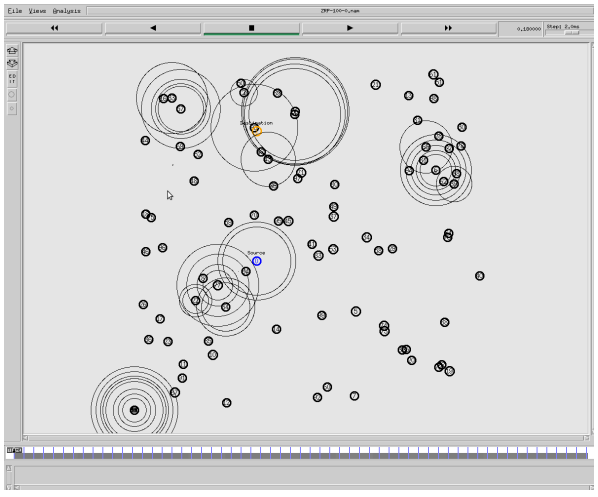


Figure 2. Network typology.

Fig. 2 nodes are free moving in the network area. Although the wireless communication range will vary across nodes, the data is propagated from source node to destination node. Accurate packet delivery is difficult to be ensured as there are a vast number of links between different parts of the network. This is substantiated in the upcoming section of the current work.

C. performance Parameters

The different routing protocols' parameters analyzed via simulation in the current study. These parameters are:

- **Packet Drop Rate (PDR)**: PDR is described as the" quantity of dropped packets per second". The dropped packets data have extracted and calculated from the simulation trace file. Every dropped packet will increase the unit time counter. Then the extracted data fed into the MATLAB, to plot the line graph for the entire simulation time. The PDR can be calculated by equation (1).

$$PDR = \frac{\sum \text{Number of packets received}}{\sum \text{Number of packets send by node}} \quad (1)$$

- **Throughput**: A network throughput represents the total number of packets that have been delivered successfully per period unit of time. The optimal protocol is the protocol that generates a higher throughput rate. In other words, throughput is essential in evaluating the effectiveness and scalability of routing protocols, it is calculated by equation (2).

$$\text{Throughput} = \frac{\Sigma \text{ received packets size}}{\text{time}} \quad (2)$$

- **End to End Delay (E2E)**: is the average time that needs to send packets to the final destination. Also, it is defined as the difference between the transit time and the arrival time of the packet and calculated by equation (3).

$$D_{avg} = Tr_{avg} - Ts_{avg} \quad (3)$$

- **Normalized routing load (NRL)**: is the total routing packets of the total data packet that are received at the destination which is called routing load (4).

$$RO = \frac{\sum \text{ Routing packets}}{\sum \text{ Packets received}} \quad (4)$$

IV. Simulation Technique

The current work investigates the behavior of reactive, proactive, and hybrid protocols. These routing protocols are AODV, DSR, DSDV, and ZRP routing protocols are examined.

A. Simulation and Metrics Performance

There are several network simulations tools, such as (NS-2) [26], (NS-3) network simulator [27], and the network simulation software QualNet (QualNet) [28]. NS-2 has been chosen as the protocol simulator for this current study because of its abundance and support of several network protocols. Four different routing protocols have been selected belonging to other families: AODV, DSR as a reactive routing protocol, DSDV as a proactive routing protocol, and ZRP as a hybrid protocol. By default, NS2 doesn't have ZRP routing protocol, unlike the other routing protocols like AODV, DSR,

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

and DSDV, where those routing protocols are automatically installed with NS2. So, the patches of ZRP routing protocols have been implemented to NS 2.35. This patch is a solution to the ZRP installation. The ZRP routing protocol was out of the NS2 scope of development, so the ZRP protocol had to be added to the NS2 to implement it. Practical simulations were carried out after the patch was added.

In addition, the current work proposed a custom Perl script to calculate metrics such as packet drop rate (PDR), throughput, average end-to-end delay, and network overhead from the trace files. Finally, after these suggested modifications, the protocols and the four MANET routing scenarios are installed and ready to be tested.

The mobility model refers to the movement pattern of the mobile nodes during the simulation study. It plays a significant role in designing and implementing an excellent wireless infrastructure because a routing protocol has performed well in one mobility model, even though it is unnecessary to perform well in other conditions.

Besides, the scripts presented in OTcl, an object-oriented language enhanced version of Tcl modeling and analyzing UDP protocols, routers, and other network items, are used to execute the NS-2 software. Tcl scripts were used to create network scenario simulation, connection settings, nodes movement, and position are implemented in the same fashion. Other modifications were implemented to adjust the transmitting and receiving power at nodes to produce an effective influence per each packet. The simulation study results are produced in a trace file that is included in the stimulation details of the network.

The MATLAB programming language generates graphs. The current study used Random WayPoint (RWP) mobility model in the network simulation parameter.

As mentioned in the section above of ”work proposed,” where the presented work for three different scenarios included four aspect speed of nodes, network overload, mobility, and network area. More comprehensive details on the simulation parameters and simulation outcomes will be given below. The illustrations are examined and empirically deduced to support how to deal with various protocols for various network conditions and which routing protocol will be adapted, convenient, and appropriate for the MANET network.

B. Result and discussion

The performance analysis results by varying network overload, node speed, and area of the network will discuss in this subsection. The parameters are used in this simulation in the current study represented in table II. For organizational purposes, we analyzed and studied the first and second scenarios together. Table II, indicates the parameter simulation used during the current study in the first and second scenarios. In this simulation, the number of nodes was varied between (40 ,80 ,100), the node speed was 20, 40, and 60, and the network area was 1000 m2 and 1500 m2—the study in the first and second scenarios, conducted for all the different routing protocols used in this work.

TABLE I: Network simulation scenario.

Parameters	***Value***
Simulator	Network Simulator 2
Simulation Time	180 seconds
Area of Different Scenarios	1000m x 1000m, 1500m x 1500m
Nodes Speed for different Scenarios	20, 40, 60
Transport Layer Protocol	UDP
Number of Nodes	40, 60, 80, 90, 100
Movement Scenario	Random Way Point (RWP)
Routing Protocols	AODV, DSR, DSDV, ZRP

Routing on the WMN is difficult to achieve because of the node’s mobility. The mobility leads to irregular alteration in the network’s topology, so selecting a suitable routing protocol is a challenging task. The present work evaluated the effects of mobility on various performance metrics in the current experiment work. The maximum mobile speeds obtained from a limited network, were between 1 m/s to 60 m/s and, the experimental networks were performed on small networks (40, 80 nodes connectivity sets) and large networks (81-100 nodes, connectivity sets).

There have been two different types of experiments and their results are summarized in the following subsections are set to the first and second scenarios.

C. Scenarios 1 & 2

Test 1- the impact of the network size and node speed

Fig. 3 & 4 illustrates the results of this experiment. The significant metric that stands out in this study is the throughput of AODV routing protocol. Where Fig. 3 & 4 found that the AODV dominates routing protocols at any condition simulation parameters. The AODV is fulfilled well than other routing protocols. It can be concluded that AODV has an Important throughput when the node’s speed is high or low.

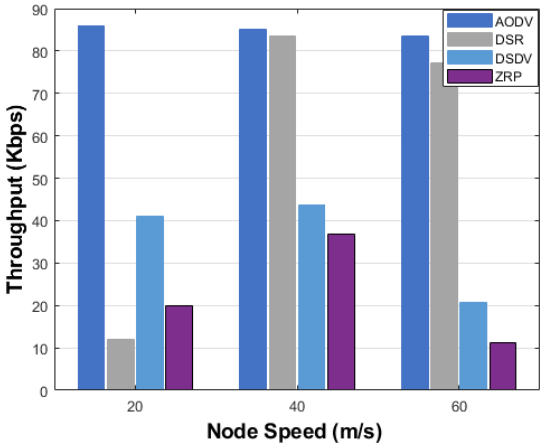


Figure 3. Throughput based variety - nodes speed - 1000 m^2.

Moreover, the throughput of AODV remains higher than other routing protocols even the network was large in 1500

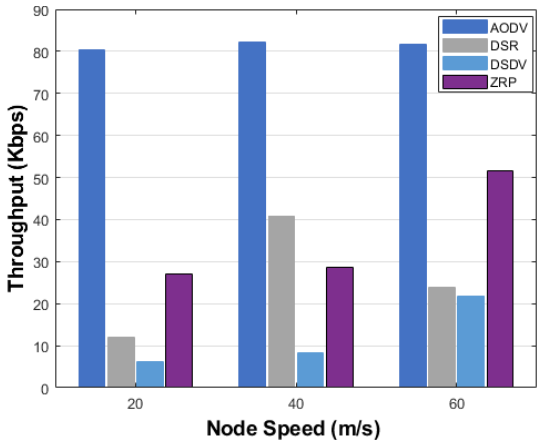


Figure 4. Throughput based variety - nodes speed - 1500 m^2.

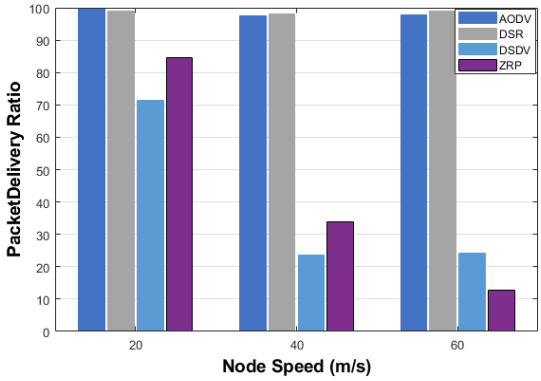


Figure 5. PDR based variety nodes - max speed - 1000 m^2.

m^2. In addition, the DSR routing protocol produces higher throughput in 1000 m^2, which is inversely proportional with the network area. ZRP has a noticeable improvement in terms of throughput with higher mobility speed and a large network area.

The results of the packet delivery ratio (PDR) analysis are summarised in Fig. 5 & 6. In this analysis, ZRP and DSDV have the lowest PDR. Contrary to ZRP and DSDV, this study found significant results related to the AODV routing protocol, where AODV AODV is the leader protocol when the PDR is considered.

Fig. 7 & 8 presents the term of end-to-end delay (E2E) metrics reinforce the assumption that the metrics will remain constant as the max speed of mobile nodes increases. DSDV performs better than all other routing protocols in this situation, providing the minimum packet delay. DSDV produces acceptable values whether the size of the network was small or large within various mobility speeds. The AODV performed an optimal ratios packet delay, whereas, the ZRP present maximum packet delay.

The Fig. 9 & 10 Shows that the ZRP routing overhead is more significant than the other routing protocols, the performance getting significantly improved with the increment of the network area and speed of nodes. The overhead routing of the protocols following affects their internal implementation. In a high nodes speed environment, when the nodes numbers are 40, 80, and 100, the results showed that the AODV routing protocol has the lowest routing overhead. However, it is preferable to reduce the routing overhead with a large network with high mobility conditions and scalable nodes. Also, AODV routing protocols present a better effect on the NRL and lowest routing overhead rather than the other routing protocols.summary, the network's speed and size area have a significant and direct impact on network efficiency. The parameters of the network performance acting more stable and improved significantly with the increment nodes speed, but the performance decrease with expanded network area.

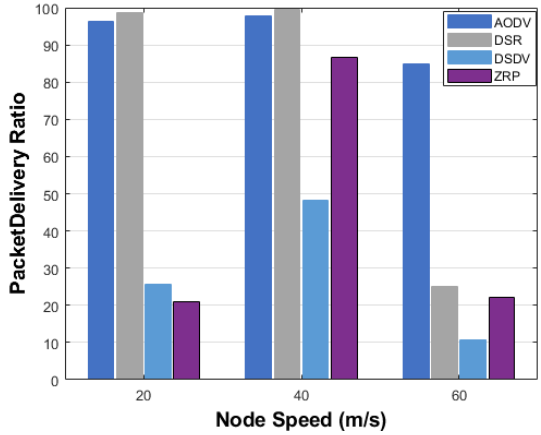


Figure 6. PDR based variety nodes - max speed - 1500 m^2.

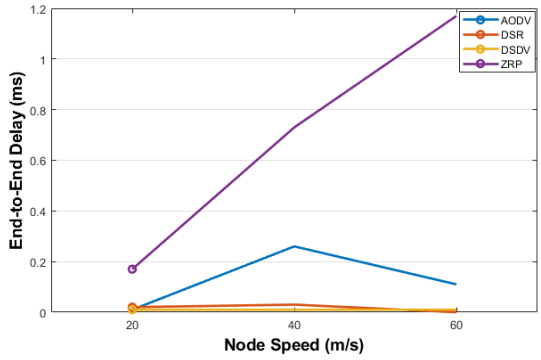


Figure 7. E2E by different - max speed - 1000 m^2.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

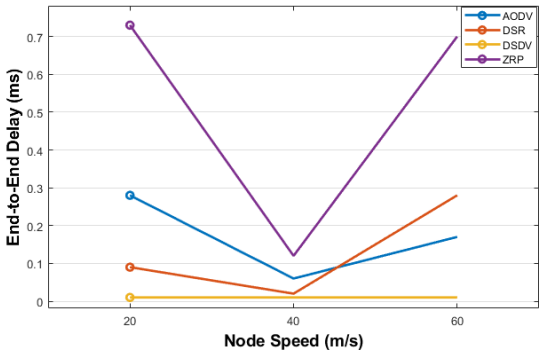


Figure 8. E2E by different - max speed - 1500 m^2.

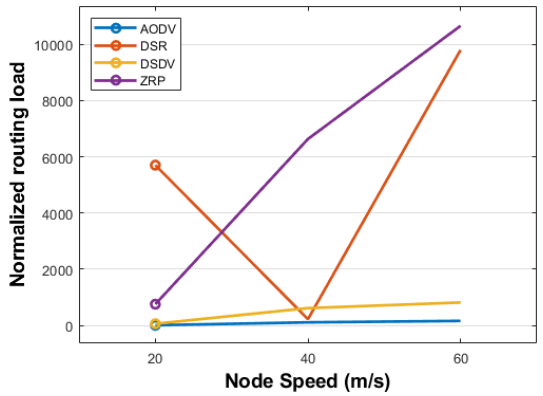


Figure 9. NRL by different nodes - max speed - 1000 m^2.

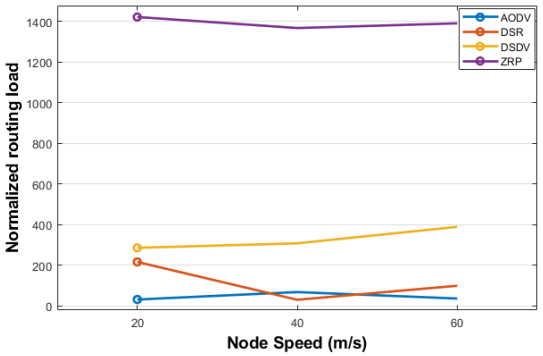


Figure 10. NRL by different nodes - max speed - 1500 m^2.

Test 2 - the impact of the nodes number and network area

The current scenario employs various network typologies. In this study, with different network areas, and maximum nodes number is 100 nodes. The plotted graphs below present the results of AODV and DSR protocols.

Fig. 11 & 12 illustrates the throughput four routing protocols for 40, 80, and 100 nodes, respectively. The DSR and DSDV protocols are worse significantly when the number of nodes within the large network area increases. However, the DSR performance remains acceptable in the small and middle network area. The AODV performance is much better than other routing protocols. Noting a slight improvement in the performance of ZRP routing protocol within a large network area. The packet delivery ratio shows in 13 & 14, is very high though AODV is doing much better in this term. In contrast, the DSR, DSDV, and ZRP serve the lowest than AODV, especially with the large network size and higher nodes numbers. Simultaneously, the results of Fig. 15 & 16 show that the AODV protocol has an insignificant increment in the average end-to-end delay with the network size. In comparison, the ZRP protocols reduced the network performance with an increase in network size and an increasing number of nodes. Terms of normalized routing load (NRL) represented in 17 & 18 the results reveal that AODV performs better than the other three routing protocols. In contrast, the DSR protocol has poor outcomes in terms of routing overhead. Moreover, DSR and ZRP have the most considerable routing load, increasing the network size and nodes.

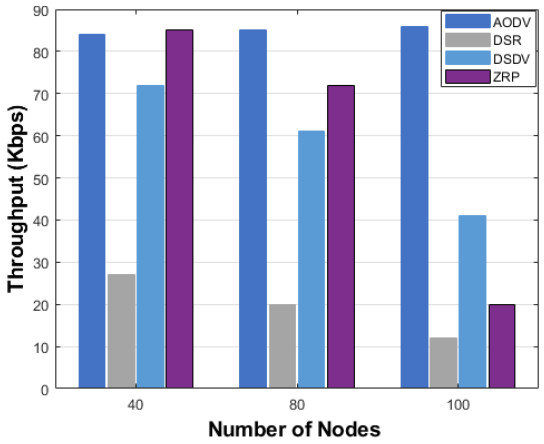


Figure 11. Throughput vs number of nodes - 1000 m^2.

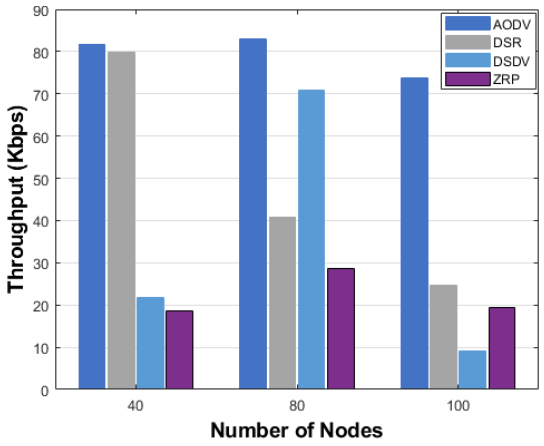


Figure 12. Throughput vs number of nodes - 1500 m^2.

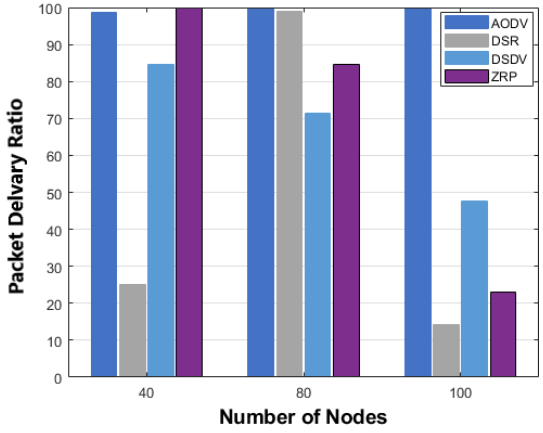


Figure 13. PDR vs number of nodes - 1000 m^2.

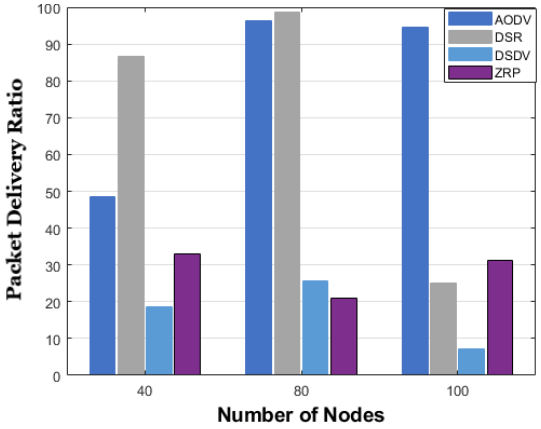


Figure 14. PDR vs number of nodes - 1500 m^2.

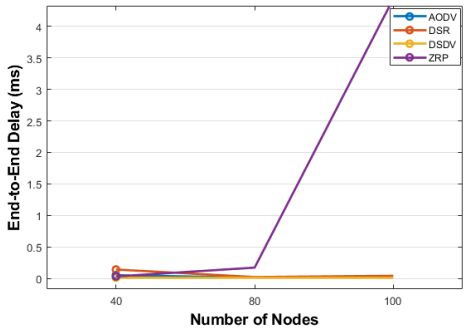


Figure 15. E2E vs number of nodes - 1000 m^2.

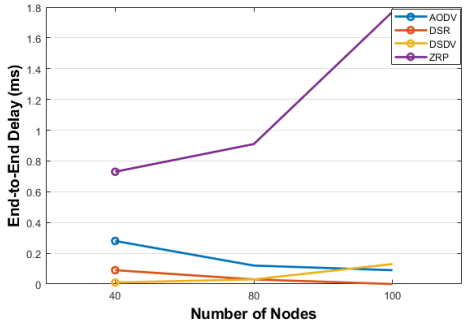


Figure 16. E2E vs number of nodes - 1500 m^2.

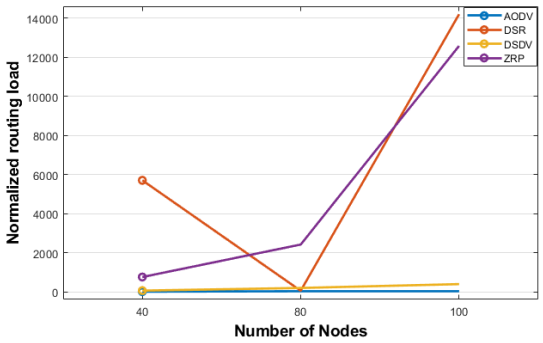


Figure 17. NRL vs number of nodes - 1000 m^2.

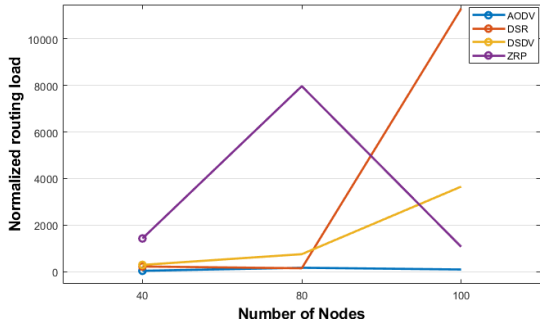


Figure 18. NRL vs number of nodes - 1500 m^2.

TABLE II: Network simulation scenario

Parameters	***Value***
Simulator	Network Simulator 2
Simulation Time	180sec, 300 sec
Area of Different Scenarios	1000^m
Nodes Speed for different Scenarios	20
Transport Layer Protocol	UDP
Number of Nodes	100
Movement Scenario	RWP
Routing Protocols	AODV, DSR, DSDV, ZRP

D. Scenario 3

The pivot of the current test in this scenario is the simulation time. In this test, the term of simulation time was the factor of axis study alongside with the node speed of the network. Table II presents the network parameters. Evaluation of the influence of simulation time on the MANET protocols. Four routing protocols have been evaluated and compared to assess their efficiency. They are PDR, end-to-to-end delay, normalized routing overhead, and throughput as shown in Fig. 19 & 20 & 21 & 22 sequentially. These protocols are seated to evaluate the performance of routing protocols based on the time alteration. The outcome of the simulation result shows that the AODV is performed better in simulation time than another routing protocol when simulation increases. Expect-

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

edly the PDR has the same value during the simulation, even as long or short time. Furthermore, the result showed that the AODV gives a valuable result with the throughput, routing overhead, and end-to-end delay. In contrast, the ZRP has the worst performance when the simulation time increases.

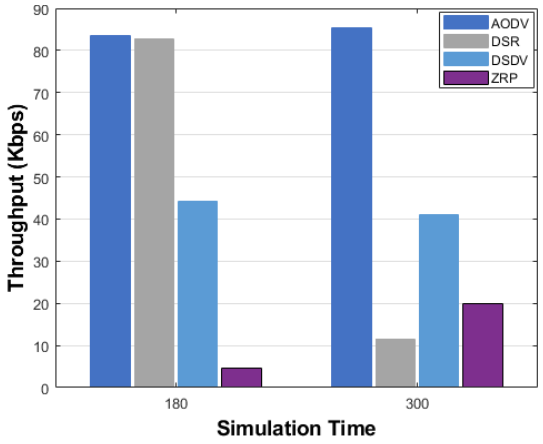


Figure 19. Throughput based on a variety of simulation time.

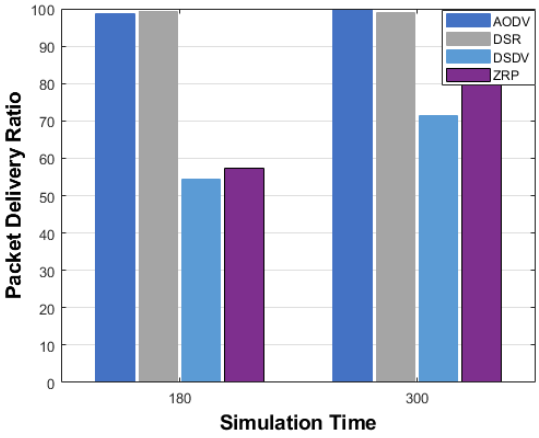


Figure 20. PDR based on a variety of simulation time.

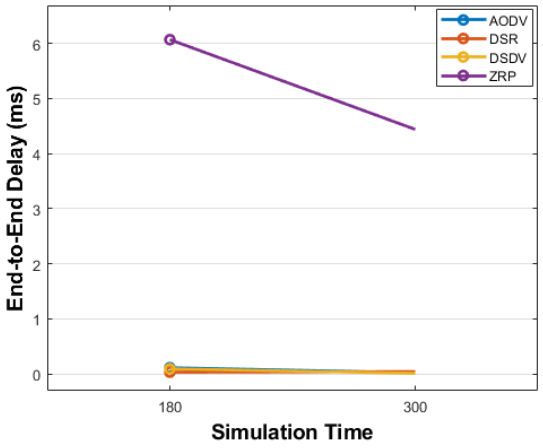


Figure 21. E2E based on a variety of simulation time.

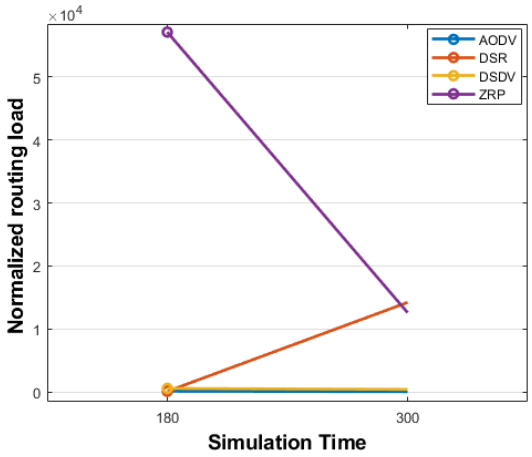


Figure 22. NRL based on a variety of simulation time.

V. Conclusion

The current investigation aimed to assess the performance of routing protocols under different metrics aspects such as throughput, PDR, end-to-end delay, and routing overhead. The current work was undertaken with NS2 simulation and the evaluation of graphics created by MATLAB with alteration the simulation time, number of nodes and network size. The simulation results obtained detailed of different metrics by employing the following routing protocols: AODV, DSR, DSDV, and ZRP. One of the more significant findings emerged from this study is that AODV routing protocols was the best performance in respect of the metrics mentioned above. On the other hand, the evaluation of the PDR was a little more valuable than the different routing protocols. A small to medium-sized network doesn't affect the performance of the routing protocols significantly, even the protocol outcomes were well performance in this type of the network. The results indicated that the vital feature of a successful routing protocol is the ability to scale. In generality, the normalized routing load is responses proportionally with the number of nodes and time simulation besides increasing network area. The normalizing load does not often change when there is increasing in the network size, the number of nodes, and time simulation, especially with AODV protocol, but the performance of the DSR and ZRP protocols was the worst, the NRL as a result of overload and network area.

The current simulation results could assist the researchers in deciding which is the best WMN routing protocol. Where the current results provide some helpful guidance outlines, that may help to select or develop a routing protocol for WMNs. Further investigations and experimentations into PDR are strongly recommended. Also, study similar experimental design to the current project should be carried out on TCP traffic instead of UDP traffic.

Acknowledgment

This paper was supported by SGS University of Pardubice project No. SGS_2021_011.

REFERENCES

[1] H. Zemrane, Y. Baddi, and A. Hasbi, "Mobile adhoc networks for intelligent transportation system: comparative analysis of the routing protocols," *Procedia Computer Science*, vol. 160, pp. 758–765, 2019.

[2] M. Sindhwani, R. Singh, A. Sachdeva, and C. Singh, "Improvisation of optimization technique and aodv routing protocol in vanet," *Materials Today: Proceedings*, 2021.

[3] S. G. Pease, L. Guan, I. Phillips, and A. Grigg, "Cross-layer signalling and middleware: A survey for inelastic soft real-time applications in manets," *Journal of network and computer applications*, vol. 34, no. 6, pp. 1928–1941, 2011.

[4] D. E. M. Ahmed and O. O. Khalifa, "An overview of manets: applications, characteristics, challenges and recent issues," 2017.

[5] U. Essays, "The characteristics and applications of manets computer science essay," *URL: http://www. ukessays. com/essays/computerscience/the-characteristics-and-applications-of-manets-computerscience-essay. php*.

[6] S. Dhar, "Manet: Applications, issues, and challenges for the future," *International Journal of Business Data Communications and Networking (IJBDCN)*, vol. 1, no. 2, pp. 66–92, 2005.

[7] U. S. e. Daya K. Lobiyal, Vibhakar Mansotra, *Next-Generation Networks: Proceedings of CSI-2015*, 1st ed., ser. Advances in Intelligent Systems and Computing 638. Springer Singapore, 2018.

[8] A. K. Sharma and M. C. Trivedi, "Performance comparison of aodv, zrp and aodvdr routing protocols in manet," in *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*. IEEE, 2016, pp. 231–236.

[9] I. A. Alameri and J. Komarkova, "A multi-parameter comparative study of manet routing protocols," in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2020, pp. 1–6.

[10] R. Skaggs-Schellenberg, N. Wang, and D. Wright, "Performance evaluation and analysis of proactive and reactive manet protocols at varied speeds," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2020, pp. 0981–0985.

[11] A. Daas, K. Mofleh, E. Jabr, and S. Hamad, "Comparison between aodv and dsdv routing protocols in mobile ad-hoc network (manet)," in *2015 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW)*. IEEE, 2015, pp. 1–5.

[12] Y. S. Devi and M. Roopa, "Performance analysis of routing protocols in vehicular adhoc networks," *Materials Today: Proceedings*, 2021.

[13] N. Sarmah, "Performance analysis of mobile ad-hoc routing protocols by varying mobility, speed and network load," 2014.

[14] M. Kumar, C. Sharma, A. Dhiman, and A. K. Rangra, "Performance variation of routing protocols with mobility and scalability in manet," in *Next-Generation Networks*. Springer, 2018, pp. 9–21.

[15] K. Mariyappan, M. S. Christo, and R. Khilar, "Implementation of fanet energy efficient aodv routing protocols for flying ad hoc networks [feeaodv]," *Materials Today: Proceedings*, 2021.

[16] M. Singh and J. Sharma, "Performance analysis of secure & efficient aodv (se-aodv) with aodv routing protocol using ns2," in *Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization*. IEEE, 2014, pp. 1–6.

[17] A. Zakrzewska, L. Koszalka, and I. Pozniak-Koszalka, "Performance study of routing protocols for wireless mesh networks," in *2008 19th International Conference on Systems Engineering*. IEEE, 2008, pp. 331–336.

[18] I. A. Alameri and J. Komarkova, "Network routing issues in global geographic information system," in *SHS Web of Conferences*, vol. 92. EDP Sciences, 2021.

[19] Y. MingChuan, G. Qing *et al.*, "End-to-end delay assessment and hybrid routing protocol for vehicular ad hoc network," *IERI Procedia*, vol. 2, pp. 727–733, 2012.

[20] T. E. Ali, L. A. K. al Dulaimi, and Y. E. Majeed, "Review and performance comparison of vanet protocols: Aodv, dsr, olsr, dymo, dsdv & zrp," in *2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA)*. IEEE, 2016, pp. 1–6.

[21] S. Mittal and P. Kaur, "Performance comparison of aodv, dsr and zrp routing protocols in manet's," in *2009 international conference on advances in computing, control, and telecommunication technologies*. IEEE, 2009, pp. 165–168.

[22] N. S. Benni and S. S. Manvi, "Impact of node failure on the routing performance in wireless mesh network," in *2016 IEEE Region 10 Conference (TENCON)*. IEEE, 2016, pp. 2120–2125.

[23] K. Sampoornam and G. R. Darshini, "Performance analysis of bellman ford, aodv, dsr, zrp and dymo routing protocol in manet using exata," in *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2019, pp. 1–5.

[24] F. Bai and A. Helmy, "A survey of mobility models," *Wireless Adhoc Networks. University of Southern California, USA*, vol. 206, p. 147, 2004.

[25] S. Samaoui, I. El Bouabidi, M. S. Obaidat, F. Zarai, and W. Mansouri, "Wireless and mobile technologies and protocols and their performance evaluation," in *Modeling and Simulation of Computer Networks and Systems*. Elsevier, 2015, pp. 3–32.

[26] M. Anand, N. Balaji, N. Bharathiraja, and A. Antonidoss, "A controlled framework for reliable multicast routing protocol in mobile ad hoc network," *Materials Today: Proceedings*, 2021.

[27] M. R. Hasan, Y. Zhao, Y. Luo, G. Wang, and R. M. Winter, "An effective aodv-based flooding detection and prevention for smart meter network," *Procedia Computer Science*, vol. 129, pp. 454–460, 2018.

[28] G. Xiang, G. Peng, L. Yu, M. Qi, L. Cheying, and S. Dan, "An interactive interface of qualnet for digsilent in power grid simulation," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2019, pp. 330–333.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Reliability analysis of a retrial queueing systems with collisions, impatient customers, and catastrophic breakdowns

Attila Kuki, Tamás Bérczes, János Sztrik, Ádám Tóth
Faculty of Informatics University of Debrecen, Hungary,4028
Email:[kuki.attila, berczes.tamas, janos.sztrik, toth.adam]@inf.unideb.hu

***Abstract*—A lot of different real-life systems can be modeled by retrial queuing (RQ) models. In this paper, RQ-systems are considered. The single server system is non-reliable, non-deterministic system failures might occur. This is a finite source system. In applications, it is more realistic, and there is no stability problem. One of the first considered system operational characteristics is the collision or the conflict of customers. When a job is under service at the server, and a new job comes, they will collide. In this case, both jobs will transport to a virtual waiting room, called orbit. The customers retry their requests from the orbit. The retrial times are random. Server failures might happen, the server might go down. While the server is down state, the new requests are transported to the orbit, or the source is blocked, that is, no customer can enter into the system. The second system characteristic is the impatient property of the customers. The customers stay in the orbit and waiting for their service. After a non-deterministic time-interval, a customer gives up retrying and leaves the system. These customers will be lost from the system, they remain unserved. This is the impatient characteristic. The third system characteristic is the catastrophic breakdown. It means, that in case of a negative event, all of the customers at the server and in the orbit leave the system, and take their places in the source. The novelty of this paper is to investigate the phenomenon of the catastrophic breakdown in a collision environment with impatient customers.**

This impatient property results, that the recursive algorithm for the time-independent probabilities can not be formulated. MOSEL-2 tool can be used for solving the system equations and calculating the system performance measures. These measures are, for example, the average sojourn time and other reliability metrics. The main goal is to investigate the effect of the impatient property under catastrophic breakdown. Numerical results are presented graphically, as well.

I. Introduction

Modeling infocommunication systems is essential to understand their dynamical behavior and find an optimal working environment. There exist a lot of tools and methods for modeling these types of systems. One of the most popular and effective tools are the retrial queueing systems (RQ-systems). Compared with the simple queueing systems, in RQ-models the customers are not lost when the system is busy. An RQ-system can have an infinite and finite number of sources. The requests arrive from the outside world or the source facility, respectively. If a customer arrives to the server, and the server is still working with an other customer, the new job enters a virtual lobby, called orbit and waits for a non-deterministic time interval (exponential), and it retries its service demand again. The orbit can be imagined as a virtual waiting facility, and it is assumed to be large enough, so an incoming job always finds a free place in the orbit. From the orbit, the jobs do not give up, they try to reach the server. Once the server is free, the jobs step into the service facility, and they will be served. This is the patient behavior of jobs.

In this paper, the impatient behavior of the customers is also considered. A job waiting in the virtual lobby (orbit) might leave the system after constant or random waiting time. Our models deal with random times, described with a distribution, e.g. the exponential distribution. These customers will remain unserved, their requests are lost.

Examples of technologies and applications, which can be modeled by an RQ-systems can be telephone centers, sensor networks, repair facilities, telecommunication environment, etc. Infinite source models have been considered and studied by numerous authors, very large number of papers were published in the literature. But, in real-life applications, there are a lot of situations, where the infinite sources are not so suitable. When there are only finite numbers of entities, which are related with the system, the finite source models are more adequate. Many examples can be mentioned, for example, mobile cellular networks, intelligent sensor networks, a lot of new IoT systems, and so on. Results on finite source retrial queueing systems are, for example, in [1], [2], [3], [4], [5], [6].

In many applications, unfortunately, the systems are non-reliable. They subject to random breakdowns. This situation also has to be investigated. Random server failures and repairs are included in the models. Generally or exponentially distributed random times and time intervals can be considered in modeling the breakdown and repair processes. A non-reliable system is very sensitive, the calculated outcomes and descriptors of the system have to be handled very carefully. Non-reliable, finite source retrial systems have been investigated by many authors, e.g. in [7], [8], [9], [3], [10].

A non-reliable $M/M/1//N$ retrial queueing system with conflict of jobs is also a part of this investigation. The phenomenon of collisions of jobs (or conflict of jobs) is a very common behavior in non-synchronized infocommunication systems with a constrained resources, for example, Ethernet transmissions and other communication facilities. In case of

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

conflict, all of the involved signals are damaged, and they need to re-send. The working and the performance of the system is sub-optimal. It is very important, to build up procedures, which try to avoid or at least decrease the effects of conflicts. Results can be found in [11], [12], [13], [14], [15], [16].

The focus of this paper is the catastrophic breakdown. Retrial queueing models in which customers are removed from the system due to catastrophic or disaster events have been investigated extensively by authors. Modeling special systems, e.g. automatic teller machines needs different types of breakdowns. A catastrophic event can be, for example, mechanical failures or power outages. Disaster events are known also as a negative arrival or a negative customer. The presence or arrival of a negative job might be dangerous for the system. The other jobs, which work for the system, are called positive jobs. These positive jobs can be damaged by the negative ones. The positive jobs might even be removed from the system. The most extreme situation is, when all of the positive jobs are removed from the system. This event is called a disaster, or catastrophic breakdown. In addition, this type of event interrupts all of the service processes at the servers and breaks down the service unit. The service in the service unit will be interrupted. The service has been done until the interruption point is lost. All of the customers from the server and the orbit are sent back to the source. Detailed studies on catastrophic breakdowns and negative customers can be found in [17], [18] [19], [20], [21].

The impatient property of customers and the feature of catastrophic breakdown make the system equations so complex, that a simple numeric solution can not be performed. That's why a computer program is used for calculating the system probabilities. Using the resulting steady-state system distribution, the most important system measures characteristics can be formulated. Figures will be provided with the effect of different parameters on these performance metrics.

II. Description of the system

In this paper, an $M/M/1//N$ model is considered. This is a finite source system, and by Kendall's description, the inter-arrival times and the service times follow the exponential law, the number of servers is one, and the number of sources is N. Two scenarios of the system can be studied and compared:

- The common break-down mode. This is the well-known non-reliable environment, that is the service facility is exposed to non-deterministic failures. The inter-event times between the failures follow the exponential law. Two different parameters can be considered for describing the failure events. In the case of a non-working service unit, the parameter is γ_0. In the case of a working server, the failure parameter is γ_1. Later, in numerical investigations, these two parameters will not be distinguished. The behavior of the job interrupted at the server can be different. The service can be continued or started again. Here, the job interrupted at the service facility is transferred back to the orbit. There is no time transition before the repair. After the failure event, the process of the repair begins immediately. The repair time is also a random variable, the distribution is exponential, and the repair parameter is γ_2. While the repair is in progress, different behavior of the source can be handled. The system is blocked, that is, no new request can enter into the system. Or, the system is not blocked, generating of new requests continues. These new jobs can not reach the server under repair, so they are transferred into the orbit. Since this is a retrial system, the jobs try to find a free server from the orbit. This retrial is described with a random variable with exponential distribution. The retrial parameter is σ/N. The customers do not give up finding a free server in up state. In the models considered below, the blocked situation will be studied.
- The catastrophic break-down mode. This is the situation when a disaster event removes all of the jobs from the system (from the orbit and from the server after interrupting the service). The repair of the system starts immediately. The same breakdown parameters are used as in the common breakdown mode, i.e. γ_0 and γ_1 for an idle server breakdown and a busy server breakdown, respectively, and γ_2 for the repair. During the down period of the system, the source is blocked, no new request can enter into the system.
- The customers have impatient behavior. The breakdown event of the server is either the regular (common) break-down mode, or it is the catastrophic breakdown. The customers are impatient. Based on the retrial property, a customer retries to reach the server after a random, exponentially distributed time, or the customer gives it up, and leaves the service unit, and is transferred back to the source after a random waiting time. The distribution is exponential, and the parameter is τ. In this case, the customer will not be served.

Let us consider the dynamic workflow of the system. A job (it can be called a customer or a request, as well) is generated from the source. The job inter-arrival times toward the server are exponentially distributed random times with parameter λ/N. Until the end of service of this job, the source will not generate a new job for this token. The generated job enters the service environment. The server can be empty (or idle), or there can be another job under service. This time the server is called as busy. When there is no job at the server, the service of the incoming job, which can arrive from the source or the orbit, starts at once. The random service time intervals are considered to follow the exponential distribution law with parameter μ. When the server is in busy state, that is a job is under service at the server, and a new request is arriving to the server, the two jobs will collide. This is the phenomenon of the collision of the customers. In this case, both jobs (the job served and the job just arrived) are moved into the orbit. From the orbit the jobs retry reaching the server again after a random time interval. The distribution of the retrial times is exponential with the expectation of $1/(\sigma/N)$. See the model on Figure 1.

Let's consider the following notations. $i(t)$ denotes the state of the system. This is the total number of jobs in the system (under service and waiting in the orbit). $k(t)$ is the server

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

status descriptor:

$$k(t) = \begin{cases} 0, \text{the service unit is up and non-working,} \\ 1, \text{the service unit is up and working,} \\ 2, \text{the service unit is failed and repair is in progress.} \end{cases}$$

.

The $P(k(t) = k, i(t) = i) = P_k(i,t)$ quantities are the system probabilities: at a time of t there are number of i jobs in the system and the value of the server state descriptor is k. With these conditions the process $X(t) = \{k(t), i(t)\}$ is a Markovian-chain with two dimensions, and a state space of $\{0,1,2\}x\{0,1,...,N\}$.

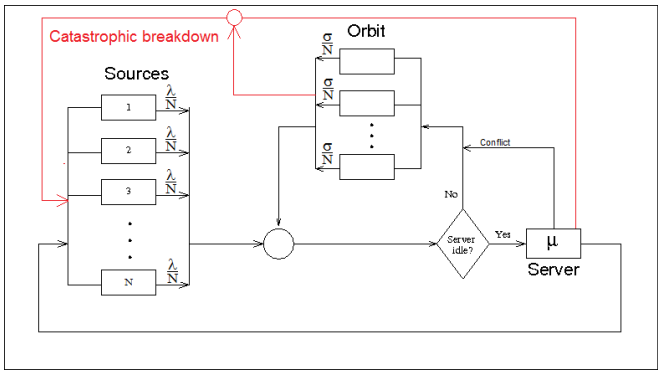


Fig. 1. System model

A successfully served job steps back to the source. All the random times, time intervals considered in the model are assumed to be totally independent of each other.

Since the $X(t) = \{k(t), i(t)\}$ process is a finite state Markov-chain, so the stability conditions hold, and the existence of the steady-state probabilities of $P_k(i,t) = P_k(i)$ is ensured.

For patient case, and with collision of jobs, the Kolmogorov balance equations for probabilities $P_k(i,t)$ can be written, as follows (see, in [13] and [14]):

$$\frac{\partial P_0(0,t)}{\partial t} = -(\lambda + \gamma_0)P_0(0,t) + \mu P_1(1,t) + \gamma_2 P_2(0,t),$$

$$\frac{\partial P_1(1,t)}{\partial t} = -\left(\lambda\frac{N-1}{N} + \mu + \gamma_1\right)P_1(1,t) + \lambda P_0(0,t) + \frac{\sigma}{N}P_0(1,t),$$

$$\frac{\partial P_2(0,t)}{\partial t} = -(\lambda + \gamma_2)P_2(0,t) + \gamma_0 P_0(0,t), \qquad (1)$$

$$\frac{\partial P_0(i,t)}{\partial t} = -\left(\lambda\frac{N-1}{N} + \sigma\frac{i}{N} + \gamma_0\right)P_0(i,t) + \mu P_1(i+1,t) + \lambda\frac{N-i+1}{N}P_1(i-1,t) + \sigma\frac{i-1}{N}P_1(i,t) + \gamma_2 P_2(i,t),$$

$$\frac{\partial P_1(i,t)}{\partial t} = -\left(\lambda\frac{N-1}{N} + \sigma\frac{i-1}{N} + \gamma_1 + \mu\right)P_1(i,t) + \lambda\frac{N-i+1}{N}P_0(i-1,t) + \sigma\frac{i}{N}P_0(i,t),$$

$$\frac{\partial P_2(i,t)}{\partial t} = -\left(\lambda\frac{N-1}{N} + \gamma_2\right)P_2(i,t) + \gamma_0 P_0(i,t) + \gamma_1 P_1(i,t) + \lambda\frac{N-i+1}{N}P_2(i-1,t).$$

Again, the $X(t) = \{k(t), i(t)\}$ process is a finite state Markov-chain, so the stability conditions hold, and the existence of the steady-state probabilities of $P_k(i,t) = P_k(i)$ is ensured.

Thus, the steady-state Kolmogorov balance equations can be formulated, as

$$-(\lambda + \gamma_0)P_0(0) + \mu P_1(1) + \gamma_2 P_2(0) = 0,$$

$$-\left(\lambda\frac{N-1}{N} + \mu + \gamma_1\right)P_1(1) + \lambda P_0(0) + \frac{\sigma}{N}P_0(1) = 0,$$

$$-(\lambda + \gamma_2)P_2(0) + \gamma_0 P_0(0) = 0, \qquad (2)$$

$$-\left(\lambda\frac{N-1}{N} + \sigma\frac{i}{N} + \gamma_0\right)P_0(i) + \mu P_1(i+1) + \lambda\frac{N-i+1}{N}P_1(i-1) + \sigma\frac{i-1}{N}P_1(i) + \gamma_2 P_2(i) = 0,$$

$$-\left(\lambda\frac{N-1}{N} + \sigma\frac{i-1}{N} + \gamma_1 + \mu\right)P_1(i) + \lambda\frac{N-i+1}{N}P_0(i-1) + \sigma\frac{i}{N}P_0(i) = 0,$$

$$-\left(\lambda\frac{N-1}{N} + \gamma_2\right)P_2(i) + \gamma_0 P_0(i) + \gamma_1 P_1(i) + \lambda\frac{N-i+1}{N}P_2(i-1) = 0.$$

From these equations, the case of a reliable server with collision of customers can be derived easily. Just give the value of zero for the parameter of γ_2, and probabilities of P_2.

Following a similar method, the system balance equations can be formulated for the case of collision, server with regular failure, and impatient jobs:

$$-(\lambda + \gamma_0)P_0(0) + \mu P_1(1) + \gamma_2 P_2(0) + \frac{\tau}{n}P_0(1) = 0,$$

$$-\left(\lambda\frac{N-1}{N} + \mu + \gamma_1\right)P_1(1) + \lambda P_0(0) + \frac{\sigma}{N}P_0(1) + \frac{\tau}{n}P_1(2) = 0,$$

$$-(\lambda + \gamma_2)P_2(0) + \gamma_0 P_0(0) = 0, \qquad (3)$$

$$-\left(\lambda\frac{N-i}{N}+\sigma\frac{i}{N}+\tau\frac{i}{N}+\gamma_0\right)P_0(i)+\mu P_1(i+1)+$$
$$+\lambda\frac{N-i+1}{N}P_1(i-1)+\sigma\frac{i-1}{N}P_1(i)+$$
$$+\tau\frac{i+1}{N}P_0(i+1)+\gamma_2 P_2(i)=0,$$

$$-\left(\lambda\frac{N-i}{N}+\sigma\frac{i-1}{N}+\tau\frac{i-1}{N}+\gamma_1+\mu\right)P_1(i)+$$
$$+\lambda\frac{N-i+1}{N}P_0(i-1)+\sigma\frac{i}{N}P_0(i)+\tau\frac{i}{N}P_1(i+1)=0,$$

$$-\left(\lambda\frac{N-i}{N}+\gamma_2\right)P_2(i)+\gamma_0 P_0(i)+\gamma_1 P_1+$$
$$+\lambda\frac{N-i+1}{N}P_2(i-1)+\tau\frac{i+1}{N}P_2(i+1)=0.$$

III. Performance Characteristics

Investigating the effect of the different parameters on the behavior of the system, the usual performance characteristics can be calculated from the steady-state probabilities.

- Average number of customers in the system $\overline{Q}$ and in the orbit $\overline{O}$

$$\overline{Q}=\sum_{i=0}^{N}iP(i),\qquad \overline{O}=\overline{Q}-P_1,$$

- Average arrival rate $\overline{\lambda}$

$$\overline{\lambda}=\sum_{k=0}^{1}\sum_{i=0}^{N}(N-i)\frac{\lambda}{N}P_k(i),$$

- Average response time $\overline{T}$ and mean waiting time $\overline{W}$ in the orbit can be obtained by the Little-formula

$$\overline{T}=\frac{\overline{Q}}{\overline{\lambda}},\qquad \overline{W}=\frac{\overline{O}}{\overline{\lambda}},$$

$$\overline{O}=\overline{Q}-P_1,$$

- Average total service time $E(T_S)$ and average total sojourn time in the source $E(\kappa)$

$$E(T_S)=\overline{T}-\overline{W},\qquad E(\kappa)=\frac{(N-\overline{Q})\overline{T}}{\overline{Q}},$$

- Average number of trials from the source $E(N_{TS})$ and from the orbit $E(N_{TO})$

$$E(N_{TS})=\frac{\lambda}{N}E(\tau),\qquad E(N_{TO})=\frac{\sigma}{N}\overline{W}.$$

$$\overline{Q}=\sum_{i=0}^{N}iP(i),\qquad \overline{O}=\overline{Q}-P_1.$$

IV. Numerical results

The steady-state equations can be solved by different methods. Here an analytical software tool, namely the MOSEL-2 was chosen. This tool formulates the underlying Markovian-equations of the system, and provides an algebraic solution for the steady-state probabilities. With the assumption of exponentiality of the system parameters, this tool is effective and quick for a reasonably large number of sources.

Figure	λ	μ	σ	τ	$\gamma_0 = \gamma_1$	γ_2
2	0.1 .. 10	1	5	Legend	0.01	1
3	03.7 .. 7.7	1	5	Legend	0.01	1
4	0.1 .. 10	1	5	Legend	0.01	1
5	1	1	5	Legend	0.001 .. 0.111	1
6	1	1	5	Legend	0.001 .. 0.111	1
7	1	1	5	Legend	0.001 .. 0.111	1

TABLE I
NUMERICAL VALUES OF MODEL PARAMETERS

The collision of customers and catastrophic breakdown features are applied for all of the following figures. In the MOSEL program the idle and the busy state failure rates (γ_0 and γ_1 are the same. The number of sources is $N = 100$. All the other system parameters are listed in Table I.

The dynamic behavior of the system can be seen on the figures. Different system characteristics are displayed in function of overall generation rate and the failure rate.

All figures compare three different cases with respect to different values of the impatient parameter, τ. The first one is a small parameter value. In this case, the expectation of patient time interval is large. This case corresponds to the patient behavior of the customers. The other two values of the impatient parameter are medium and large.

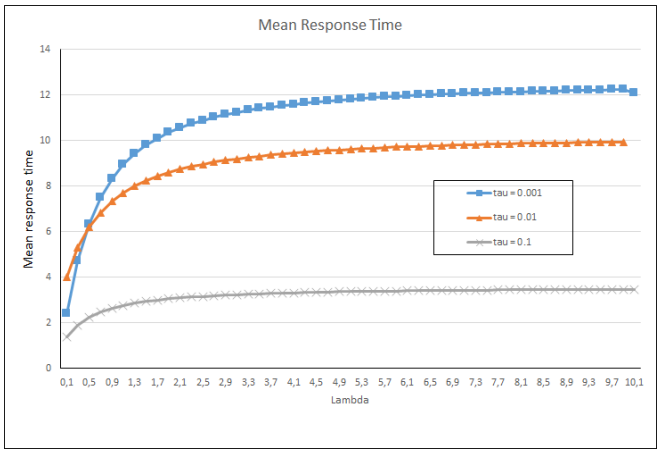


Fig. 2. Mean response time vs. generation rate

On Figures 2, 3, and 4 the overall generation rate, λ is the running parameter on the X-axes. On Figure 2 the mean response time is displayed. For the patient behavior, this performance measure has larger values. In this case, the customers do not leave the system from the orbit.

Figure 3 shows the utilization of the server. For larger generation rates larger utilization can be observed. It is interesting, that for larger impatient parameters larger increments of the

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

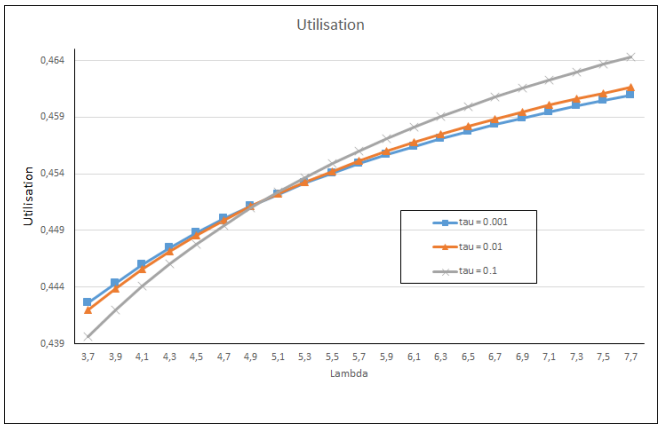


Fig. 3. Utilization vs. generation rate

corresponding curves are present. In addition, the curves have a single intersection point, which means, for a given generation rate (λ=5), the same utilization can be observed with a different impatient rate. This phenomenon might be resulted from a special coincidence of the collision, the catastrophic breakdown, and the impatient properties.

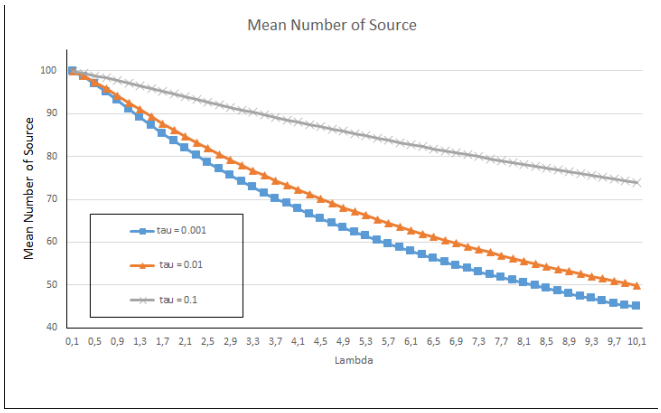


Fig. 4. Mean number of source vs. generation rate

On Figure 4 the number of free tokens in the source can be seen. The curves have a decrement in their slopes. The higher generation rate implies a higher number of collisions, so, the customers will fill up the orbit. The patient customers remain in the orbit, consequently, the corresponding curve has the smallest values.

On Figures 5, 6, and 7 the failure rates are the running parameter. The figures display the same system characteristics, as Figures 2, 3, and 4. Due to the increasing catastrophic failure generation rates, the mean response time and the utilization have decreasing curves. Because of the frequent catastrophic breakdown, all of the customers leave the systems at every breakdown event, thus the number of sources will increase. These general trends are refined with the patient / impatient behavior of the customer.

V. Conclusion

In this paper, the interaction of the behaviors of collision, catastrophic breakdown, and impatient customers has been investigated. The impatient property and the catastrophic breakdowns result extra dimension in the Kolmogorov balance equations, so the recursive numerical solution can not be provided. The MOSEL-2 tool was used for solving the system equations and calculating the steady-state probabilities. Due to the finite state space, these probabilities exist, there are no stability problems. With the help of the system probabilities, reliability investigations can be performed. In this paper, the most important and interesting system measures were presented. A lot of system parameters were tried. Those parameters were chosen, where the considered performance measures display significant deviation between the scenarios.

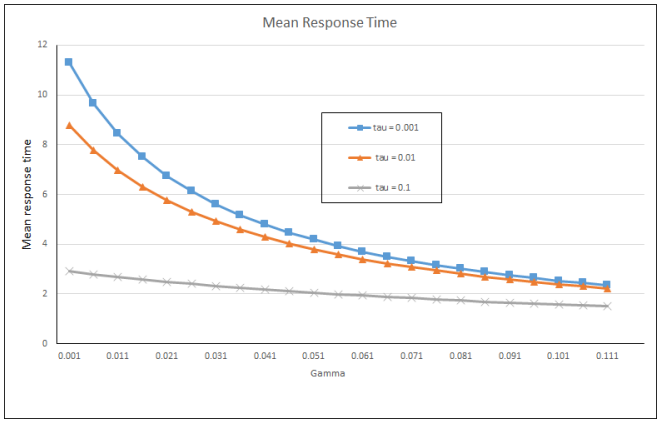


Fig. 5. Mean response time vs. failure rate

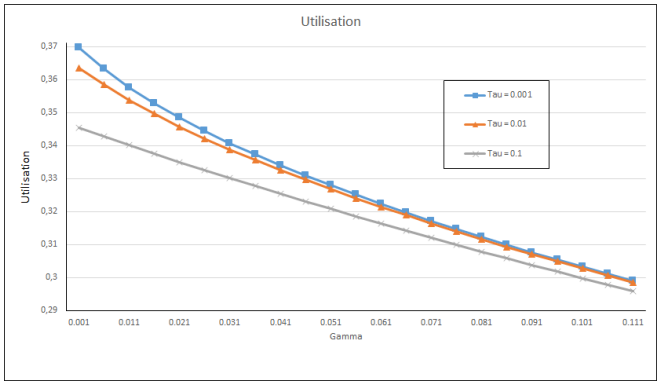


Fig. 6. Utilization vs. failure rate

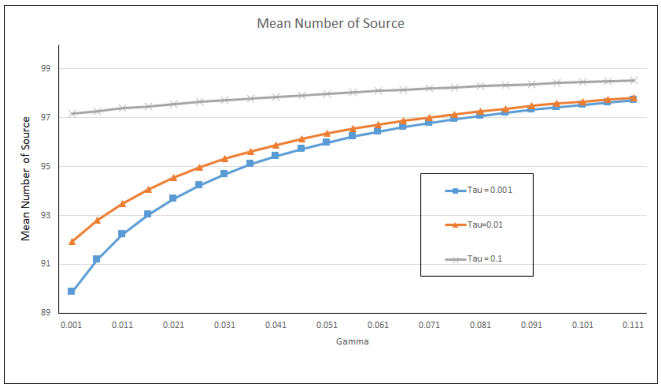


Fig. 7. Mean number of source vs. failure rate

ACKNOWLEDGMENT

The research work was supported by the construction EFOP - 3.6.3 - VEKOP - 16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

The research work was supported by the Austro-Hungarian Cooperation Grant No 106öu4, 2020.

REFERENCES

[1] V. Anisimov, "Asymptotic analysis of highly reliable retrial systems with finite capacity," *In: Queues, Flows, Systems, Networks. Proceedings of the International Conference Modem Mathematical Methods of Investigating the Telecommunicational Networks*, pp. 7–12, 1999.

[2] J. Artalejo and A. G. Corral, *Retrial Queueing Systems: A Computational Approach*. Springer, 2008.

[3] J. Wang, L. Zhao, and F. Zhang, "Analysis of the finite source retrial queues with server breakdowns and repairs," *Journal of Industrial and Management Optimization*, vol. 7, no. 3, pp. 655–676, 2011.

[4] V. I. Dragieva, "A finite source retrial queue: number of retrials," *Communications in Statistics-Theory and Methods*, vol. 42, no. 5, pp. 812–829, 2013.

[5] J. Kim and B. Kim, "A survey of retrial queueing systems," *Annals of Operations Research*, vol. 247, no. 1, pp. 3–36, 2016.

[6] B. Almási, T. Bérczes, A. Kuki, J. Sztrik, and J. Wang, "Performance modeling of finite-source cognitive radio networks." *Acta Cybern.*, vol. 22, no. 3, pp. 617–631, 2016.

[7] B. Almási, J. Roszik, and J. Sztrik, "Homogeneous finite-source retrial queues with server subject to breakdowns and repairs." *Math. Comput. Modelling*, vol. 42, no. 5-6, pp. 673–682, 2005.

[8] V. I. Dragieva, "Number of retrials in a finite source retrial queue with unreliable server." *Asia-Pac. J. Oper. Res.*, vol. 31, no. 2, p. 23, 2014.

[9] N. Gharbi and C. Dutheillet, "An algorithmic approach for analysis of finite-source retrial systems with unreliable servers," *Computers & Mathematics with Applications*, vol. 62, no. 6, pp. 2535–2546, 2011.

[10] F. Zhang and J. Wang, "Performance analysis of the retrial queues with finite number of sources and service interruptions," *Journal of the Korean Statistical Society*, vol. 42, no. 1, pp. 117–131, 2013.

[11] A.-A. Ali and S. Wei, "Modeling of coupled collision and congestion in finite source wireless access systems," in *Wireless Communications and Networking Conference (WCNC), 2015 IEEE*. IEEE, 2015, pp. 1113–1118.

[12] S. Balsamo, G.-L. Dei Rossi, and A. Marin, "Modelling retrial-upon-conflict systems with product-form stochastic Petri nets," in *International Conference on Analytical and Stochastic Modeling Techniques and Applications*. Springer, 2013, pp. 52–66.

[13] A. Kvach and A. Nazarov, *Sojourn Time Analysis of Finite Source Markov Retrial Queuing System with Collision*. Cham: Springer International Publishing, 2015, ch. 8, pp. 64–72.

[14] A. Nazarov, A. Kvach, and V. Yampolsky, *Asymptotic Analysis of Closed Markov Retrial Queuing System with Collision*. Cham: Springer International Publishing, 2014, ch. 1, pp. 334–341.

[15] T. V. Lyubina and A. A. Nazarov, "Research of the non-markov dynamic retrial queue system with collision (in russian)," *Herald of Kemerovo State University*, vol. 1, no. 49, pp. 38–44, 2012.

[16] Y. Peng, Z. Liu, and J. Wu, "An $M/G/1$ retrial G-queue with preemptive resume priority and collisions subject to the server breakdowns and delayed repairs." *J. Appl. Math. Comput.*, vol. 44, no. 1-2, pp. 187–213, 2014.

[17] S. Subramanian *et al.*, "A stochastic model for automated teller machines subject to catastrophic failures and repairs," *Queueing Models and Service Management*, vol. 1, no. 1, pp. 75–94, 2018.

[18] B. Thilaka, B. Poorani, and S. Udayabaskaran, "Performance analysis for queueing systems with close down periods subject to catastrophe," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 7, pp. 39–57, 2018.

[19] U. Gupta, N. Kumar, and F. Barbhuiya, "A queueing system with batch renewal input and negative arrivals," in *Applied Probability and Stochastic Processes*. Springer, 2020, pp. 143–157.

[20] D. Piriadarshani, S. Narasimhan, M. Maheswari, B. James *et al.*, "A retrial queuing system operating in a random environment subject to catastrophes," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 2, pp. 5029–5032, 2020.

[21] S. I. Ammar, A. Zeifman, Y. Satin, K. Kiseleva, and V. Korolev, "On limiting characteristics for a non-stationary two-processor heterogeneous system with catastrophes, server failures and repairs," *Journal of Industrial & Management Optimization*, vol. 17, no. 3, p. 1057, 2021.

Attila Kuki Born in 1964. Graduated in 1989 in mathematical statistics and English-Hungarian translator. Got PhD in 2007 in mathematics. Current position is a lecturer at University of Debrecen, Faculty of Informatics, Hungary. Educational tasks: statistics, stochastic simulation, networking, information systems. Research fields: stochastic simulation, performance evaluation of infocommunication systems. International cooperations (visits): Barcelona, 2 weeks, Erlangen, 4 and 2 weeks, Peking, 2 weeks.

Ádám Tóth received his MSc Degree in Informatics in 2016, at the University of Debrecen, Hungary. He is a graduaded PhD student (2021) and member of the Department of Informatics Systems and Networks of the same university. His main research field is the performance analysis of retrial queues with finite number of sources and their application.

Tamás Bérczes received his MSc Degree in Mathematics in 2000 at the University of Debrecen, Hungary. He received a Ph.D. degree in 2011, and habilitation in 2017. He is currently Associate Professor at the Department of Informatics Systems and Networks of the same university. His primary research interests are performance analysis of retrial queues and their application.

János Sztrik is a Full Professor and studied mathematics at University of Debrecen 1973-1978. Obtained the M.Sc. in 1978, Ph.D. in 1980 both in probability theory and mathematical statistics from the University of Debrecen. Received the Candidate of Mathematical Sciences degree in probability theory and mathematical statistics in 1989 from the Kiev State University, USSR, habilitation from University of Debrecen in 2000, Doctor of the Hungarian Academy of Sciences in 2002. His research interests are in the field of production systems modelling and analysis, queueing theory, reliability theory, and computer science.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Reflection of the COVID-19 pandemic in mass media

Kirill Yakunin, Ravil I. Mukhamediev, Marina Yelis
Satbayev University (KazNRTU)
Almaty, Kazakhstan
ravil.muhamedyev@gmail.com

Adilkhan Symagulov, Yan Kuchin, Elena Muhamedijeva
Institute of Information and Computational Technologies
Almaty, Kazakhstan
ykuchin@mail.ru

Jan Rabcan
University of Zilina
Zilina, Slovakia

Aubakirov Margulan
Maharishi International University
Fairfield, Iowa

***Abstract*— The paper analyzes the "reflection" of the COVID-19 theme in the mass media of the Republic of Kazakhstan and the Russian Federation. Used a corpus of mass media of about 1 million publications and WHO data. Using topic modeling and analysis of correlation dependences, both the similarity and some difference in the publication activity of mass media were established. In particular, in both countries, media activity correlates with the absolute indicators of the number of new infected and the number of deaths. Correlation coefficient is 0.6-0.8. However, mass media take into account the number of positive rate, reproduction rate weaker. Correlation coefficient is 0.4-0.6. In the Russian media, there is a high correlation (0.75) with changes in measures to limit the spread of infection taken by the state, while in Kazakhstan this correlation is much lower (0.4).**

Keywords*— *COVID-19, mass media analysis, topic modeling, BigARTM

I. INTRODUCTION

The health care system as a factor of sustainable and stable growth of welfare is one of the main priorities in Kazakhstan. The report of the Head of State "The Third Modernization of Kazakhstan: Global Competitiveness" dated March 3, 2017, includes health issues [1], the State Program of Health of the Republic of Kazakhstan "Densaulyk", the main goal of which is to strengthen public health to ensure sustainable socio-economic development of the country. However, both in Kazakhstan and on a global scale, health care systems are faced with a multitude of problems causing increased demand for health services, high public expectations, and rising costs [2]. Relative inefficiency and low productivity in the health sector have led to increased fiscal restrictions, which in turn have contributed to increased social tensions and a decline in economic growth during pandemics, which have been highlighted by COVID-19 [3]. Not only economic but also social and medical efficiency is important in the health care system; "medical measures of therapeutic and preventive nature may be economically unprofitable, but medical and social effects require them" [4].

At the same time, Inadequate representation of health authorities in the media space contributes to the spread of rumors and misinformation [5], affects the mental health of the population [6]. The work assesses the media effect of COVID-19, which manifested itself during the pandemic in Kazakhstan and Russia. It is analyzed as the similarities and differences in the reflection in the mass media of processes in health care and society associated with COVID-19.

In the process of the research the following corpus have been used [7]. The architecture of the system used during the research and methodology fundamentals are described in [8].

II. METHODS AND DATA

Corpus of news from Russian and Kazakhstani media sources from 2000 to 2020 from 30 major sources, including social networks (VK.com, YouTube, Instagram and Telegram) and news websites [7]. It includes 4,233,990 documents from Kazakhstani sources and 2,027,963 documents from Russian sources (Figure 1).

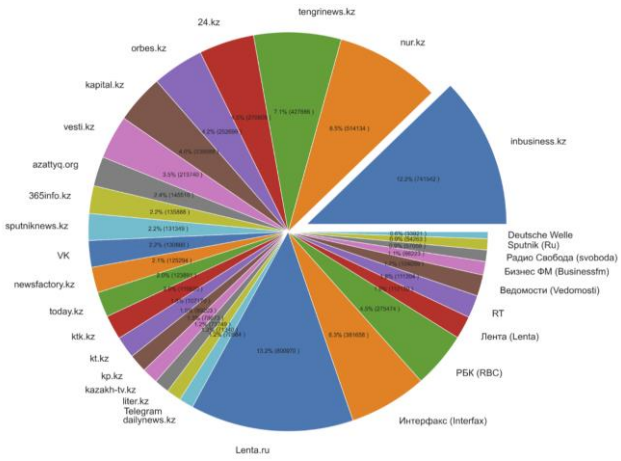


Figure 1. Major sources of the corpus.

A. COVID data

In order to analyze the objectivity of reflection of COVID-19 pandemic in mass-media, data from Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [8] was used; namely, the following daily indicators were considered separately for Kazakhstan and Russian Federation:

- Number of new cases smoothed according to CSSE-methodology.
- Number of new deaths smoothed according to CSSE-methodology.
- Virus reproduction rate.
- Number of new test.
- Positive rate – ratio of positive tests results among all tests performed.
- Tests per case – the indicator is inversed positive rate.
- Stringency index – measure of strictness of restriction due to pandemic, as per WHO-methodology.

Our approach of media analysis is based on topic modeling (TM) [9]. TM is a method based on the statistical characteristics of document collections, which is used in the tasks of automatic summarization, information retrieval and clustering. TM transforms to the algorithm the intuitive understanding that documents in a collection form groups in which the frequency of occurrence of words or word combinations differs. The basis of TM is the statistical model of natural language. Probabilistic TM describes documents (M) by a discrete distribution on a set of topics (T) and topics by a discrete distribution on a set of the term [9].

To build a thematic model of the corpus of documents, a very popular latent Dirichlet allocation (LDA) [10] is used. Some generalization of LDA is additive regularization of topic models (ARTM), implemented in the form of BigARTM library [11].

We have applied BigARTM topic model on corpus of 1 116 272 news publication from over 30 major Russian and Kazakhstani internet media sources.

The time window that was considered spans from 01.01.2020 to 25.02.2021. Initially a topic model with 200 topics was computed, which we refer to as initial topic model or level-0 topic model. Then topics related to medicine, pandemic and healthcare were manually selected by experts and a sub-corpus was filtered by selecting only news publications, which related to the selected topics with weight from θ-matrix higher than a pre-defined threshold (0.05 in this case). This topic model is referred to as level-1 topic model on medicine. However, the level-11 topic model on medicine contained a portion of irrelevant or controversial documents and topics, and in order to improve the quality of the sub-corpus relating to medicine and healthcare, the described process was repeated for the purpose of obtaining level-2 and level-3 models.

When performing final analysis, three topic models were used:

- Level-0 (initial) topic model, which contains general topics on main areas of human activity, including economy, politics, culture, education, etc.
- Level-2 – contains topics mostly related to medicine, pandemic, etc., but also contains some overlapping topics, such as restrictions in education, sports and public life due to the pandemic, overall economy situation, etc.
- Level-3 – contains only topics relevant to medicine and healthcare.
- Level-1 topic model didn't seem to be applicable for the planned analysis, due to large proportion of irrelevant topics.

Than for each of the topic models under consideration topics relative weight daily dynamics were calculated. Relative weight of a topic is a ratio between sum of weights of all documents published on a given day to the given topic to sum of the weights of the documents to all topics. In other words, it's a ratio of a column of θ-matrix, representing the given topic to the sum of the whole θ-matrix. So, this indicator shows ratio of information that relates to the given topic on a given date and ranges from 0 to 1.

B. Method

To extract the thematic structure of the corpus, a method based on the BigARTM thematic model is used. BigARTM is an open source library for thematic modeling of large collections of text documents and transactional data sets. The thematic model determines which topics each document belongs to, and what words describe each topic. The model is trained without a teacher (unsupervised learning) [11]. Then Pearson correlation coefficient is calculated. This coefficient is used to describe the correlation between two groups of different data i.e., the correlation coefficient between each independent variable and the dependent variable. Pearson correlation coefficient is calculated between the topics daily relative weights and the abovementioned WHO indicators, and then to find the topics with the highest correlation. The advantage of using this method is that the relationship between variables can be clearly defined and quantified. Then the obtained information can be used for analysis and interpreted by experts. It is proposed to perform this analysis due to the following filters and parameters:

1) For Russian Federation and the Republic Kazakhstan separately.
2) For each of the three levels of topic models separately.
3) For each of the 7 WHO indicators selected for analysis.

III. Results and Discussions

By applying the proposed method 42 sorted topics lists with correlation to corresponding indicators where obtained, which constitute analysis across 2 countries, 3 topic models and 7 different indicators. Than the lists of top and bottom topics with corresponding Pearson correlation values were manually analyzed by exerts in order to attempt to draw some

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

conclusions. Table 1 illustrates an example of data under analysis.

TABLE I. CORRELATION BETWEEN NUMBER OF NEW DEATHS FROM COVID AND MASS MEDIA TOPICS FROM LEVEL-0 (INITIAL) TOPIC MODEL

Correlation	*Topic name (top-words)*	*Topic volume (documents)*
0.91	Vaccine, Vaccination, Drug, Coronavirus, Test, Satellite, Russian	15434
0.86	Petersburg, Saint Petersburg, Petersburg, Leningrad region, Moscow, report_deaths, COVID	1495
0.77	Health, Product, Doctor, Alcohol, Organism, Nutrition, Healthy	8318
0.74	Tell, Photo, Arrive, Depart, Tourism, Return, Go	2693
0.67	Temperature, Degree, Night, Snow, Weather, Air, Strong	8196

The following conclusions were made in the course of interpreting the obtained results.

- If we consider different COVID indicators, new deaths and new cases tend to have higher maximum correlation across countries and topic models (typically around 0.6-0.8). However more objective indicators, such as reproduction rate, positive rate and number of tests per case tend to have lower maximum correlation (typically around 0.4-0.6). Hence, we argue that usually mass media tend to react in biased fashion, inflating of deflating public opinion on epidemiological situation on the basis of absolute numbers. For example, number of new cases is useless, unless it is known how many tests were performed, which makes tests positive rate more useful. However, media in Kazakhstan and Russia don't seem to consider such data.
- In Russian Federation there are topics with high correlation to stringency index (0.75), while in Kazakhstan the highest correlation to stringency index is only 0.43. This indicates that in Kazakhstan there is a dissonance between pandemic restrictions and how they are reflected in media.

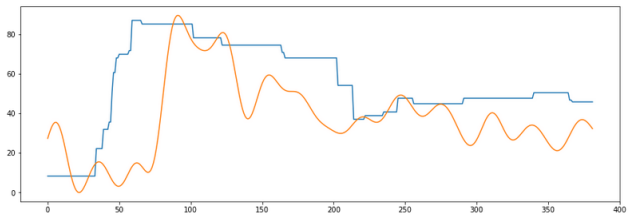


Figure 2. Blue line – stringency index in Russia, yellow line – topic with highest correlation (0.75), relating to reports about COVID cases across different regions.

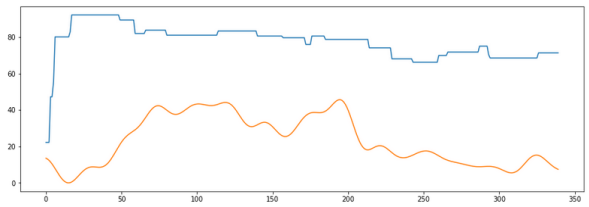


Figure 3. Blue line – stringency index in Kazakhstan, yellow line – topic with highest correlation (0.43), relating to reports about COVID cases across different regions.

- In Kazakhstan a topic with the highest correlation (0.8) to number of new deaths is a topic regarding "oxygen, investigation, embezzlement", which may be interpreted as one of the reasons for spikes in number of deaths. At the same time, in Russia the topics with highest correlation to number of new deaths (0.82-0.84) are related to vaccination and new Sputnik-V vaccine. There seemed to be an attempt to smooth the panic among people by providing them with perspectives of ways to stop the epidemy.
- While counterintuitive, there's a considerable negative correlation (-0.3 - -0.4) between number of new deaths and cases and topics related to economy and banking. Seems like although the negative impact of the pandemic on economy is obvious, during periods of harsher epidemiological situations media didn't pay much attention to economy and politics, while other, more neutral topics, such as celebrities, lifestyle, culture and art don't seem to have any considerable correlation with epidemiological indicators.

IV. CONCLUSION

To conclude, a method for analyzing how objective statistical indicators related to COVID are being reflected in mass media was proposed. The method is based on applying BigARTM topic model in order to obtain topical structure of the corpus, to consequently perform correlation analysis with COVID indicators, such number of new cases, positive tests rate, stringency index and others.

The proposed method exhibits high potential for obtaining insights on how COVID is presented in the media and what type of statistics tend to drive media activity. For example, it has been identified, that media in Russia and Kazakhstan tend to focus on absolute values of new cases/deaths per day, while more informative indicators like positive tests rate and virus reproduction rate are not being covered and don't show significant correlation with publication activity.

Moreover, the method might be applied to assess how severity of restrictions is being reflected in media, identify topics which might indicate reasons for degradation of epidemiological situation,

Directions of further research include:

- Perform sentiment analysis of news and compare them with COVID data.

- Attempt to recognize possible statistical inaccuracies of official COVID-19 indicators by using mass media data.
- Build and analyze topical profile of COVID-19 pandemic and how it changed in time, in order to attempt to asses it's impact on economy, education, politics, tourism, etc.

ACKNOWLEDGMENT

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. IRN AP09259587 "Developing of methods and algorithms of intelligent GIS for multi-criteria analysis of healthcare data").

This work was partially supported by the Slovak Research and Development Agency, no. APVV PP-COVID-20-0013 "Development of methods of healthcare system risk and reliability evaluation under coronavirus outbreak".

REFERENCES

[1] Message from the President of the Republic of Kazakhstan N. Nazarbayev to the people of Kazakhstan. 31 January 2017 [Electronic resource]. - Access mode: https://www.akorda.kz/ru/addresses/addresses_of_president/poslanie-prezidenta-respubliki-kazahstan-nnazarbaeva-narodu-kazahstana-31-yanvarya-2017-g

[2] Atun R. "Transitioning health systems for multimorbidity," The Lancet, vol. 386., no. 9995. pp. 721-722, 2015.

[3] Baldwin, Richard, and B. Weder Di Mauro. "Economics in the time of COVID-19: A new eBook." VOX CEPR Policy Portal,2020.

[4] Orlov. E.M., Sokolova, O. N. "Category of effectiveness in the health care system." Fundamental research, no.4, 2010 (In Russian).

[5] Tasnim S., Hossain M. M., Mazumder H. "Impact of rumors or misinformation on coronavirus disease (COVID-19) in social media," Journal of preventive medicine and public health, vol. 53. no. 3, pp. 171-174, 2020.

[6] Gao J. et al. "Mental health problems and social media exposure during COVID-19 outbreak," Plos one, vol. 15, no. 4, p. e0231924, 2020. Yakunin, K., Kalimoldayev, M., Mukhamediev R., et al. "KazNewsDataset: Single Country Overall Digital Mass Media Publication Corpus," Data, vol. 6, no.3, p.31, 2021.

[7] Mukhamediev, R.I.; Yakunin, K.; Mussabayev, R.; Buldybayev, T.; Kuchin, Y.; Murzakhmetov, S.; Yelis, M. "Classification of Negative Information on Socially Significant Topics in Mass Media." Symmetry, no.12, p.1945, 2020

[8] Dong E, Du H, Gardner L. "An interactive web-based dashboard to track COVID-19 in real time," The Lancet infectious diseases, vol. 20, no. 5, pp. 533-534, 2020.

[9] Vorontsov, K.V., Potapenko, A.A. "Regularization, robustness and sparseness of probabilistic thematic models," Comput. Res. Modeling, vol. 4, pp. 693–706, 2012.

[10] Hamed, J.; Yongli, W.; Chi, Y.; Xia, F. "Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey," Multimed. Tools Appl., vol. 78, pp.15169–15211, 2017.

[11] Vorontsov, K.; Frei, O.; Apishev, M.; Romov, P.; Dudarenko, M. "BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections." In International Conference on Analysis of Images, Social Networks and Texts; Springer: Cham, Switzerland, pp. 370–381, 2015.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Software based support of curriculum mapping in education at medical faculties

Jaroslav Majerník[1], Martin Komenda[2], Andrzej Kononowicz[3], Inga Hege[4], Adrian Ciureanu[5]
[1] Department of Medical Informatics, Pavol Jozef Šafárik University in Košice, Faculty of Medicine, Košice, Slovakia, jaroslav.majernik@upjs.sk
[2] Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, Czech Republic
[3] Department of Bioinformatics and Telemedicine, Uniwersytet Jagiellonski, Krakow, Poland
[4] Medical School, Universitaet Augsburg, Augsburg, Germany
[5] Medical Informatics and Biostatistics, University of Medicine and Pharmacy of Iasi, Iasi, Romania

***Abstract*—The quality of education closely relates to the quality of curriculum and its management. This also applies to medical education, which is also changing with the development of modern technologies. Therefore, we summarised the best experience and knowledge of curriculum mapping and brought to our faculties a pioneering concept, including expertise in selected medical disciplines. The main results of our efforts include the delivery of methodologies and a unified ICT platform developed for optimisation of curricula. Based on users' needs and requirements, the anatomy as one of the core preclinical study discipline and five different medical disciplines were completely mapped and optimised to prove modernizing effects as well as practical application of our platform. Thanks to the open and modular approach of platform's framework, the interdisciplinary study programmes can be easily designed, mapped and managed, and the redundancies and the gaps in curricula can be detected too. Curricula designers are equipped by the visual tools and feedbacks allowing them effective administration of particular study blocks as well as to communicate with teachers and educators to adapt and develop curricula to the recent trends and knowledge.**

Keywords—medical education; curriculum management; curriculum mapping, software platform

I. INTRODUCTION

Development of modern information and communication technologies and advances in data processing techniques brought various systems and platforms supporting curriculum management. However, to our knowledge, there is still no systematic solution based on proven pedagogical approaches and methodologies designed to define, create, manage and analyse the curricula of medical and healthcare institutions of higher education within one robust system [1]. Moreover, there is a global need for a uniform curriculum model providing a general and standardized way to describe the building blocks and the attributes of education using predefined parameters [2]. On the other hand, the research teams and academic institutions develop and improve their systems to meet most of the local as well as national requirements, depending on their educational systems. The electronic curriculum management and mapping systems reported during the period of last decade presented advances that covers more and more features, graphical outputs and many additional supporting tools for curriculum designers. An example of such systems is Learning Opportunities, Objectives and Outcome Platform (LOOOP) [3]. The system was developed by Charité – Universitätsmedizin Berlin hospital, and is currently in use by several German medical faculties and used as tool to further development of NKLM (German National Medical Competencies Catalogue). Another system for collaborative authoring and managing of learning objectives developed at the RWTH Aachen (Germany) using Semantic Web technologies is the ACLO-Web system [4]. This system is based on a MediaWiki platform and enables adding learning objectives for all courses in the curriculum. System's module ACLO CM is dealing with consistency checking of the curriculum. System "medtrics" is in use at California University of Science and Medicine and it is a proprietary product implemented by a commercial company [5]. The systems support curriculum analysis by showing the relationship across different levels of learning outcomes and their associated learning events, pedagogy, and assessment. A joint project of several German universities led by University of Tübingen resulted in the system MERLIN [6]. Using this system, six faculties completed mapping of the whole curriculum, implemented several visual analytics presentations and integrated with above mentioned NKLM. Prudentia, the web-based curriculum mapping system developed at the University of Notre Dame, Australia highlighted the need for an interface combining the functionality of curriculum mapping system with Blackboard Learning Management System [7]. Such interfaces help in implementing curriculum mapping in blended learning curricula which is postulated at an increasing number of higher education institutions. A modular curriculum mapping system OPTIMED (OPTImized MEDical education) was developed at Masaryk University, Brno, Czechia [8]. This system consists of four components: a learning outcome register; learning unit register; curriculum browser and reporting and export module. OPTIMED facilitated outcome-based education compatibility through employment of structured entry of new learning objectives, where the users are able to select action verbs for the statements in accordance with the recommended vocabulary of Bloom's taxonomy [9]. Building upon experiences of Masaryk university team, we created a cooperation of five universities

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

with the aim to transfer and share curriculum management and mapping knowledge. For the purposes of speeding and improving the long-term process of medical and healthcare curricula harmonization, we decided to develop an innovative and well-structured electronic system for curriculum optimisation. Development was based on detailed needs analysis, which generated a set of local institutional requirements related to the goals, aspirations and current features of curriculum organization.

II. Materials and Methods

In order to identify individual expectations and requirements on Curriculum management system in details, the combination of online survey with personal discussions was realized. The online survey was developed using Google Form. Personal discussions were conducted based on bilateral meetings either directly at the institution from which we collected the requirements or using the videoconferencing technologies. Curriculum system's key characteristics and features derived from the needs analysis include: online availability; visual overview of curriculum, integration of different user roles; export of curricula by course, study field, department, faculty; visual relations between various components of curriculum; possibilities to search by keywords; integration of international recommendations; possibility to modify reports and outputs according to the institutional requirements; evaluation of learning objectives; identification of redundancies in learning objectives; outcome-based education compatibility and complex reporting based on available curriculum building blocks. To meet all required characteristics a systematic search of literature published in the last five years, enriched with experiences of the five project partners was conducted and resulted in a comprehensive overview of what and how is nowadays implemented in curriculum mapping software [10].

Considering all the above mentioned requirements and information about recent trends in curriculum mapping we developed an EDUportfolio platform. This web-based platform represents a completely new solution and was developed under direction of Masaryk university. All partners participated in the process of implementation by their contributions to both the technical and methodical aspects. One platform for all universities is used, as we agreed it is more effective compared to the operation of several independent platforms, where the inter-institutional analysis and comparisons of curricula will require external modules and/or systems. In addition, the localizations for all partner languages were developed and implemented. The entire system is based on the curriculum description, which makes it possible to define particular teaching blocks in a parametric and thus structured way (e.g. for study program, medical discipline, course, learning unit, learning outcome) in accordance with international standards provided by the MedBiquitous association.

The EDUportfolio was built on the modern PHP Symfony 4.4 framework together with the Twig template engine and the Doctrine ORM library for object mapping. In terms of databases, the PostgreSQL open-source object-relational database system was used. Yarn was used to manage dependencies on the frontend, and Composer was used to manage backend dependencies for third-party libraries. The Zurb Foundation framework, using the jQuery library, was used to develop a responsive frontend. The asset administration is dealt by the webpack's derivate so-called webpack-encore, which comes with the Symfony. Various JavaScript libraries were used for the purposes of individual modules, such as d3.js, NVD3 and Datatables (interactive visualisations of data), or select2.js, sweetalert2.js, jstree.js and featherlight.js (improvements in user experience).

The newly developed and implemented curriculum management platform EDUportfolio, including its local instances in five different languages (German, Czech, Polish, Romanian and Slovak) was developed in full compliance with the requirements of individual institutions. To verify this, it was necessary to perform various types of tests before the final version of the platform is released to all partner institutions. The individual functional as well as user acceptance tests were performed continuously and repeatedly together with the progress of the web-based platform development and piloting of its practical usage at all partner institutions. Basic unit tests, integration tests and system tests of EDUportfolio, aimed to verify that the individual system components and modules perform tasks as designed. The tests were done by the technical development team during the implementation. After the development and release of the system modules for the consortium, which used them to map the curriculum of Anatomy as well as curriculums of all complementary disciplines, the exploratory testing and the user acceptance testing were conducted. In addition, a set of semi-automated and automated scripts was developed and applied to simplify the verification of EDUportfolio functionalities.

III. Results

EDUportfolio platform is based on the proven methodology, which follows curriculum designers during a complicated process of definition of learning outcomes and learning units devoted to particular courses. EDUportfolio uses several descriptive attributes to specify curriculum in a structured form suitable for data processing (Tab. 1).

TABLE I. Descriptive attributes used in EDUportfolio

Parameters	Description
Category	Program-level, Competency-level, Sequence block-level
Assessment form	Form of assessment of students' knowledge and skills
Duration of teaching	In teaching hours
Type of teaching	E.g. lecture, seminar, clinical practice, self-study, etc.
Importance	Initial brief explanation (why a building block occurs, what is its relevance and/or how it contributes to the learning outcomes)
Description	Short summary of a building block along with educational goals and instructional description
Keywords	Free text based on standardized keyword form
Significant terms	Fundamental topic contained and explained during teaching period in a tree structure
Study materials	Recommended literature or e-learning information source

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

A complete curriculum consists of individual building blocks that represent basic units for curriculum development. The table 2 summarises main components of curriculum building blocks as it is used in EDUportfolio.

TABLE II. COMPONENTS OF CURRICULUM BUILDING BLOCKS

Components of building block	Content/Value
Study program	Medicine/General medicine (depending on partner country)
Medical and health discipline	Anatomy
Sequence block	Course, Module, Unit, Block, Clerkship
Event	Instructional or Assessment Session
Competence	Learning outcomes, competencies, learning objectives, professional roles, topics, classifications (measurable description of what students are able to demonstrate in terms of knowledge, skills and values)

The central instance of EDUportfolio, localised in English, was used to manage curricula of Anatomy (Fig. 1). In addition to this central EDUportfolio platform, we implemented another five instances in partner's local languages. The platforms are available for the needs of individual partner institutions to map and describe curricula of complementary disciplines. However, all partners can independently manage any other curricula of their study disciplines or courses using the provided methodology.

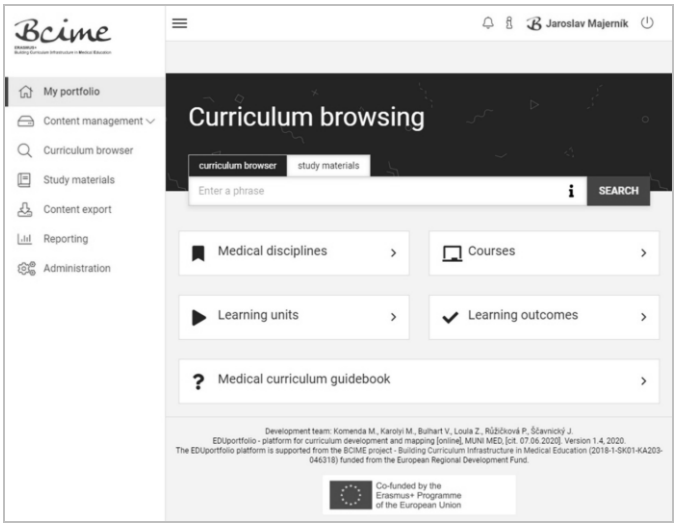


Figure 1. Central instance of EDUportfolio.

EDUportfolio platform consists of modules, through which the individual parts of curriculum are managed by the authorised users. The modules were designed with the aim to be clear and easy to use, even for novices who want to transform for example their paper-based approach of curriculum management to the modern electronic and structured form with all its advantages. The individual modules offer features and functions supporting activities that were identified in the curriculum management related processes. The list of modules include user administration module, curriculum development modules, module for curriculum browsing, module for study materials browsing, module for curriculum reporting, and module for publications.

There are currently four different user roles specified in EDUportfolio: administrator - person authorised to manage system and its core features; student - person authorised to see and browse the content of completed parts of curriculum; teacher - person authorised to see and browse curriculum (finished and not finished parts) assigned to him/her; and curriculum designer - person authorised to see and browse curriculum assigned to him/her and to design his/her parts of curriculum (curriculum designers and guarantors of individual courses).

Content management structures its options (Curriculum development modules) in the top-down principle, i.e. showing the global parts of the curriculum as first (study program) and going down the particular elements of curriculum items (e.g. details of learning outcomes in individual learning units). The following views for the modification of all building blocks of the curriculum are available: study programmes - the list of study programs available at particular education institution; medical disciplines - medical disciplines taught in given study program(s); courses - courses (compulsory, elective) taught and offered to the students of particular medical disciplines and study programs; learning units - components (lectures, practical lessons, seminars etc.) of individual courses; learning outcomes - competencies of the students after completing particular learning unit, course or whole study programme; and study materials - information about relevant study literature (e.g. books, textbooks and manuals, educational websites and atlases, digital videos, presentations and animations, casuistic in images, e-learning courses (LMS) etc.) that can be linked to existing curriculum building blocks (learning units and learning outcomes).

The curriculum browsing module itself is available in the main menu options too. The results of user's searching are returned according to the relevance of a given search phrase from the most to the least appropriate. The search itself uses several attributes of the learning unit, as well as other curriculum building blocks to which it is linked. Each relevant result consists of: the title of the learning unit (which can be opened and the user can see its content); information about the field of study, course to which the learning unit is added (the user can use its hyperlink to see the full list of learning units linked to this course), the form of teaching and its range and the keywords of the learning unit.

In addition to the initially identified requirements of curriculum designers, given by the needs analysis, the EDUportfolio allows users to browse content not only within the curriculum, but also within the recommended study materials. This module was added to the system with the aim to extend its browsing capabilities, to allow learners identification of relevant study literature directly in the curriculum description and/or to be redirected to eLearning courses or other multimedia materials via embedded hyperlinks. Users can search within books and multimedia descriptions and the results are also returned to them according to relevance of a

given search phrase. Furthermore, there is a set of interactive filters for choosing specific programme, course or semester too that can be used to refine results the user is interested in. Like the Curriculum browser, this module is also available to all logged-in users at the home page of EDUportfolio.

Curriculum reporting is a standalone module that allows users to examine aggregated statistics about teaching. The link to Curriculum reporting module with individual summary reports is located in the main menu options, similar to other EDUportfolio modules. The summary report modules were developed to allow users to check: overview of learning units' elaboration, overview of learning outcomes' elaboration, overview of learning outcomes' assessment forms, overview of teaching range, numbers of outcomes' and learning units' links to courses and number of learning outcomes' links to learning units. Each module of all above mentioned reports consists of two interactive parts: a filter panel and a visualisation window. The filter panel allows users to filter input data (where relevant) based on the chosen study programme, section, medical disciplines, courses, semester and/or teaching type and to decide whether to show or hide categories with no values. Individual parts of filters are shown or hidden, depending on the type of selected report. In all filters, it is also possible to reset all previously set preferences using the "reset filters" button. Visualisation via bar chart represents a graphical overview of selected parts of curriculum (learning units, learning outcomes, assessment forms, teaching range, links between curriculum elements) and their level of elaboration. The visualisation itself is a vertical or horizontal multi-bar chart that is built using the NVD3 JavaScript library for interactive web visualisations.

Module for publications is a freely accessible module for overview of all our relevant activities and publications, which relate to the domain of curriculum development, management and mapping.

IV. Conclusion

The challenge of our research was to perform innovations of curricula in teaching domains formalised with the use of an unambiguous parametric description, and entities adopted from the outcome-based concept. In general, such innovations enhance the transparency and continuity of the environment in which the teachers, guarantors, curriculum designers and faculty management as well as students, work on a daily basis. Here, the fundamental curriculum building blocks were introduced from the global perspective in a form of entity-relationship data model.

The whole process of curricula management was improved, fully digitalized and simplified. The benefits of our platform will be also transformed to the students via optimal distribution of courses, lessons and taught topics. Respecting student centered approach, the optimised curricula integrated not only recent professional skills, but as we suppose, also the modern teaching methods based on new modern technologies. Additional added value of our approach was obtained through interinstitutional cooperation of international teams and therefore the students of all partner institutions can benefit from higher quality of educational materials, that can be produced by international teams. Thus, the priority to support effective and efficient higher education systems using curriculum management platform can be achieved as well.

Furthermore, the open structure and the methodology itself allow to use mechanisms not only in medical study disciplines, but also in any other study branches and disciplines offered across higher education in all participating countries. Thus, this concept is interdisciplinary and can be applied without any methodological restriction also at other education institutions in partner regions.

Acknowledgment

Results presented in this paper were obtained with the support of the Erasmus+ project 2018-1-SK01-KA203-046318 "Building Curriculum Infrastructure in Medical Education" and the national agency's grant KEGA 011UPJS-4/2019 "Increasing of competences and critical thinking level in students of medical study programs using simulation tools of Problem-Based Learning and Evidence-Based Medicine".

References

[1] M. Karolyi, J. Ščavnický, V. Bulhart, P. Růžičková, and M. Komenda, "EDUportfolio: Complex platform for curriculum management and mapping", Proceedings of the 11th International Conference on Computer Supported Education, Volume 2: CSEDU, SciTePress 2019, pp. 352-358, https://dx.doi.org/10.5220/0007722103520358.

[2] M. Karolyi, J. Ščavnický, and M. Komenda, "First step towards enhancement of searching within medical curriculum in czech language using morphological analysis", Proceedings of the 10th International Conference on Computer Supported Education, SciTePress 2018, pp. 288-293, https://dx.doi.org/10.5220/0006757902880293.

[3] F. Balzer, W.E. Hautz, C. Spies, A. Bietenbeck, M. Dittmar, F. Sugiharto, et al., "Development and alignment of undergraduate medical curricula in a web-based, dynamic Learning Opportunities, Objectives and Outcome Platform (LOOOP)", Med Teach, 2016, 38(4), pp. 369–377, https://doi.org/10.3109/0142159X.2015.1035054.

[4] C. Spreckelsen, S. Finsterer, J. Cremer, H. Schenkat, "Can social semantic web techniques foster collaborative curriculum mapping in medicine? J Med Internet Res, 2013, 15(8), e169, PMID: 23948519, https://doi.org/10.2196/jmir.2623.

[5] G. Al-Eyd, F. Achike, M. Agarwal, H. Atamna, D.N. Atapattu, L. Castro, et al. "Curriculum mapping as a tool to facilitate curriculum development: A new School of Medicine experience", BMC Med Educ, 2018, 18(1):185, https://doi.org/10.1186/s12909-018-1289-9.

[6] O. Fritze, M. Lammerding-Koeppel, M. Boeker, E. Narciss, A. Wosnik, S. Zipfel, J. Griewatz, "Boosting competence-orientation in undergraduate medical education - A web-based tool linking curricular mapping and visual analytics", Med Teach 2019, 41(4), pp. 422-432, https://doi.org/10.1080/0142159X.2018.1487047.

[7] C. Steketee, "Prudentia: A medical school's solution to curriculum mapping and curriculum management", J Univ Teach Learn Pract 2015, 12(4), pp. 1–10, http://ro.uow.edu.au/jutlp/vol12/iss4/9.

[8] M. Komenda, D. Schwarz, C. Vaitsis, N. Zary, J. Štěrba, L. Dušek, "OPTIMED platform: Curriculum harmonisation system for medical and healthcare education", Stud Health Technol Inform 2015, 210, pp. 511–515, PMID: 25991200.

[9] M. Komenda, M. Karolyi, A. Pokorna, C. Vaitsis, "Medical and healthcare curriculum exploratory analysis", Stud Health Technol Inform 2017, 235, pp. 231–235, PMID: 28423788.

[10] A. Kononowicz, Ł. Balcerzak, A. Kocurek, A. Stalmach-Przygoda, I.A. Ciureanu, I. Hege, M. Komenda, J. Majerník, "Technical infrastructure for curriculum mapping in medical education: a narrative review", De Gruyter, 2020, https://doi.org/10.1515/bams-2020-0026.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Checking the writing of commas in Slovak

Matej Meško, Patrik Hrkút, Štefan Toth, Michal Ďuračík, Dominik Kornhauser
Department of Software Technologies
University of Žilina, Faculty of Management Science and Informatics
Žilina, Slovakia
matej.mesko@fri.uniza.sk

Abstract—**This article focuses on the design of a procedure for solving the correction of commas (incorrectly placed or missing) in the text written in the Slovak language. At the beginning of the article, the text deals with existing third-party applications that address the same issue. Subsequently, we focus on the ways in which the field of neural networks is currently applied to the problem of finding mistakes in the text. Based on the acquired knowledge, we proposed a solution using neural networks. The next part of the article deals with experiments with our neural network and different lengths of the input vector. The last part contains an evaluation of the experiments and an overall summary.**

Keywords—comma; Slovak language; comma checker

I. INTRODUCTION

Current software offers quite intelligent tools for word processing, whether for recognizing commands, transforming a spoken word into text, or finding and correcting errors in text written by humans. The problem, however, is that there are many languages. For English, text correction and overall intelligent text recognition are best designed, which is understandable given its prevalence.

In general, the correction of errors in the text is a very extensive scientific field. In our team, we focus on correcting errors in our native Slovak language. One of the first tasks we decided to solve is to correct comma writing. It is not an easy task, as Slovak is not a grammatically simple language because it contains a huge number of spelling rules and exceptions. Although it is not difficult to learn many rules in generally, there are various exceptions in grammar which in many cases allow or do not allow writing correct commas. And this is a problem for many people, especially students, to write a comma in the right place in a sentence. Alternatively, they miss a comma or put a comma where it does not belong in a sentence.

I. STATE OF THE ART

In the case of simpler text correction, the Slovak user has the option to use *dictionary-based corrections* within the available software (the database contains several forms of one word). This approach is used by various text editors (e.g., MS Word, Open Office, LibreOffice).

In the case of a more comprehensive inspection, we can mention the *korektor.sk* [1] tool. According to the authors, this tool: "Corrects the most frequent errors we make in the Slovak language – missing commas and diacritics, the form of words (such as the form of the adjective in the plural nominative), etc.". It checks grammar and spelling, and of course commas. This tool corrects words using the morphological database of the *Slovak National Corpus*, where all words are described by some marks (they express which forms of the word are spelled correctly). Other available tools for correcting Slovak text include *LanguageTool* [2] and *Corrector* [3]. They allow you to check spelling for multiple languages, including Slovak. However, these three tools do not guarantee or achieve high accuracy, especially when correcting commas, as Slovak is a language with relatively complex rules.

A survey of available articles dealing with the issue of adding punctuation to the text showed that one of the most common reasons for solving the addition of punctuation is the subsequent use of modified text for further machine processing. Such processing includes, for example: translation into other languages, sentiment analysis or information extraction. [4, 5] These tools are mostly trained on written texts, which also include punctuation. For correct results, this punctuation is needed as part of the input data. Input data are often obtained by automatically transcribing text from audio or video recordings, so they usually do not contain punctuation. Another reason may be to make it easier for a reading by a person.

In the following short paragraphs, we summarize the most common algorithmic procedures for solving the problem of error detection in the text.

A. N-grams

The easiest way to create a spell checker for a language is to create a dictionary. As we mentioned earlier, this is a very commonly used method.

Sample texts are divided into individual words. The words are grouped together, which are generally referred to as *n-grams*. The letter *n* indicates how many words are in one dictionary entry. The value of its frequency occurrence is added to this record. The whole dictionary can consist of several layers, where one layer is a dictionary with one dimension of n-gram. For example, it may contain dictionaries with *3*, *2* and *1-gram*.

When comparing, the search term in all layers of the dictionary is compared, and the higher number of n-grams and the frequency of occurrence have priority. Of course, this approach has its pros and cons.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

B. LSTM Neural Networks

The most used architecture for word processing when using neural networks is the long short-term memory (LSTM) network.

According to [5], it has several advantages over the traditionally used n-gram language model. One of them is the ability to generalize a context that was not seen by the model during training, a feature known for neural networks. Another advantage is the larger size of the context and the ability to change it dynamically for different punctuation marks.

Gale and Parthasarathy use the LSTM network in their work [4] in combination with character n-grams, which serve as input to the model and reduce the complexity of the task that the model must learn. In combination with LSTM, they also use convolutional neural networks (CNNs) to capture a broader context.

The work [4, 6] also uses bidirectional LSTM, which can move in both directions in the source text. This allows the model to use a variable-length context before or after the current position in the text. In this way, the entire input text can serve as a context for predicting punctuation and should therefore achieve better results. However, the disadvantage of this model is its complexity and the need to access the entire input text when restoring punctuation.

C. DDN – Deep Neural Networks

Another method used is *deep neural networks* (DNN). This approach is used by Che et al. in [7], where the first model consists of three fully interconnected hidden layers, other models also use convolution before entering the network itself. All data is collected from text files and in the first step they are transformed into one long sequence of words. In the mentioned work, the input for this network are five-element vectors of words, in which it is classified whether after the third word in the sequence there should be a punctuation mark and which of the signs should be given for this position.

D. Other approaches

In [8], the authors formalize comma writing as a binary classification task for the boundary between two consecutive words. They use several sequential models for this type of problem.

The Hidden-Event Language Model (HELM) is a common n-gram language model, which in the mentioned work is trained on the sequence of words with defined boundaries between words – C for a comma, N for a non-comma. The *Factored Hidden-Event Language Model* (fHELM) is a HELM extension that can handle multiple property classes. Its feature is that it can omit less frequent traits in favor of more frequent traits. The last model that the authors use in their work is *Conditional Random Fields* (CRF). In this model, the conditional probability of a sequence of text descriptions, i.e., commas and non-commas between words, is estimated using a log-linear model that can be effectively trained. CRF, according to the authors, have the advantage over other language models of higher performance compared to HELM, resp. fHELMs are not as well scalable for larger datasets.

Another way is the approach used in [4], where the input data for the model are individual characters of the text in which the punctuation is restored. Such a character vector has a length of $2n + 1$, where n represents the width of the context, i.e., the number of characters before and after the character on which the model focuses. The model then decides whether there should be a punctuation mark at the examined position.

According to Gale and Parthasarathy, this method has several advantages, including simple tokenization, as both input and output are sequences of UTF-8 characters, so the same model can easily be used for other languages. Another advantage is that it is not necessary to deal with tokens outside the dictionary, i.e., words that are not in the defined dictionary. However, the disadvantage of this approach is the need for a much broader context. While in a word-level approach, the model suffices with a few consecutive words, a larger number of characters is needed to capture a sufficient context in this case.

II. Rules of Writing Commas in Slovak Grammar

The publication *Rules of Slovak Spelling* [9] states that in the grammatical and semantic structure of sentences, the comma, as one of the punctuation marks, fulfills the following main functions (we will state them in abbreviated form, as they are complicated):

1. *Delimiter* – commas are used to exclude interjections, interpolators, salutations, contact expressions, evaluating particles, introductory sentences, free sentence elements,
2. *Connection* – using this function, a comma indicates the connection of a multiple sentence member,
3. *Separation* – in this case comma separates the marked, independent, and connected sentence member or sentence structure.

In addition to the above functions, the comma also performs other functions – it indicates the reverse order in names, terminological conjunctions, or it separates decimal values from integers (as a dot in English texts), or individual values in enumerations of symbols and numbers.

III. Model and Input Data

We decided to solve the problem as a classification task, in which a decision of placing (or non-placing) a comma in front of a space character is made. The input for this model is a vector of characters obtained from the input text, and we used a deep neural network as a classification element. The input vector is formed by a sequence of characters, which contains the *character of interest (a space)* and then the characters before and after it (context for the neural network). Since we have planned various experiments, the description of this vector could generally be written as:

$$b + 1 + a$$

where b is the length of the context before the space, a is the length of the context after the space and the number 1 represents the space character itself. This vector then

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

determined how the input text was divided into individual pieces according to the locations of spaces.

Before dividing the text into separate pieces, we removed the following characters from the input text: all characters except letters of the Slovak alphabet, numbers, spaces, and punctuation marks – dot, comma, question mark, exclamation mark, and colons. They are assumed not to have much of an effect on writing a comma in a sentence.

After clearing, the input text was divided into a set of pieces, which were even more thin down. For this part, a frequency analysis of the number of words preceded by a comma was performed. All input data was used here, and the threshold was set to 3 such words occurrences.

The reason for this filtering was that the input data can also contain errors, as we drew from publicly available sources, and we trained the network only with inputs that contained a comma.

Subsequently, each piece had to be prepared for our network, as it could contain a comma at the target location. If it contained it at the desired location, this information (the *y* flag) was stored so that neural network weights could be corrected. In the next step, all spaces were removed (the context could also contain them) and the text was then moved to their spaces (Fig. 1).

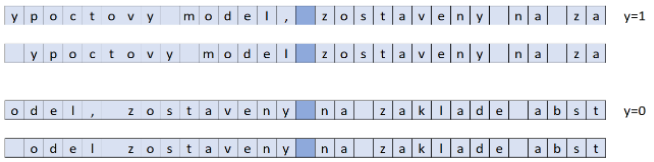


Figure 1. Preparation of the text for training and validation

If there was a lack of characters after commas removal, the missing characters were replaced by a fill – empty spaces.

A. Data sources

To achieve the highest possible accuracy of comma correction in a wide range of texts, it is necessary to have a large amount of data available for training. For the needs of our research, several freely available sources of the Slovak text were used.

We used a set of all articles of *Slovak Wikipedia*[1], *data collected from various Slovak websites* (various texts from articles, discussion forums, advertisements, etc.), the Slovak mutation of the *Digital corpus of the European Parliament*[2] and the *Golden Fund of Slovak Literature*[3].

IV. Neural Network Structure

The model itself is composed of an input layer, the size of which is given by the length of the input vector. The network also contains several hidden fully interconnected layers, the number and size of which will be the subject of our experiments. In these hidden layers, as in the input layer, a *ReLU* activation function [10] is used, with the range from zero to infinity.

The output layer of the network size is 1, as the comma should or should not be at that location. A *sigmoid* activation function is used for this layer, which is suitable for models in which the probability is predicted as an output [10].

During training, the data will be divided into a training (2/3) and a test set (1/3). The trained model will be tested on previously unused data that were not part of the original trained or validation set.

V. Experiments

The neural network was coded using the *TensorFlow*, *Keras* and *NVIDIA CUDA Deep Neural Network* libraries in *Python*. In this article we will not present the exact implementation of the network to maintain the specified size of the text of the article.

The network model is *Sequential* and consists of several layers. We changed their number during the experiments. The output always represents one neuron.

The main parameter in the experiments is the size of the input vector, more precisely the length of the context before and after the space.

A. Different context lengths

For the first experiment with the size of the vector, we used data obtained only from the Slovak version of Wikipedia. Here, the model contained six hidden layers measuring 512 neurons. We changed the size of the input dimension from 5 to 20 in steps of size 5.

In our experiments, we examined two standard statistical indicators: *accuracy* and *balanced accuracy*. From the results of the first experiment (*Table I.*), surprisingly, the best results were obtained by models with shorter contexts. The reason for this may be that only one word before or after the comma may be enough to reveal most of the correct locations of commas.

TABLE I. Experiment Result for Different Context Lengths (Indicated in %)

before / after	5	10	15	20
5 chars	97.94	97.61	97.93	97.75
10 chars	92.37	91.82	89.61	93.47
15 chars	91.38	91.57	90.88	91.83
20 chars	92.55	92.27	93.23	93.17

Accuracy

before / after	5	10	15	20
5 chars	87.20	86.54	87.23	86.65
10 chars	79.08	81.44	71.36	83.82
15 chars	79.11	79.19	78.39	79.40
20 chars	78.92	79.37	77.38	77.47

Balanced accuracy

Subsequently, we applied the network to all input data and tried to increase the size of the input vector even more. However, the balanced accuracy is significantly lower for lengths of 50 characters before and after the monitored space.

[1] https://dumps.wikimedia.org/skwiki/latest/

[2] https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html

[3] https://zlatyfond.sme.sk/autori

TABLE II. RESULTS FOR FURTHER EXPERIMENTS USING ALL DATA

Context before, after	10, 10	20, 20	30, 30	40, 40	50, 50
Accuracy	86.70	93.69	93.70	93.83	90.29
Balanced accuracy	65.24	76.62	76.98	75.90	67.83

According to *Table II*. it can be said that a larger context does not have such a significant effect on the overall accuracy of a large amount of data from different sources. Also, a context that is too short or too long can have a bad effect on the results. For context lengths of 20, 30, and 40 characters, the differences in accuracy and balanced accuracy were not significant.

B. *Amount and type of input data*

In this experiment, we tried to verify the assumption that sources of different types will give the model a broader scope and the model itself will be able to better generalize both formal and informal texts.

The model consisted of six layers, each with a size of 512 neurons. The length of the context in this case was 40 characters before and after the observed space.

For the first training we used the data from Slovak Wikipedia (skwp). Subsequently, a text from the Golden Fund of Slovak Literature (gfsl), the digital corpus of the European Parliament (dcep) and finally data from the Slovak websites (skws) were added.

From *Table III*. is obvious that although the accuracies of models with increasing amounts of data of different kinds do not differ much, the balanced accuracies are significantly different.

TABLE III. EXPERIMENTAL RESULTS FOR INCREASING AMOUNTS OF INPUT DATA

Data source	skwp	+ gfsl	+ dcep	+ skws
Accuracy	89.07	91.43	88.57	90.29
Balanced accuracy	77.36	80.70	66.89	67.83

The highest accuracy on test data was achieved by a model trained on data from Wikipedia and the Golden Fund. Because the test data is still the same and contains text of different styles, it was assumed that the model trained on the diverse data would yield better results. However, the problem may be that with the addition of other resources to the training data, noise or sentences with misspelled commas may have been added to them, which could negatively affect the results.

After running another experiment with the same data, with the same context length, but with the addition of another hidden layer of size 512, better results were obtained. The success rate of such a model on test data reached 95.18% and the balanced success rate 81.89%. We can deduct from this that with the increased amount of data, the smaller neural network was not able to learn the necessary relationships to complete the correct location of commas.

C. *Number of hidden layers and their size*

Previous experiments were performed with six layers and a layer size of 512 neurons. We therefore decided to try to change these sizes and run experiments with the number of layers 1, 6, 7 and 8 and once we tried to increase the number of neurons to 2048.

Based on the results found in *Table IV*. it can be stated that the best results were obtained with 7 and 8 hidden layers of 512 neurons.

The eight-layer model with a size of 2048 achieved worse balanced accuracy compared to the models with a smaller number of neurons. The reason for this is the high rate of false positives and false negatives.

TABLE IV. RESULTS OF AN EXPERIMENT WITH A DIFFERENT NUMBER OF LAYERS

Number of hidden layers	1	6	7	8	8
Number of neurons in one layer	512	512	512	512	2048
Accuracy	88.25	90.29	95.18	94.81	90.77
Balanced accuracy	78.31	67.83	81.89	82.80	67.72

Low accuracy was also achieved with the six layers model. In this case, the reason may be too few neurons for the model to be able to learn to generalize many comma completions cases. The model with one hidden layer brought the biggest surprise. On the test set of data, this model managed to achieve the lowest accuracy, but a relatively high balanced accuracy compared to other models. This could be caused by the high number of positives that makes the calculated balanced accuracy so high. This model therefore places too many commas, but in most cases to the incorrect locations. So, it sometimes puts the commas to the places where they should not be located.

D. *Comparison with the existing solution*

To compare the proposed solution, we used text that was obtained from various articles on the Internet and was not part of the training data for the model. Altogether, the text contained more than 64,000 words, of which 8% of the words contained a comma in front of them. Text devoid of commas entered to our best model and the online tool korektor.sk. Then their outputs were compared with the original text. The results of our model are shown in *Table V*.

TABLE V. CONFUSION MATRIX OF OUR BEST MODEL

		Actual comma	
		Positive (comma)	**Negative (not-comma)**
Predicted comma	**Positive (comma)**	True Positive (TP) 5,015	False Positive (FP) 4,954
	Negative (not-comma)	False Negative (FN) 149	True Negative (TN) 53,957

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

As you can see, the neural network with our best model found 5,015 correct commas (TP) and did not find 149 commas (FN). On the other side, the model found 4,954 commas (FP) where they were not to be placed. This value is too high, so the precision is only 50.31%.

The results of some selected statistics obtained from the experiment with our model and online tool *korektor.sk* are shown in the following Table VI.

TABLE VI. COMPARISON OF OUR BEST MODEL WITH THE ONLINE TOOL KOREKTOR.SK

	Our best model	korektor.sk
Accuracy	92.04	96.20
Balanced accuracy	94.35	71.45
Sensitivity	97.11	43.23
Specificity	91.59	99.66
Precision	50.31	89.39

The results of our model are not the best. The accuracy of our model is worse than korektor.sk, although balanced accuracy is better. Also, the sensitivity of our model is better. On the other side, the precision is very low. The model shows too many false alarms.

VI. CONCLUSION

We designed and tested proposed neural networks. The most surprising results were obtained from the experiments with the length of the context before and after the observed space, as the models with a shorter context in it achieved better results on the test data. The best of these models, with context lengths of 5 characters before and 5 characters after the comma, was also tested on a larger data set and achieved similar accuracy, 98.02% and a balanced accuracy of 86.94%.

However, a closer look with experiments shows that our model is not outperformed in all areas. Although in some statistics it was better than korektor.sk. From the results of our research, we found that the input data and their nature greatly affect the success and relevance of the proposed solution. The results of testing the proposed solution pointed to the need for implementation and the need to find other options for improving and modifying the research of the problem. In the following research, we will re-evaluate the further direction and, above all, we will focus on improving the achieved results in relation to the current solutions used and the use of new or modified algorithms.

ACKNOWLEDGMENT

This work was supported by Grant System of University of Zilina No. KOR/1124/2020.

REFERENCES

[1] korektor.sk, [Online]. Available: https://korektor.sk/ [Cit. March 2021].

[2] LanguageTool, [Online]. Available: https://languagetool.org/ [Cit. March 2021].

[3] Corrector, [Online]. Available: https://www.corrector.co/ [Cit. March 2021].

[4] W. Gale, S. Parthasarathy, „Experiments in Character-level Neural Network Models for Punctuation," rev. INTERSPEECH, Stockholm, 2017.

[5] O. Tilk, T. Alumäe, „LSTM for Punctuation Restoration in Speech Transcripts," rev. Sixteenth annual conference of the international speech communication association, Dresden, 2015.

[6] O. Tilk a T. Alumäe, „Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," rev. Interspeech, San Francisco, 2016.

[7] X. Che, C. Wang, H. Yang a C. Meinel, „Punctuation Prediction for Unsegmented Transcript Based on Word Vector," rev. Proceedings of the Tenth International Conference on Language Resources and Evaluation, Portorož, 2016.

[8] B. Favre, D. Hakkani-Tür a E. Shriberg, „Syntactically-informed Models for Comma Prediction," rev. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.

[9] Ondrejovič, S. a kol., Pravidlá slovenského pravopisu, Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied, 2000.

[10] S. Sharma, „Activation Functions in Neural Networks," 6 September 2017. [Online]. Available: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6. [Cit. 18 April 2020].

[11] R. Vohl, „Deep Learning Methods for Drum Transcription and Drum Pattern Generation", 2018.

Analysis of methods for planning data processing tasks in distributed systems for the remote access to information resources

Topic: Communication and control systems and networks

Oksana Nass
Higher School of Information Technologies
Zhangir Khan West Kazakhstan Agrarian-Technical University
Uralsk, Republic of Kazakhstan
nass55@mail.ru

Gaukhar Kamalova
Higher School of Information Technologies
Zhangir Khan West Kazakhstan Agrarian-Technical University
Uralsk, Republic of Kazakhstan
gokhakam@gmail.com

***Abstract*—The article examines the possibility of reducing the time spent on performing tasks in distributed data processing systems, which are necessary for remote access to information resources. For this, the content and volume of annotation data of resources are analyzed; the main methods of scheduling task processing in distributed systems are identified and analyzed: Least Recently Used Node/Server, Resource consumption prediction, Random Selection, Queue metrics based and Elapsed-time prediction; SJF, SPN and SRTF; Time-sharing and Least Attained Service; FIFO, LIFO, RSS. The revealed techniques are analyzed under conditions of incompleteness of information about the resource requirements of distributed systems.**

Keywords— informatization, distributed systems, big data, resource requirements, data processing techniques

I. INTRODUCTION

Currently, the process of informatization characterizes society, affecting all spheres of human activity [13].

The quantity of information in the modern world is increasing, according to the most conservative estimates, exponentially. Information systems and technologies have become part of our life. Today, more and more electronic services are received by people via the Internet, thanks to information systems for various purposes [9].

The widespread use of social networks, e-commerce, ubiquitous remote access to information resources and other information services are characterized by a significant increase for data, as well as an increase in computing loads. Following the period of creation and use of local and corporate information systems, the period of creation and use of distributed data processing systems (DDPS) begins.

Distributed data processing systems can be distributed spatially and / or functionally. They can solve a common problem only together, combining their local capabilities and agreeing on the adopted particular solutions. Usually they are a collection of application software, database, system-wide software, implemented on the basis of a computing system, with the aim of solving some application for data processing or control [10].

The difficulty of creating and using DDPS lies in the fact that the automatic transfer of well-proven solutions in the field of local informatization to distributed systems often does not give the desired results [1, 12].

In this regard, the problem of improving methods for scheduling tasks and processing big data in DDPS is urgent.

II. ANALYSIS OF THE CONTENT AND VOLUME OF DATA FOR PROCESSING IN DDPS

As an example, consider the process of planning applied tasks of data processing in DDPS, necessary for remote access to information resources.

The types of information resources, depending on the access mode, are determined by GOST 7.82-2001. This GOST is allocated: resources of local and remote access (network resources) [4].

We concretize the division of resources according to this criterion.

Access is understood as access to an electronic document located on a device other than the user's device. Local access is access to an electronic resource located on a physical medium. Remote access is access to electronic resources located on a server accessible through information and telecommunication networks [6].

Remote access to information resources can be paid, limited, permanent, temporary, etc.

We use annotation as a means of providing remote access to information resources. Name, author of the resource; languages in which information is presented; the date and place of creation, the organization responsible for the resource, as

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

well as other information will identify the information resource.

Annotation of Internet resources is a process of analytical and synthetic information processing, the purpose of which is to obtain a generalized characteristic of an Internet resource in terms of content, language, form and novelty of the information contained in it, as well as service capabilities provided to a remote user. The structure of the annotation for an Internet resource includes four blocks: identification, thematic (content), service and chronological [14].

The input of the distributed data processing system in real time receives a stream of annotations for processing.

Annotations contain four blocks. They can differ in their intended purpose (reference, advisory) and have a different amount of data (extended or short annotations).

The nodes of the distributed processing of the cluster are engaged in processing the received tasks, working with large amounts of data.

Cluster is a group of computers connected by high-speed communication channels working together to run common applications, representing from the user's point of view a single hardware resource [5].

The task scheduling subsystem is responsible for mapping tasks to the available computational resources of the DDPS. Based on the choice of a task for execution with large amounts of data, we will choose a procedure for sorting tasks in accordance with their priorities on each specific DDPS node.

For greater clarity, these processes will be presented schematically in Figure 1.

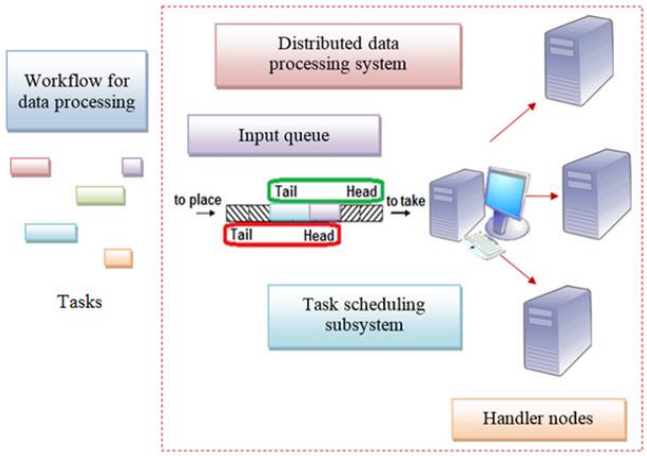


Figure 1. Process of planning data processing tasks in DDPS.

In this way, the tasks being processed operate on Big Data. For their planning, real-time DDPS is required, not limited by the number of tasks to be processed; applied tasks may have significantly different requirements for the computational resources of the DDPS; in addition, it is not known in advance what resources DDPS will have at each moment of time.

III. Scheduling Techniques to Reduce Task Execution Time in DDPS

Let us examine which planning techniques will help reduce the time spent on completing tasks in real-time DDPS in conditions of incomplete information about resource requirements.

By the task scheduling policy we mean a set of rules that are used to determine when and how to select a new task for processing, under what conditions the processing of the task flow is stopped, what to do with partial processed data.

The discipline of service means the order of service of applications [2].

The main task of optimizing the operation of any queuing system is queuing management, that is, to serve the maximum number of requests entering the queuing system per unit of time and thereby qualitatively service each requirement and get the maximum possible profit for the enterprise. The process of work of queuing systems is a random process with a discrete state and continuous time, since the demands in the system arrive in a random, unpredictable way [2].

1) The method of scheduling tasks, when resource requirements are known in advance and all tasks are processed in a universal way, involves hierarchical scheduling of applied tasks [11, 15]:

- The task with the lowest resource requirements is selected first, based on the Shortest Job Next (SJN) service discipline, also known as Shortest Job First (SJF) or Shortest Process Next (SPN);
- The task for which the least service time is selected first, based on the service discipline shortest remaining processing time, also known as Shortest Remaining Time First (SRTF).

So, the method of scheduling tasks, when resource requirements are known in advance, takes into account the heterogeneity of tasks and has low planning overhead, but requires all information about a multitasking system already at the development stage, which may not be feasible for complex systems.

2) The method of scheduling tasks with support for preservation of condition allows you to interrupt the execution of the current task and switch to the execution of another task, which have higher priority [3, 7].

This technique is an integral part of any operating system and is used to schedule interrupt processes. It allows several tasks at different stages of execution on each computational node.

An interrupt is a temporary halt in the execution of one program in order to quickly execute another, at the moment more important. The interrupt controller serves interrupt procedures, receives an interrupt request from external devices, determines the priority level of this request and issues an interrupt signal to the microprocessor [7].

This technique is based on the following disciplines:

- Time-sharing, which implies multitasking and multiprogramming of tasks for the sharing of computing resources by several tasks;
- Least Attained Service, which is based on mathematical laws of the theory of probability, for example, when a task is selected for processing that supposedly requires the least service time.

So, this method of scheduling tasks involves the implementation of saving the state of execution of tasks, but it requires additional costs, in addition, interrupts can be unacceptable when processing big data and computing time-consuming tasks.

3) Method for assigning resources of DDPS nodes to tasks without support preservation of condition [8]:

- Least Recently Used Node / Server presupposes select the least loaded or unused node;
- Resource consumption prediction, allowing you to select the node that best matches the resource consumption forecast;
- Random Selection assumes that the handler node is randomly selected;
- Queue metrics based allows you to select the node with the shortest queue length or average processing latency;
- Elapsed-time prediction allows you to select the node with the shortest predicted execution time.

This technique requires additional resources to predict the choice from the queue of tasks for processing in accordance with priority or other criteria.

4) Method of scheduling tasks when resource requirements are unknown [16]:

- The tasks are processed in the order of their arrival, based on the FIFO (First in First Out) service discipline;
- The tasks are processed in reverse order, based on the LIFO (Last in First Out) service discipline;
- The task is selected at random, based on the RSS (Random Selection for Service) discipline;
- The task is selected at Time Sharing – part of the time is allocated to each task in turn;
- The task that will receive the Least Attained Service is selected.

The disadvantages of this method include the following: selection of a DDPS node that meets resource requirements is not carried out; the same type of handler nodes are used (the amount of resources on each node is the same); it is assumed that tasks are of the same type in terms of resource requirements (computational and spatial labor intensity).

For greater clarity, we present the identified techniques in Figure 2.

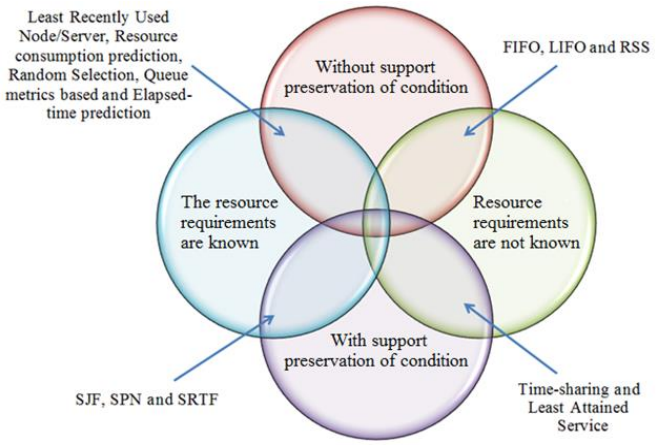


Figure 2. Basic methods of scheduling tasks in DDPS.

In this way, for planning applied data processing tasks in real-time DDPS, using annotation as a means of providing remote access to information resources, it is proposed to use the method of assigning resources of DDPS nodes to tasks without maintaining state in conditions when resource requirements for DDPS are unknown. The service disciplines selected for this method (FIFO, LIFO, and RSS) are implemented at the level of middleware between application and system software.

IV. CONCLUSION

The analysis of the possible content and amount of processed data was carried out in conditions when annotations are used as a means of providing remote access to information resources.

Big data stream consists of equal priority annotation processing tasks. The processed annotations can differ in their intended purpose (reference, recommendation) and have a different volume (extended or short annotations), however, they are characterized by the same resource requirements for DDPS, since they have the same structure, which includes four blocks (identification, thematic, service and chronological).

Scheduling techniques have been identified that can reduce the time spent on completing tasks in real-time DDPS:

- Least Recently Used Node/Server, Resource consumption prediction, Random Selection, Queue metrics based and Elapsed-time prediction methods do not take into account the heterogeneity of tasks;
- SJF, SPN and SRTF methods take into account the heterogeneity of tasks, however they are designed for tasks with known resource requirements;
- Time-sharing and Least Attained Service methods implement the preservation of the task condition, which allows processing several tasks at different stages of execution on one node, however, the

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

resources of one computer do not allow processing time-consuming tasks;

- FIFO, LIFO, RSS service disciplines are used in conditions of incomplete information about resource requirements, however, they are applicable only to equal priority tasks with the same type of resource requirements and assume the same type of processing nodes.

Thus, to plan data processing tasks in DDPS, using annotation as a means of providing remote access to information resources, it is proposed to use service disciplines: FIFO, LIFO, RSS, which are implemented at the middleware level between application and system software.

ACKNOWLEDGMENT

This publication is the result of the project implementation ERASMUS+ ACeSYRI: Advanced Center for PhD Students and Young Researchers in Informatics reg.no. 610166-EPP-1-2019-1-SK-EPPKA2-CBHE-JP.

REFERENCES

[1] Distributed data processing. Construction of distributed models in the SimInTech system: guidelines / compilers S.P. Khabarov, M.L. Shilkina; executive editor A.M. Zayats. – St. Petersburg: SPbGLTU, 2018. – 124 p. (in Russian).

[2] Dobrokhotov, Yu. N. Fundamentals of the theory of queuing: a training manual / Yu. N. Dobrokhotov. – Cheboksary: ChGSKhA, 2018. – 82 p. (in Russian).

[3] olubev I.A. Scheduling tasks in distributed computing systems based on metadata: Dissertation of the candidate of technical sciences: 05.13.11 / Federal State Budgetary Educational Institution of Higher Professional Education “St. Petersburg State Electrotechnical University named after V.I. Ulyanov (Lenin) ". – St. Petersburg, 2014. – 135 p. (in Russian).

[4] GOST 7.82-2001 System of standards for information, librarianship and publishing. Bibliographic record. Bibliographic Description of Electronic Resources. [Electronic resource]. URL: http://online.zakon.kz (in Russian).

[5] Khabarov S. P. Computing machines, systems and networks / S.P. Khabarov, M. L. Shilkina. – St. Petersburg: SPbGLTU, 2017. – 240 p. – ISBN 978-5-9239-0888-6 (in Russian).

[6] Kolkova, N.I. Information support of automated library and information systems: textbook / N.I. Kolkova, I.L. Skipor. – Kemerovo: KemGIK, 2018. – 356 p. – ISBN 978-5-8154-0419-9 (in Russian).

[7] Kuznetsov, E.N. Elementary base and functional units of information measuring and control systems: a tutorial / E.N. Kuznetsov. – Penza: PSU, 2019. – 348 p. – ISBN 978-5-907102-89-7 (in Russian).

[8] Methods and models for the study of complex systems and processing of big data: monograph / I.Yu. Paramonov, V.A. Smagin, N.Ye. Kosykh, A.D. Khomonenko; edited by V.A. Smagin and A.D. Khomonenko. – St. Petersburg: Lan, 2020. – 236 p. – ISBN 978-5-8114-4006-1 (in Russian).

[9] Models and research methods of information systems: monograph / A.D. Khomonenko, A.G. Basyrov, V.P. Bubnov and others ; edited by A.D. Khomonenko. – St. Petersburg: Lan, 2019. – 204 p. – ISBN 978-5-8114-3675-0 (in Russian).

[10] Nabieva G.S. Block-symmetric models and methods for designing data processing systems: PhD dissertation: 6D070200 / Kazakh National Technical University named after K.I. Satpayev. – Almaty, 2010. – 115 p. (in Russian).

[11] Osmolovskiy S.V., Ivanova E.R., Fedorov I.R. Hierarchical scheduling of tasks in real-time systems on multi-core processors // Bulletin of the Samara Scientific Center of the Russian Academy of Sciences, volume 18. – 2016. – No. 1 (2). – P. 405-411 (in Russian).

[12] Prikhodko M.A. Principles of modeling processes in conditions of incompleteness of initial information: Synopsis of the dissertation of doctor of technical sciences: 05.13.01 – "System analysis, control and information processing" (industry). – Moscow, 2011. – 32 p. (in Russian).

[13] Robert, I.V. Theory and technique of education informatization. Psychology-pedagogical and technological aspects / I.V. Robert. – Moscow: Knowledge laboratory, 2014. – 400 p. (in Russian).

[14] Ryabtseva, L.N. Analytical and synthetic processing of information: Annotation and abstracting: a tutorial / L.N. Ryabtseva. – Kemerovo: KemGIK, 2019. – 103 p. – ISBN 978-5-8154-0480-9 (in Russian).

[15] Standard of the Republic of Kazakhstan ISO 15888-2000 Space data and information transfer systems. Standard formatted data units. Referencing environment, IDT. [Electronic resource]. URL: http://online.zakon.kz (in Russian).

[16] Zhumataev N.S. Principles of modeling processes in conditions of incompleteness of initial information: PhD dissertation: 6D075100 / South Kazakhstan State University named after M. Auezov. – Shymkent, 2011. – 108 p. (in Russian).

Tripping of F-type RCDs for High-Frequency Residual Currents

Hanan Tariq and Stanislaw Czapp
Faculty of Electrical and Control Engineering
Gdańsk University of Technology
Gdańsk, Poland
hanan.tariq@pg.edu.pl; stanislaw.czapp@pg.edu.pl

***Abstract*—Residual current devices (RCDs) are apparatus commonly used for protection against electric shock in low-voltage electrical installations. They protect people in the case of an earth fault or even in the case of direct contact with the live parts. However, to be effective protective devices, RCDs have to detect residual currents of various waveform shapes which appear in modern electrical installations. For this purpose, RCDs are classified into four types: AC; A; F and B. This paper is focused on F-type RCDs provided for the detection, in particular, of mixed-frequency residual currents. According to the standard referring to the F-type RCDs, they are tested by manufacturers under the non-sinusoidal waveform having components generated by control equipment supplied from a single-phase. In this paper, results of two tripping tests (other than normative) of F-type RCDs are presented. During the first test, waveforms having components generated by control equipment supplied from three phases were forced. During the second test, high-frequency pure sinusoidal residual currents were generated. Results of these tests have shown that F-type RCDs may detect mixed-frequency residual currents other than the normative but may not react to sinusoidal currents of frequencies higher than 1 kHz.**

Keywords—earth current; high frequency; RCDs; tripping test

I. INTRODUCTION

In low-voltage electrical installations, residual current devices (RCDs) have been used on a larger scale for over 60 years [1]. Earliest types of RCDs were able to detect only sinusoidal residual currents of a network frequency. Further development of rectifiers was an impulse to construct RCDs detecting pulsating direct residual currents [2] or even a smooth direct current [3]. Currently, in modern electrical installations, the new challenge for RCDs appears, i.e., mixed-frequency residual currents, having various spectra. Such currents may be a cause of improper tripping of RCDs. The mixed-frequency residual currents are generated not only in industry – they flow in domestic installations as well. In various household appliances such as a washing machine or food processors, variable-speed drives are used. An earth fault in a circuit with a variable-speed drive may result in both residual pulsating direct current and especially the aforementioned mixed-frequency residual currents [4, 5].

The international standards [6, 7] classify RCDs into the following types, regarding their ability to detect various shapes of residual waveforms:

- AC-type: for sinusoidal (50/60 Hz) residual currents only,
- A-type: for sinusoidal (50/60 Hz) residual currents and pulsating direct residual currents,
- F-type: for sinusoidal (50/60 Hz) residual currents, pulsating direct residual currents, and mixed-frequency residual currents generated by control equipment supplied from a single-phase,
- B-type: for sinusoidal residual currents having frequency up to 1000 Hz, pulsating direct residual currents, smooth direct residual currents, and mixed-frequency residual currents.

Regarding F-type RCDs, their ability to the detection of mixed-frequency residual current is verified under the normative waveform [7] composed of the following three components (I_{10}, I_{50} and I_{1000}):

1) $I_{10} = 0.035I_{\Delta n}$: the component having frequency equal to 10 Hz; it reflects the output frequency (relatively low frequency) of the converter which controls the rotor speed of the motor; $I_{\Delta n}$ corresponds to the rated residual operating current of the RCD (e.g., 30 mA) at the rated network frequency,
2) $I_{50} = 0.138I_{\Delta n}$: the component having frequency equal to the network rated frequency (here assumed 50 Hz),
3) $I_{1000} = 0.138I_{\Delta n}$: the component having frequency equal to 1000 Hz; it reflects the switching frequency of the aforementioned converter (inverter).

The above-described mixed-frequency normative waveform is visualized in Fig. 1. The permissible range of the tripping current for such a waveform is $(0.5–1.4)I_{\Delta n}$.

Analysis of the provisions of the standard [7] dedicated for F-type RCDs shows that the mixed-frequency normative waveform reflects components occurring in a specific case. It is considered that a converter is supplied from a single-phase only. However, the market offer related to F-type RCDs provides both two-pole and four-pole devices. Therefore, it seems reasonable to verify the behavior of these devices in other conditions of supply, especially in the case of the three-phase supply.

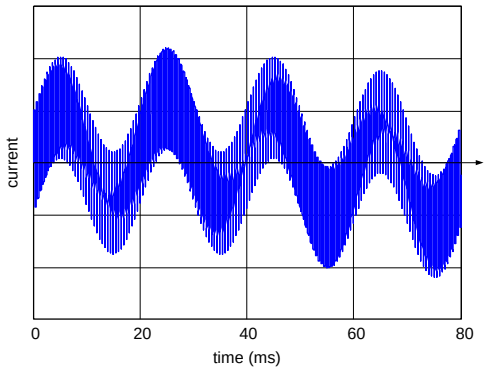


Figure 1. The normative mixed-frequency waveform for testing of the F-type RCDs, reflecting the earth fault current in a converter circuit supplied from a single-phase. A waveform composed of: 10 Hz, 50 Hz and 1000 Hz.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

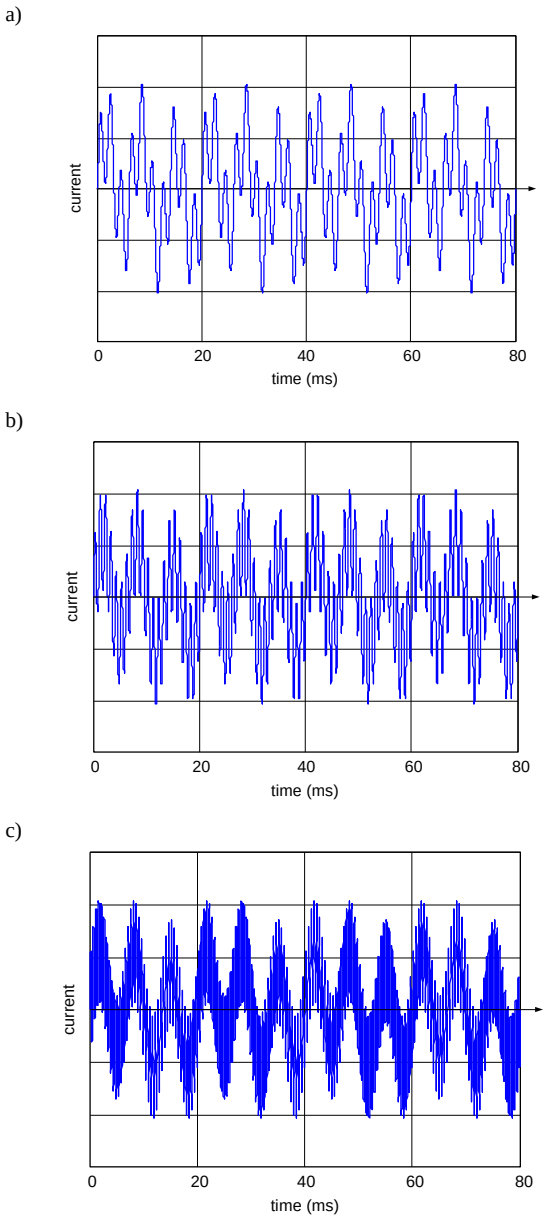


Figure 2. Mixed-frequency waveforms reflecting the earth fault current in a converter circuit supplied from three phases. A waveform composed of: a) 50 Hz, 150 Hz and 500 Hz; b) 50 Hz, 150 Hz and 1000 Hz; c) 50 Hz, 150 Hz and 2000 Hz.

In practice, there are many converter circuits supplied from three phases. In three-phase circuits with frequency converters, the following three components dominate [8]:

1) I_{output}: the component of the output frequency of the converter to obtain desired rotor speed of the motor; it can have values ≤50 Hz (for the 50 Hz, the nominal speed is achieved),
2) I_{150}: the component that refers to the voltage-to-earth of the rectifier neutral point (specific in the case of the supply from three phases),
3) $I_{switching}$: the component that refers to the switching (PWM – pulse width modulation) frequency of the converter.

As it is seen, these components are different than in a single-phase circuit. Moreover, the switching (PWM) frequency may have a value significantly higher than normative 1000 Hz for the F-type RCDs according to [7].

Fig. 2 presents the results of simulations of the earth fault current composed of components that may occur in a converter circuit supplied from three phases. In the simulative study, the following content of these components was assumed:

1) $I_{50} = 0.035I_{\Delta n}$ (the output frequency: 50 Hz),
2) $I_{150} = 0.138I_{\Delta n}$ (the 150 Hz component that refers to the voltage-to-earth of the rectifier neutral point – a three-phase supply),
3) $I_{switching} = 0.138I_{\Delta n}$ (the switching frequency: 500 Hz (Fig. 2a), 1000 Hz (Fig. 2b) and 2000 Hz (Fig. 2c)).

Taking into account the published results of the frequency behavior of RCDs [4, 9–12] as well as normative requirements referring to F-type RCDs [7], the following extended tests of F-type RCDs have been performed:

1) Verification of the response of RCDs to mixed-frequency residual currents which may occur in a converter circuit supplied from three phases.
2) Verification of the response of RCDs to the pure sinusoidal residual current of frequency up to 50 kHz; the very high frequency in this test reflects the effect of the possible switching frequency of the converter.

II. Description of the Laboratory Stand

Fig. 3 is presenting a typical laboratory platform utilized for the testing of RCDs. The equipment used for the purpose can be described as:

- AC source (230 V, 50 Hz) which is supplying power to the system,
- a generator responsible for creating a mixed-frequency signal,
- an ammeter (true RMS) for the measurement of residual current,
- a variable resistor to have a flexible and desired residual current value in the circuit,
- an RCD, meant to be verified.

The laboratory generator is coupled with a PC computer having a dedicated software. This software enables to set a pure sinusoidal waveform or parameters of the particular components (amplitude and phase angle) of a mixed-frequency waveform. After programming the mixed-frequency waveform, the laboratory generator maintains the proportion of its components constant for every value of the total residual current.

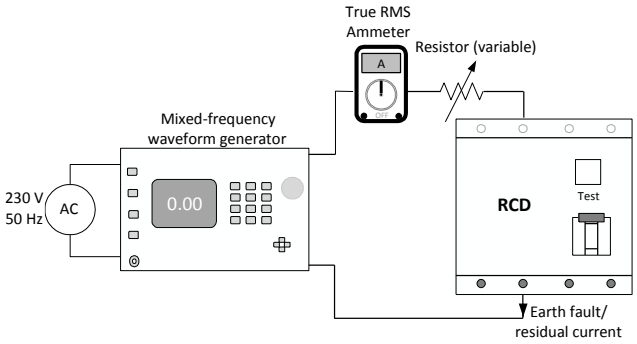


Figure 3. Testing platform for RCDs.

III. The Scope of the Tests and Their Results

A. General Information

RCDs that are meant to be tested have the rated residual operating current value of 30 mA ($I_{\Delta n}$) and the subjective RCDs were abruptly exposed to the preset values of residual current i.e., $I_{\Delta n}$; $2I_{\Delta n}$; $5I_{\Delta n}$; $10I_{\Delta n}$; $15I_{\Delta n}$ and $20I_{\Delta n}$. These residual current values can be further elaborated as 30 mA, 60 mA, 150 mA, 300 mA, 450 mA and 600 mA respectively. Testing of the response of RCDs to the suddenly applied residual current value reflects a real accident when a person touches an enclosure of the electric equipment in the case of its insulation fault. After touching this enclosure, a body current of a relatively high value may suddenly flow. Two RCDs (F-type) were chosen for the purpose and Table I presents the symbols for selected RCDs and manufacturers.

TABLE I. Symbols of Tested RCDs

Serial No. of RCD	Manufacturer symbol	RCD symbol
1	Manufacturer_X	RCD_X
2	Manufacturer_Y	RCD_Y

Both RCDs' (RCD_X and RCD_Y) behavior has been verified by subjecting them to two different types of testing mechanisms:

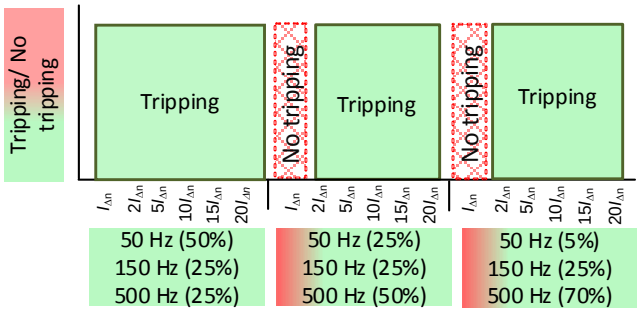
1) The tripping test for mixed-fequency waveforms; each waveform was composed of three components (I_{50}, I_{150}, $I_{switching}$); the content of the I_{50} and $I_{switching}$ was variable.
2) The tripping test for pure sinusoidal waveforms of high frequency.

B. The Test for Mixed-Frequency Waveforms (Three Components/Frequencies)

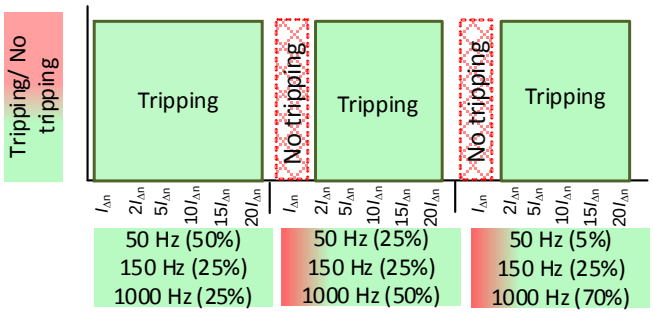
Primarily, the delivered testing waveform was comprised of three different frequencies i.e., fundamental frequency (50 Hz) and two high-frequency figures. For the high-frequency part, the first component (150 Hz) is typical for a three-phase supply. The second high-frequency component was assumed to be: 500 Hz; 1000 Hz and 2000 Hz consecutively. Results of the aforementioned testing type have been shown in Fig. 4 and Fig. 5. In each graph, the percentage content of the particular components is presented.

The testing results of the F-type RCD (RCD_X) in Fig. 4 depict the fact that despite being the modernised type (regarding high frequency), the performance of F-type RCDs can only be considered as relatively satisfying. It is worth mentioning that two out of three frequency components are same (50 Hz and 150 Hz) for all stages (Fig. 4a)b)c)) whereas third and the highest frequency component is variable i.e., 500 Hz, 1000 Hz and 2000 Hz for Fig. 4a, Fig. 4b and Fig. 4c, respectively. The RCD (RCD_X) showed its positive performance for the case when the share of high-frequency component was low i.e., 50 Hz (50%), 150 Hz (25%) and 500/1000/2000 Hz (25%).

a)



b)



c)

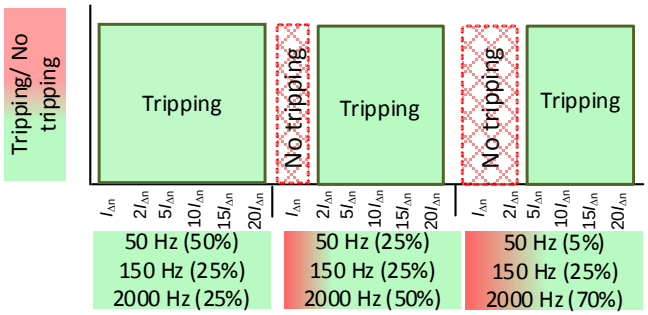


Figure 4. Tripping results of 30 mA F-type RCD (RCD_X) in the presence of the following high-frequency components mixed with fundamental frequency (50 Hz): a) 150 Hz and 500 Hz, b) 150 Hz and 1000 Hz, c) 150 Hz and 2000 Hz.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

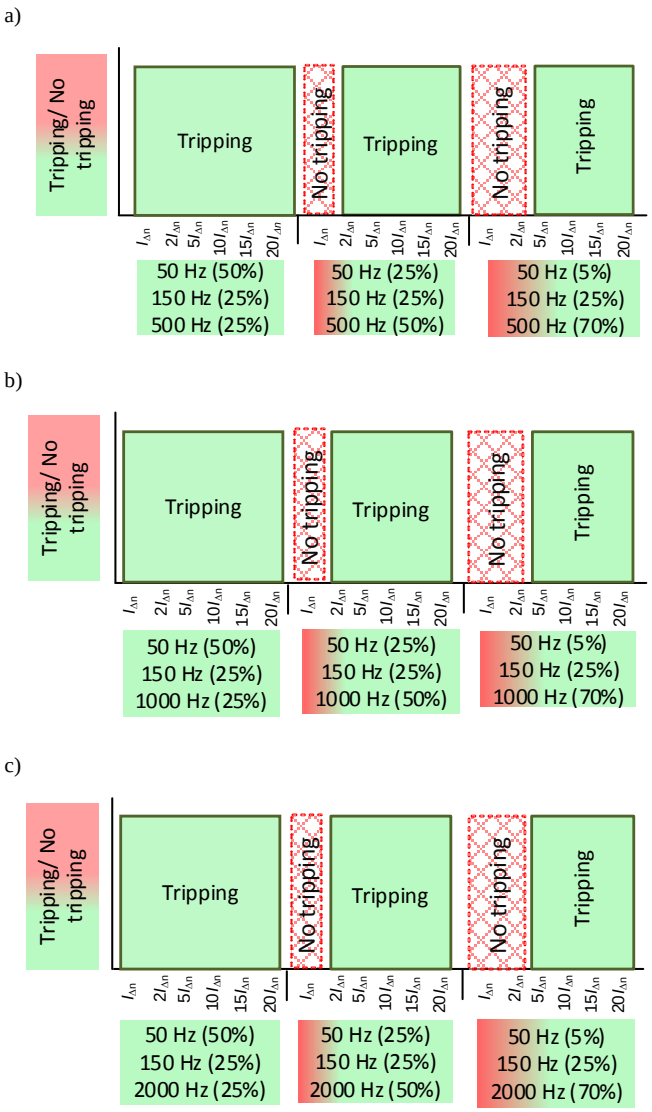


Figure 5. Tripping results of 30 mA F-type RCD (RCD_Y) in the presence of the following high-frequency components mixed with fundamental frequency (50 Hz): a) 150 Hz and 500 Hz, b) 150 Hz and 1000 Hz, c) 150 Hz and 2000 Hz.

Afterwards, the increased share of the high-frequency component i.e., 50 Hz (25%); 150 Hz (25%) and 500/1000/2000 Hz (50%) caused the RCD (RCD_X) to depict no reaction at the residual current value of $I_{\Delta n}$ (30 mA) but the same RCD (RCD_X) showed tripping when exposed to the residual current value of $2I_{\Delta n}$ (60 mA) and sustained its operation up to $20I_{\Delta n}$ (600 mA). However, for the final case, 50 Hz (5%); 150 Hz (25%); and 500/1000/2000 Hz (70%), RCD_X remained untripped for $I_{\Delta n}$ and additionally for $2I_{\Delta n}$ in the case of 2000 Hz (70%), which was the worst behavior comparative to previously mentioned ratios.

Behavior of another F-type RCD (RCD_Y) subject to the aforementioned testing phenomenon has been verified (Fig. 5a)b)c)). Similar to the previously mentioned case (Fig. 4), RCD_Y has also not shown adequate outcomes. In the case of the less share of high-frequency component i.e., 50 Hz (50%); 150 Hz (25%) and 500/1000/2000 Hz (25%), the RCD (RCD_Y) behaved normally and displayed no tripping irregularity. When this RCD (RCD_Y) was exposed to a greater share of high-frequency component i.e., 50 Hz (25%), 150 Hz (25%) and 500/1000/2000 Hz (50%), it didn't trip for 30 mA ($I_{\Delta n}$) of supplied residual current but behaved well on the rest of the residual current levels ($2I_{\Delta n}$; $5I_{\Delta n}$; $10I_{\Delta n}$; $15I_{\Delta n}$ and $20I_{\Delta n}$).

Ultimately, the case where high-frequency component had the greatest share i.e., 50 Hz (5%), 150 Hz (25%) and 500/1000/2000 Hz (70%), RCD_Y remained untripped for $I_{\Delta n}$ and $2I_{\Delta n}$ as well. It can be labelled as the worst response among both tested RCDs (RCD_X and RCD_Y).

C. *The Test for Pure Sinusoidal Waveforms of High Frequency*

The latter part of the research was dedicated to the tripping verification of the F-type RCDs (RCD_X and RCD_Y), tested while exposing a sinusoidal (harmonic-free) frequency signal of 50 Hz; 500 Hz; 1000 Hz; 2000 Hz; 5000 Hz; 10,000 Hz; 20,000 Hz and 50,000 Hz. The residual current was abruptly applied with the levels of $I_{\Delta n}$ (30 mA); $2I_{\Delta n}$ (60 mA); $5I_{\Delta n}$ (150 mA); $10I_{\Delta n}$ (300 mA); $15I_{\Delta n}$ (450 mA) and $20I_{\Delta n}$ (600 mA). Testing results are shown in Fig. 6 and Fig. 7. Both RCDs (RCD_X and RCD_Y) have shown quite similar behavior in this part of testing (Fig. 6 and Fig. 7). One thing that needs to be highlighted that both RCDs (RCD_X and RCD_Y) tripping results were quite unfavorable for very high frequencies. It can be clearly seen that RCDs' (RCD_X and RCD_Y) behavior was somehow better on lower frequencies (Fig. 6a and Fig. 7a) as compared to the higher ones (Fig. 6b and Fig. 7b). However, while comparing the results of lower frequencies (Fig. 6a vs. Fig. 7a), in both cases the results were quite pleasing for only 50 Hz frequency.

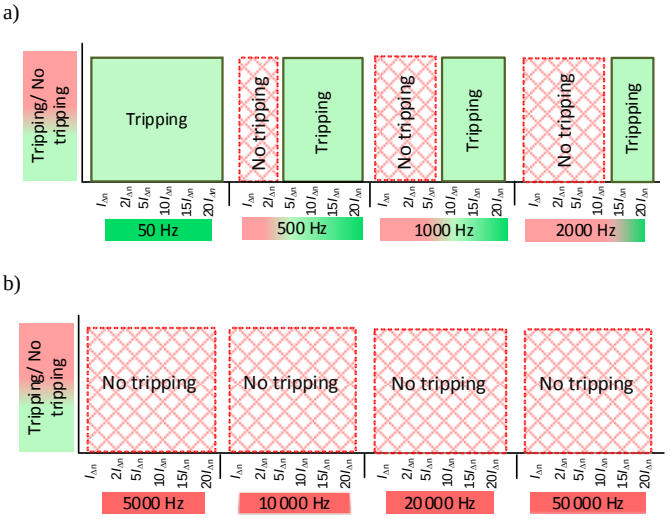


Figure 6. Tripping results of 30 mA F-type RCD (RCD_X) for sinusoidal waveform having frequency: a) 50 Hz, 500 Hz, 1000 Hz and 2000 Hz, b) 5000 Hz, 10,000 Hz, 20,000 Hz and 50,000 Hz.

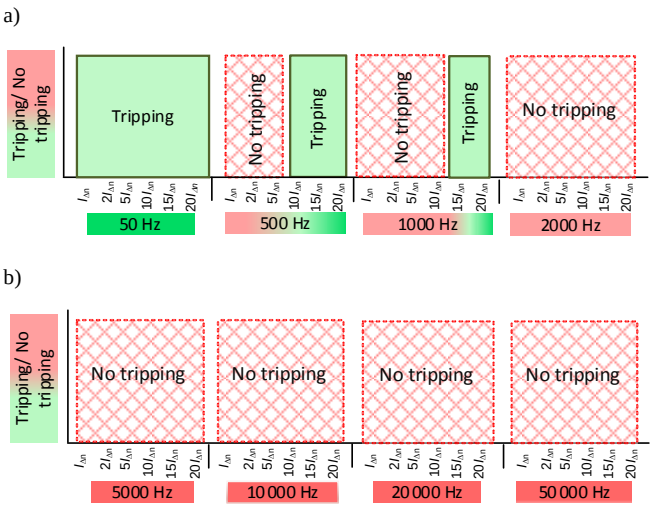


Figure 7. Tripping results of 30 mA F-type RCD (RCD_Y) for sinusoidal waveform having frequency: a) 50 Hz, 500 Hz, 1000 Hz and 2000 Hz, b) 5000 Hz, 10,000 Hz, 20,000 Hz and 50,000 Hz.

As soon as the testing frequency reached 500 Hz (Fig. 6a and Fig. 7a), both RCDs (RCD_X and RCD_Y) failed to trip to a certain residual current limit ($2I_{\Delta n}$ and $5I_{\Delta n}$ respectively). Similar unfavourable performance was observed for 1000 Hz and 2000 Hz, where RCDs (RCD_X and RCD_Y) functioning was a question mark for the protection obligation. For higher frequencies within the range 5 kHz – 50 kHz (Fig. 6b and Fig. 7b), the tested RCDs didn't detect even the highest residual current value of $20I_{\Delta n}$ (600 mA).

Thus, the response of the F-type RCDs to very high frequency currents (higher than 1 kHz) is doubtful.

IV. Conclusions

Results of the tests presented in this paper have shown that F-type RCDs may detect mixed-frequency residual currents composed of components other than normative. However, these RCDs are not suitable for the detection of residual currents having a very high frequency. The first one among the tested RCDs did not react to the sinusoidal waveforms of frequencies 5 kHz, 10 kHz, 20 kHz and 50 kHz, even if the testing current was 20 times higher than the rated residual operating current of the RCD. In the case of second RCD among the tested ones, no reaction was observed for the sinusoidal waveforms of frequencies 2 kHz and higher. Since the switching frequency of converters in modern electrical installations has tendency to be increased, the present requirements of the standard referring to the tests of the F-type RCDs seem to be insufficient.

References

[1] Residual Current Devices. Application Guide, Eaton, Vienna, Austria, 2017.

[2] H. Rösch, "Current-operated ELCBs for AC and pulsating DC fault currents," Siemens Power Eng., no. 8–9, pp. 252–255, 1981.

[3] R. Solleder, "Allstromsensitive Fehlerstrom-Schutzeinrichtung fur Industrieanwendung," ETZ, vol. 115, pp. 896–901, 1994.

[4] S. Czapp and H. Tariq, "Behavior of residual current devices at frequencies up to 50 kHz," Energies, vol. 14, no. 6, 2021, https://doi.org/10.3390/en14061785.

[5] S. Czapp and J. Guzinski, "Electric shock hazard in circuits with variable-speed drives," Bulletin of the Polish Academy of Sciences: Technical Sciences, vol. 66, no. 3, pp. 361–372, 2018, DOI: 10.24425/123443.

[6] Residual current operated circuit-breakers without integral overcurrent protection for household and similar uses (RCCBs) – Part 1: General rules, IEC 61008-1, 2010.

[7] Type F and type B residual current operated circuit-breakers with and without integral overcurrent protection for household and similar uses, IEC 62423, 2009.

[8] S. Czapp, "The impact of higher-order harmonics on tripping of residual current devices," Proc. Int. Power Electronics and Motion Control Conference EPE-PEMC 2008, Poznan, Poland, 1–3 Sept. 2008, DOI: 10.1109/EPEPEMC.2008.4635569.

[9] Z. Erdei, M. Horgos, C. Lung, A. Pop-Vadean, and R. Muresan, "Frequency behavior of the residual current devices," IOP Conf. Ser. Mater. Sci. Eng., vol. 163, no. 012053, 2017.

[10] Y. Shopov, S. Filipova-Petrakieva, and B. Boychev, "Investigation of residual current devices in high frequencies," In Proceedings of the 10th Electrical Engineering Faculty Conference (BulEF), Sozopol, Bulgaria, 11–14 Sept. 2018.

[11] T.M. Lee and T.W. Chan, "The effects of harmonics on the operational characteristics of residual current circuit breakers," In Proceedings of the International Conference on Energy Management and Power Delivery, Singapore, 21–23 Nov. 1995, pp. 715–719.

[12] S. Czapp and J. Horiszny, "Simulation of residual current devices operation under high frequency residual current," Przeglad Elektrotechniczny, no. 2, pp. 242–247, 2012.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Computational study of red blood cell behaviour in shear flow for different bending stiffness of the membrane

Mariana Ondrušová[1,*], Ivan Cimrák[1,2]

[1] Cell-in-fluid Biomedical Modelling and Computations Group, Faculty of Management Science and Informatics, University of Zilina, Slovakia

[2] Research Centre, University of Zilina, Slovakia
[*] Corresponding author: mariana.ondrusova@fri.uniza.sk

***Abstract* — While a red blood cell moves in the vessel, it often changes its shape, adapting it to many obstacles or other objects. This behaviour is influenced by elastic properties of the cell. One need to take into account these elastic properties when modelling of red blood cells. The elastic properties are implemented in the models by means of elastic moduli. One of these moduli is the bending modulus. This modulus is important for maintaining the shape of the cell but especially it is ensuring a gradual return to its original state in case of deformation. The bending modulus is based on keeping the angles between each pair of adjacent triangles in the red blood cell's mesh. In this article we focus on rotation of the cell in shear flow. By appropriately selected simulations, we want to investigate the rotation frequency of the cell and the influence of bending modulus on the rotation frequency.**

Keywords— computational modelling; red blood cell; shear flow; simulations

I. Introduction and Motivation

One of our long-time goals is simulation of microfluidic devices, in order to examine the flow of fluid in devices of very small sizes, or to explore behaviour of various biological cells. The flow of blood can be modelled as a flow of continuous fluid with cells as elastic objects immersed in that fluid. These objects are red blood cells, white blood cells and blood platelets that are together in the blood plasma. The first step of the research of the behavior of red blood cells is mainly biological experiments. Experiments are performed in vitro. One of biological experiments is, for example, stretching of the red blood cell [1]. This experiment helps us to calibrate some of the elastic parameters of the red blood cell.

Our computer model has been introduced in [2] and is based on fluid interaction with elastic objects that represent red blood cells. The model of the fluid is based on the lattice-Boltzmann method. Cell deformations is governed by elastic properties defined by elastic moduli. Validation of the model was published in [3-5]. Our current values that we use are the results of this calibration. In this article we focus on the bending modulus, which affects the bending ability of the cell and thus changes its behaviour in the flow. All simulations are carried out under an open-source software ESPResSo [6].

The general behaviour of simulations is influenced by numerical properties of cells, shape of the microfluidic device or character of the fluid flow. In this case we work with shear flow. Shear flow occurs when the fluid velocity is different at the top and bottom of the cell or if the fluid at the top of the cell has the same velocity but the opposite direction as on the bottom side. From the beginning, the cell begins to move under the influence of fluid – this movement is called tumbling. At higher velocities the cell rotates, and the centre of the cell appears static and only membrane rotates – this type of motion is called tank-treading. With tumbling and tank-treading we have already done simulations, for example, in [7,8]. However, we do not know how our model behaves during the transition between these two states. And that is another goal we have set.

In this article, we are focused on the influence of the elastic modulus on the behaviour of our computational model in shear flow. We described simulations where we create one cell that rotates within shear flow. In Section II we will describe the basics of our computational model with a more precise focus on the bending modulus. In Section III we provide basic simulation settings. Section IV we will describe the results, and, in the end, we will summarise the results into conclusions.

II. Model of Red Blood Cell

In this section we will shortly describe our simulation model. During the simulation, mutual interactions occur at each time step. Those mutual interactions are:
1. between the object and the fluid,
2. between the fluid and the channel walls,
3. between the objects and the walls of the channel

Also, for simulations with multiple cells, there are interactions:
4. between objects.

Fluid in the channel is described by the Lattice-Boltzmann method. The fluid space consists of points in the cubic grid. The grid remains fixed throughout the simulation time. The fluid is

made up of fictitious points that collide with one another and thus transmit information about themselves - about the velocity and direction of movement. More information is in [9].

Models of blood cells consist of a flexible triangular network, which have a shape of the cell's surface. We then insert elastic moduli into the network. Elastic moduli represent elastic properties. Elastic moduli allow cell to deform but also to return into its original shape.

Simulation of the membrane in our model is ensured by six moduli. Five of them are representing the elastic behavior: stretching, bending, local and global area conservation, volume conservation. The sixth one is representing cell's membrane viscosity. Analysis of individual moduli can be found for example in [10-12].

For our further work we will focus on the bending modulus. This modulus is important for preserving the shape of the cell, but also ensures a gradual return to its original shape when it is deformed. The bending modulus is based on keeping the angles between each pair of adjacent triangles (e.g. triangles ABC and BCD) and is thus computed for each edge. The force to maintain the angles is calculated as follows:

$$F_b\,(A,B,C) = k_b \frac{\Delta\theta}{\theta^0}\, n_{ABC}, \qquad (1)$$

where k_b is bending modulus, θ is angle between two triangles, which have a common edge BC, θ^0 is their relaxed angle, $\frac{\Delta\theta}{\theta^0} = \frac{\theta-\theta^0}{\theta^0}$ is the relative offset of the angle between two adjacent triangles with the common edge BC and n_{ABC} is the unit normal vector of the triangle ABC. If we suppose that BC is the common edge, the force applied to A has the opposite direction and double magnitude to forces applied to B and C.

Interaction between the fluid and the cells is governed by the coupling represented by the friction coefficient, which at any time step penalizes locally the differences in the velocity of the object and the velocity of the fluid.

III. Simulation Setup

If we create a setting for the cell, in which we will always change only one parameter, we can easily compare the change in cell behavior. The basis for our cell validation research was based on [13], where we deal with tumbling, tank-treading and also part of the cell membrane bending. For the simulation, we used a 20 x 20 x 20μm channel. When modeling the shear flow in the fluid, the cell started to rotate under the influence of the opposite directions of the fluid flow at opposite boundaries of the simulation box. The fluid is started from the upper and lower channel boundaries, in the opposite direction, which means that if at the upper velocity is "v", on the bottom will be "-v". For our channels, the velocity in y-direction is zero. Distribution of the velocity in the in x-direction is linear.

The current value for the bending modulus is $3x10^{-3}$ N.m. To evaluate the effect of the bending modulus on cell behavior, we have tested the following values: $3x10^{-4}$, $3x10^{-3}$, $1.2x10^{-3}$, and $4.8x10^{-3}$ Nm. After that, based on [3], we have selected six values of fluid velocities from $0.0002ms^{-1}$ to $0.002ms^{-1}$. Other important values for simulation are shown in Table 1.

TABLE I. Simulation Setup

timestep	10^{-7} s
stretching coefficient	$8x10^{-6}$ N/m
local area conservation coefficient	$6x10^{-6}$ N/m
global area conservation coefficient	$9x10^{-4}$ N/m
volume conservation coefficient	$5x10^{2}$ N/m^2
viscosity coefficient	$1.5x10^{-6}$ Ns/m
radius of RBC	$4x10^{-6}$ m
friction coefficient	$3.39x10^{-9}$ Ns/m
fluid density	10^{3} kg/m^3
fluid viscosity	$1.5x10^{-6}$ m^2/s

IV. Results

As already mentioned, the cell under the influence of the shear flow starts to rotate. We assumed that rotation would be uniform for a given velocity. However, visualizations have shown that the cell rotational is influenced by its initial inclination in regard to the direction of the fluid flow – Fig.1. The first picture represents the cell in its initial position, at the beginning of the simulation. Later during the rotation, the cell gradually accelerates up to the moment, when it returns to its original position parallel to the flow of fluid. Here, the cell slows down its rotation significantly. This behaviour can be observed during the whole simulation, also when multiple rotations occur.

For a better understanding we present cell's rotation in the Fig.2. The x-position of one fixed point around the cell is in the range 6 to 14 μm. We already know that the radius of the cell is 4μm, the diameter is 8μm, which corresponds to the observed range of numbers.

We can find description of the cell's behavior during tumbling and tank-treading in several publications. However, we were concerned about our cell behavior during the "transition" between these two states. We have seen a several of simulations to find the right rate for changing cell behavior. One example is Fig. 3 with a cell that still has a part of the tumbling but at the same time, the cell membrane is already turning around the main axis, the base for tank-treading. The axes show the rotation of the cells in the simulation channel.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

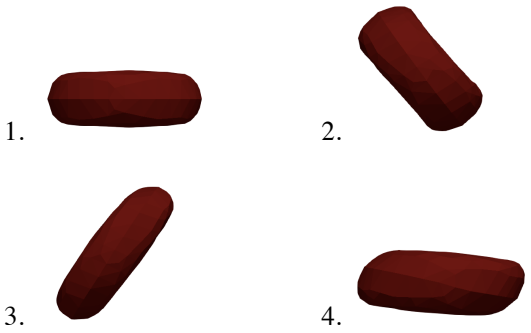


Figure 1. Rotation of cell in shear flow. Numbers 1-4 show cell positions in time.

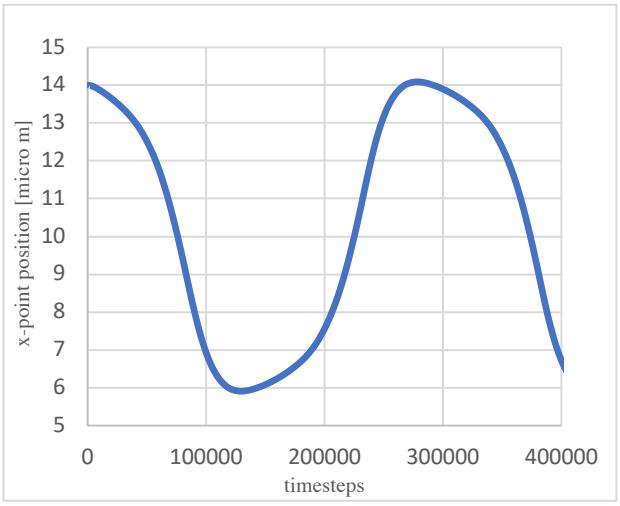


Figure 2. Representation of rotation of cell during the simulation.

We can find description of the cell's behavior during tumbling and tank-treading in several publications. However, we were concerned about our cell behavior during the "transition" between these two states. We have seen a several of simulations to find the right rate for changing cell behavior. One example is Fig. 3 with a cell that still has a part of the tumbling but at the same time, the cell membrane is already turning around the main axis, the base for tank-treading. The axes show the rotation of the cells in the simulation channel.

As we have already written in Section II., the bending modulus affects the overall stiffness of the cell. For our simulations, we've selected values of the bending modulus in this range, in order to help the cell to preserve its original shape. Smaller values for bending modulus didn't bring us enough visible differences, at higher values the cell became so rigid, that didn't correspond to the model which represents the red blood cell. Therefore, the minimum value was set to $3x10^{-4}$ and maximum to $4.8x10^{-3}$ Nm.

The following four simulations represent influence on the rotation of the point around the cell. Observed differences can be seen on selected figures – Fig. 4.-7. We can see the difference between the highest and the lowest value. In Fig. 4, there are only two values compared to other figures - lowest and highest. Therefore, because at low velocities the differences are minimal and not clearly visible.

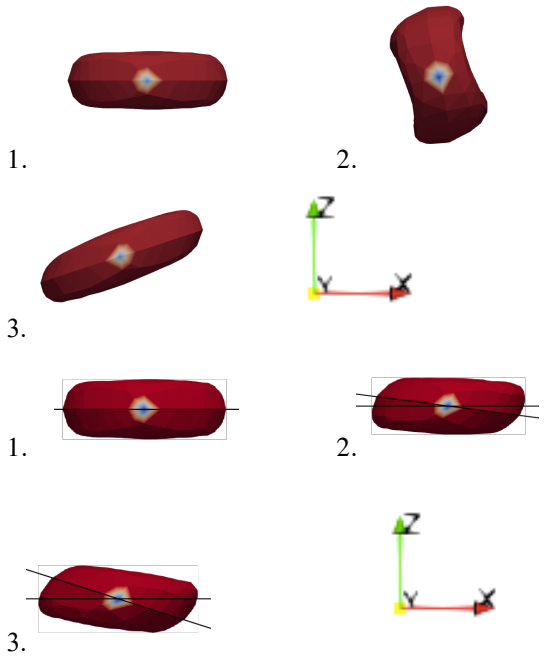


Figure 3. Behavior of the cell during the „transition". The first three pictures indicate tumbling, second three tank-treading.

Figures also show the dependence between cell stiffness and point rotation - the lower is the cell stiffness, the more slowly is the cell rotating in the flow. It is logical to assume, that the higher stiffness of the cell has a negative effect on the rate of rotation in the flow. However, the higher stiffness of the cell causes, that the cell is less affected by the shape change. What it means is that the forces that return the cell to its original shape are lower. By less force, the cell is less slowed down and hence its rotation is higher.

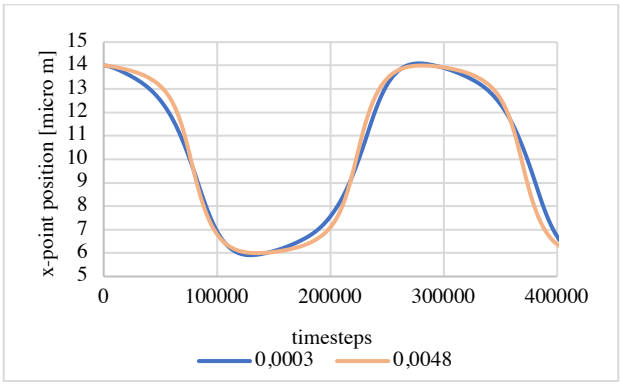


Figure 4. Rotation of the cells is representing by position of one tracked point on the cells surface. We can see that the angular velocity is influenced by bending modulus. Color lines represent different values of k_b. Velocity of fluid in this experiment was $0.0006ms^{-1}$.

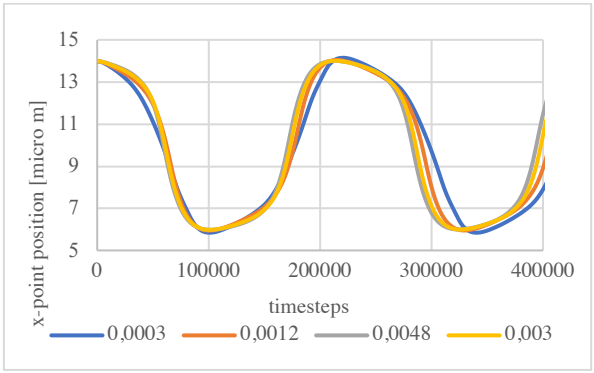


Figure 5. Color lines represent different values of k_b. Velocity of fluid in this experiment was $0.0008ms^{-1}$.

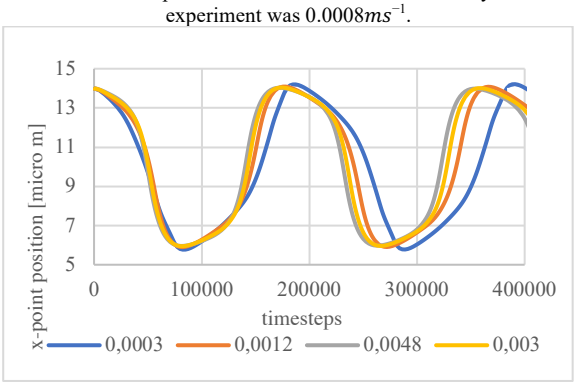


Figure 6. Color lines represent different values of k_b. Velocity of fluid in this experiment was $0.001ms^{-1}$.

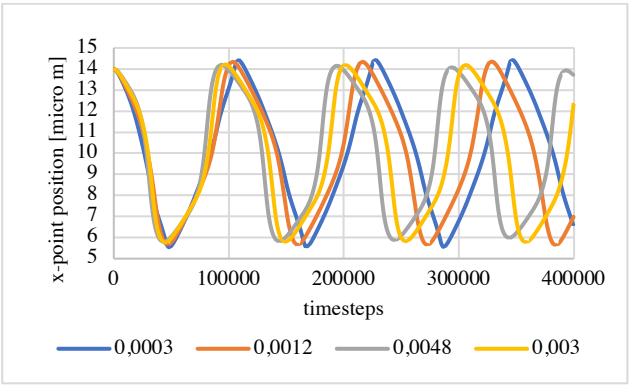


Figure 7. Color lines represent different values of k_b. Velocity of fluid in this experiment was $0.002\ ms^{-1}$.

V. Conclusions

The first results point out to inequality of cell rotation velocity in shear flow. If the cell is located along the direction of the flow of fluid, the cell rotates significantly more slowly than at the transitions to this state. It is caused by the fact that fluid is run by a motion of the channel boundaries. The closer the cell is to the boundaries, the more important is the impact of the velocity of fluid to the cell.

In the previous section, we were interested in the behavior of our model in transitions between tumbling and tank-treading. The cell still maintains rotation around the axis, but we can observe also a slight rotation of the membrane around the axis of rotation.

Another goal was to show the effect of the bending module on the overall stiffness of the cell. The essence was not to find the best value for bending modulus, but to show the effect of the modulus on the rate of rotation of the cell in the shear flow. On four figures we saw, the higher is the bending modulus, the period of cell point rotation is reduced. Even after these findings, we cannot determine which module value is the best. For as we wrote in the Section I., the values of the modules must be defined in accord with others elastic moduli, in order to meet the results from biological experiments.

Acknowledgment

This work was supported by Operational Program "Integrated Infrastructure" of the project "Integrated strategy in the development of personalized medicine of selected malignant tumor diseases and its impact on life quality", ITMS code: 313011V446, co-financed by resources of European Regional Development Fund.

References

[1] M. Dao, C.T.Lim, S. Suresh: Mechanics of the human red bood cell deformed by optical tweezers. *J. Mech. Phys. Solids* 51(11), 2259-2280 (2003)

[2] I. Cimrák, M. Gusenbauer, T. Schrefl: Modelling and simulation of processes in microfluidic devices for biomedical applications. *Comput. Math. Appl.* 64(3), 278-288 (2012)

[3] M. Ondrušová: Dynamical properties of red blood cell model in shear flow. *IDT*, s. 288-292 (2017)

[4] M. Bušík: Development and optimization model for flow cells in the fluid. [Dissertation thesis].114 p. (2017)

[5] K. Bachratá, H. Bachratý: On modeling blood flow in microfluidic devices. *10th International Conference on ELEKTRO*, pp. 518–521 (2014)

[6] ESPResSo (Extensible Simulation Package for Research on Soft matter. [Online]. Available: http://espressomd.org. [Cit. 08.09.2018]

[7] T. Krüger, M. Gross, D. Raabe, F. Varnik: Crossover from tumbling to tank-treading-like motion in dense simulated suspensions," *Soft Matter,* pp. 9008-90015 (2013)

[8] T. M. Fisher: Tank-tread frequency of the red cell membrane: Dependence on the viscosity of the suspending medium, *Biophys. J.* 93, 2553–2561 (2007)

[9] B. Dunweg, A. J. C. Ladd: Lattice-Boltzmann simulations of soft matter systems. Ad*vanc. in Polymer Science* 221, 89–166 (2009)

[10] R. Tóthová, I. Jančigová, M. Bušík: Calibration of elastic coefficients for spring-network model of red blood cell. *IDT*, pp. 376-380, (2015)

[11] M. Ondrušová: Sensitivity of red blood cell dynamics in a shear flow. *Cen. Europ. Res. J.,*pp. 28-33 (2017)

[12] K. Kovalčíková, A. Bohinikova, M. Slavík, I. Mazza Guimaraes, I. Cimrák: Red Blood Cell Model Validation in Dynamic Regime. *Lec. Not. in Comp. Science*, 10813, pp. 259-269 (2018)

[13] D. A. Fedosov, B. Caswell, G.E. Karniadais: Cell deformation in shear flow. *Comput. Hydrodyn. of Caps. and biolog. Call,* pp. 204-209 (2010)

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Numerical Experiment Characteristics Dependence on Red Blood Cell Parameters*

Kristina Kovalcikova
Research centre
University of Zilina
Zilina, Slovakia
kristina.kovalcikova@uniza.sk

Hynek Bachraty, Katarina Bachrata, Katarina Buzakova
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia

***Abstract*—The aim of this paper is to confirm sensitivity of simulation experiments on a used model of the red blood cell (RBC). Different triangulations have been used in the experiment, which has been designed in accordance with the real medical experiment. These triangulations consisted of 374 and 141 nodes. We compared several parameters of RBCs for the simulations with different types of RBCs using statistics. These parameters were characteristics of cells' velocity and deformation. The acquired results show a certain correspondence of these characteristics. Therefore, we can substitute more complicated models of RBCs by a simpler models in some types of simulations. This substitution reduce the computational time of simulations. Thus simulations can run faster or can contain a more important amount of the RBCs, or channel can have a bigger size.**

Keywords—Red blood cell, KS test, deformation, velocity

I. INTRODUCTION

Our research group is dealing with numerical models of blood cells immersed in blood or a similar fluid. The numerical tool to make models is described in [1], [4] or [5]. Our aim is to explore the behavior of blood cells, and to design microfluidic devices with special purposes. In order to spare time and financial resources, such a microfluidic device should be explored firstly by numerical manners, in order to understand its features and improve its functionality. After that it can be constructed and tested in laboratory conditions.

One of the limitations of this approach is complexity of the microfluidic devices. Some of the microchips need to have considerable dimensions and to contain complicated structures. In some cases, they can be simplified, so the numerical model is simple enough and it can be run in a reasonable time. However, in some cases it is not possible to make such a simplifications, and we need to handle with big models. Thus, if the simplification of the channel is not possible, we can simplify the numerical model of the RBC. This can be done only if the simplified model is reliable enough, and if the simulation results with the simplified model are comparable to the results with undepreciated model.

The choice of the simulation channel used for this study, which was inspired by an existing laboratory experiment described in [7], is made with intention of latter comparison of simulation results with results of the original experiment.

This work was supported by Operational Program "Integrated Infrastructure" of the project "Integrated strategy in the development of personalized medicine of selected malignant tumor diseases and its impact on life quality", ITMS code: 313011V446, co-financed by resources of European Regional Development Fund.

II. DESIGN OF THE SIMULATION

The simulation is inspired by a laboratory experiment explained into details in [7]. The aim of the experiment was observation of RBC velocity-deformation correlation in a microchip with narrow slits formed by obstacles. The liquid used in the experiment was blood diluted by standard saline with the blood-saline ratio of 1:50, so the final hematocrit is about 1%.

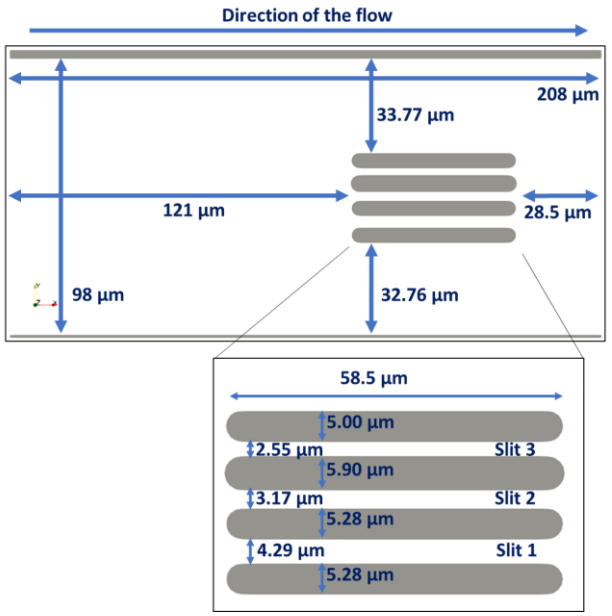


Figure 1. Schema of the simulation channel with its dimensions

The simulation channel with its dimensions is presented in the Figure 1. The internal dimensions of the channel are 208 µm × 98 µm × 3.5 µm. In the simulation, the in-flow and out-flow extremity of the simulation channel are constructed as a periodic boundaries. That means, that any cell which is leaving the simulation channel, is entering the same channel from the other side. To avoid the influence of these periodic boundary conditions, the simulation channel was constructed longer in x-axis. This way, the simulation channel is composed basically from 2 parts: The first one, which is without obstacles, is used to define random initial position of the cells.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The second one, with obstacles and narrow slits, is completely free of cells at the beginning of the simulation (Figure 2), and it is observed during the simulation to explore the behavior of the RBCs. The periodic data are treated and used only for one of the characteristics described below. Apart from that we considered only the first flow of RBCs through the channel, in order to clear the dataset from periodic effects.

The amount of RBCs in the simulation is fixed to be 38, in order to fit the hematocrit from the laboratory experiment.

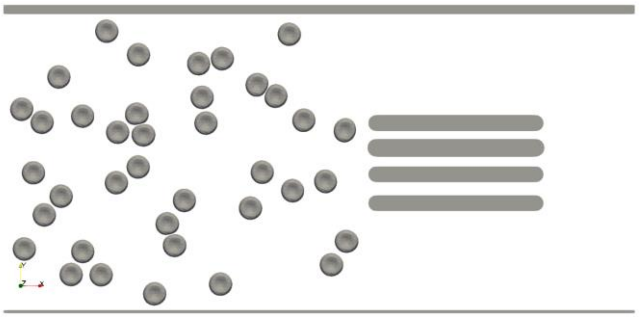


Figure 2. Initial position of the cells in the channel, example of a random seeding

III. TECHNICAL DESCRIPTION OF SIMULATION PARAMETERS

There are two types of simulation, which were run. The simulation box was the same for the both simulations, but the model of cell was different. The principal difference between the models are their finesse – first model was an approximation of the cells surface by 374 points and the second one was made from 141 points. Both models are depicted in Figure 3.

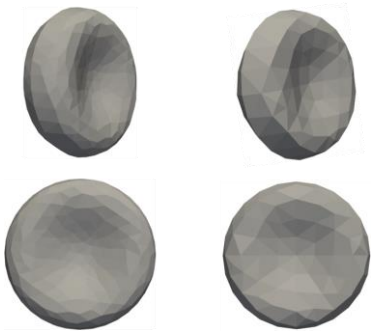


Figure 3. The RBC models used in simulation. From left to right model with 374 nodes and 141 nodes

The elastic coefficients had to be set separately for each model. In theory, those elastic parameters should not be dependent on the triangulation of the cell. But, as far as the triangles on the cells surface are not perfectly equilateral for any of the two models, the elastic coefficients have to be adjusted for each model.

The calibration of those elastic coefficients was made by stretching experiment. The detailed explanation of the calibrating process is explained in [6]. The obtained elastic parameters for our models are summarized in TABLE I. Another type of parameters which had to be set separately for each cell model were the interaction parameters between cells, which define the behavior of the cells during their collision. Those parameters are summarized in TABLE II. The numerical parameters of simulation liquid are depicted in TABLE III. The TABLE IV. summarize the nomenclature convention of realized simulations. Number at the beginning of the name signify the model used for RBC, the letter in the end signify the initial seeding of the cells.

IV. OUTPUTS FROM THE SIMULATION

For the comparison of the simulation results for different RBC models, we used these following characteristics: Cells' velocity in different part of the simulation box and deformation of the cells. For this purpose, the records from simulations include for each cell the following information:

- Actual timestep,
- Vector of position of the center of the cell,
- Dimensions of circumscribed box.

The simulation channel was created with periodic boundary conditions in in-flow and out-flow boundary. During the simulation, some of the cells run twice (or more times) through the simulation box. In order to avoid the effect of periodic boundaries, the second run (and all other runs) was cut from the output. The comparison of the simulation characteristics was done only with data, which were recorded while the cells run first time through the observed part of the channel. The only characteristics which use the data from repetitive passages through the channel were the ones which were dealing with deformation and velocity of the cells passing through the narrow slits. In general, only few cells passed through these slits, so we used all of the available data from those passages to make the comparison.

V. COMPARISON OF CHARACTERISTICS FOR DIFFERENT SIMULATIONS

A. General comparison of cell velocities

The speed of the individual RBCs is an important and sensitive characteristics of their behavior in channel respectively simulation experiment. As noted above, we have an immediate RBC center position in each recorded simulation step. These values allow us to monitor the velocity and the trajectory of each RBC in detail. However, the complete data description of these movements is too extensive and it is not easy to perform simple yet solid comparison.

Therefore, to get first notion about cells behavior, we compared the velocities of several cells in simulations with equal seedings. We focused on x-component of the velocity, as the liquid was flowing principally in this direction.

Figure 4 presents a comparison of four cell couples (two cells within one couple starting from the same position) from simulations 374a and 141a. We can see for all of the four cases, that the cells velocity in x-direction is slightly smaller in simulations with 141node cells. The difference of the velocity for these cells ranges from 1 to 15%. More detailed explanation of this fact will be described below.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE I. ELASTIC PARAMETERS OF THE CELL MODELS USED IN OUR EXPERIMENTS, ESTABLISHED BY SIMULATION OF STRETCHING EXPERIMENT

	374 model		141 model	
	LB units	SI units	LB units	SI units
Radius	$3.91Lm$	$3.91*10^{-6}m$	$3.91Lm$	$3.91*10^{-6}m$
Stretching coefficient k_s	$6*10^{-3}LN/Lm$	$6*10^{-6}N/m$	$5*10^{-3}LN/Lm$	$5*10^{-6}N/m$
Bending coefficient k_b	$8*10^{-3}LNLm$	$8*10^{-18}Nm$	$8*10^{-3}LNLm$	$8*10^{-18}Nm$
Coefficient of local area conservation k_{al}	$1*10^{-3}LN/Lm$	$1*10^{-6}N/m$	$1*10^{-3}LN/Lm$	$1*10^{-6}N/m$
Coefficient of global area conservation k_{ag}	$0.9LN/Lm$	$9*10^{-4}N/m$	$0.9LN/Lm$	$9*10^{-4}N/m$
Coefficient of volume conservation k_v	$0.5LN/Lm^2$	$5*10^2N/m^2$	$0.5LN/Lm^2$	$5*10^2N/m^2$
Membrane viscosity	$0Lm^2/Ls$	$0m^2/s$	$0Lm^2/Ls$	$0m^2/s$

TABLE II. PARAMETERS OF INTER-CELLULAR INTERACTIONS FOR DIFFERENT CELL MODELS

	374 model		141 model	
	LB units	SI units	LB units	SI units
a	$2*10^{-3}(-)$	$2*10^{-3}(-)$	$2*10^{-3}(-)$	$2*10^{-3}(-)$
n	$1.5Lm$	$1.5*10^{-6}m$	$1.5Lm$	$1.5*10^{-6}m$
cutoff	$0.4(-)$	$0.4(-)$	$0.7(-)$	$0.7(-)$
offset	$0(-)$	$0(-)$	$0(-)$	$0(-)$

TABLE III. THE NUMERICAL PARAMETERS OF SIMULATION LIQUID

Parameter	LB units	SI units
Density	$1Lkg/Lm^3$	$1*10^3kg/m^3$
Kinematic viscosity	$1Lm^2/Ls$	$1*10^{-6}m^2/s$
Friction coeff. 374 cell	$1.15(-)$	$1.15(-)$
Friction coeff. 141 cell	$3.06(-)$	$3.06(-)$

TABLE IV. THE NOMENCLATURE CONVENTION OF REALISED SIMULATIONS. NUMBER AT THE BEGINNING OF THE NAME SIGNIFY THE MODEL USED FOR RBC, THE LETTER IN THE END SIGNIFY THE INITIAL SEEDING OF THE CELLS

Simulations with	374-nodes cells	141-nodes cells
Series with equivalent seeding of cells	374a	141a
	374b	141b
	374c	141c
	374d	141d
Series with unique seeding of cells		141e
		141f
		141g
		141h

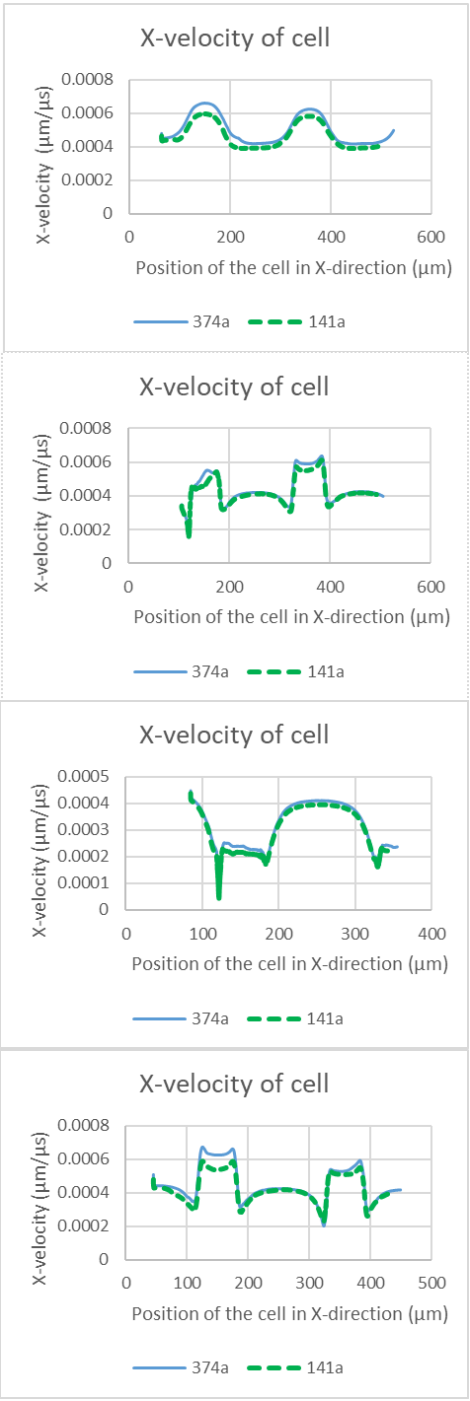


Figure 4. Comparison of x-velocity of four representative couples of cells in simulations with 141-node cells (141a) and in simulations with 374-node cells (374a).

B. Statistical processing of cells velocities

At this point, we focus on further statistical comparison of RBC velocities in channel. In study [2], we designed and verified statistical methods that allow easy and methodologically clear processing and comparison of these data. We have also used this methodology in this case. For statistical processing and comparing of the velocities of the entire RBC population, we have shown that it is sufficient to record only the minimum, maximum, and average (average from recorded velocities) value of each RBC velocity. Each of the three datasets of the minimum, maximum, and average speeds of all RBCs can be considered as simplified but significant characteristics of their movement in the channel. Confirmation or rejection of the hypothesis of a consistent behavior of RBC velocities in two simulations can then be replaced by a standard statistical test of hypothesis that the two obtained datasets come from the same, nonparametric, distribution. For this we can use, for example, a relatively rigorous Kolmogorov-Smirnov test (KS-test), or a more "tolerant" compliance tests dealing with mean value of the empirical distribution.

For the practical use of this method, we performed the following preprocessing of the obtained data. First of all, we merged speed data for experiments 141a to 141d in one group 141a_d, for experiments 141e to 141h in one group 141e_h and for experiments 374a to 374d in the group 374a_d. We obtained larger data sets, each with data for 152 RBC. In group 141a_d and 374a_d, the cells have the same starting positions. Comparing the properties of these two datasets will be the core of the answer to the main issue of the impact of the RBC model to the simulation characteristics. The data groups 141a_d and 141e_h, differing only by the seed of the monitored RBCs, are used to verify the used comparative methods. We expect and confirm that the observed characteristics are the same for these groups.

Before the comparison of the datasets, we made the following adjustment: as already mentioned, even a simple comparison of x-velocities of several RBC pairs showed a systematically lower 141-node cell velocity. A simple recalculation of the complete record of all speeds for the 141a_d and 374a_d experiments shows that the speeds in the second group are approximately 8% higher. We have therefore multiplied by coefficient 1.08 the datasets containing 152 minimum, maximum and average velocities for groups 141a_d and 141e_h before performing statistical compliance tests.

As we can see from Figure 5, TABLE V. and TABLE VI. , a relatively good match was obtained between datasets 141a_d and 374a_d.

The presence of 8%-error signify, on one hand, the precision in which the simulation results can be interpreted. In case we can afford this imprecision, we can run the simulations with those simplified cells and have results within acceptable error bar. On the other hand, we can use this information to adapt the velocity of the flow or of the cells in the simulation channel, and to obtain results with better precision.

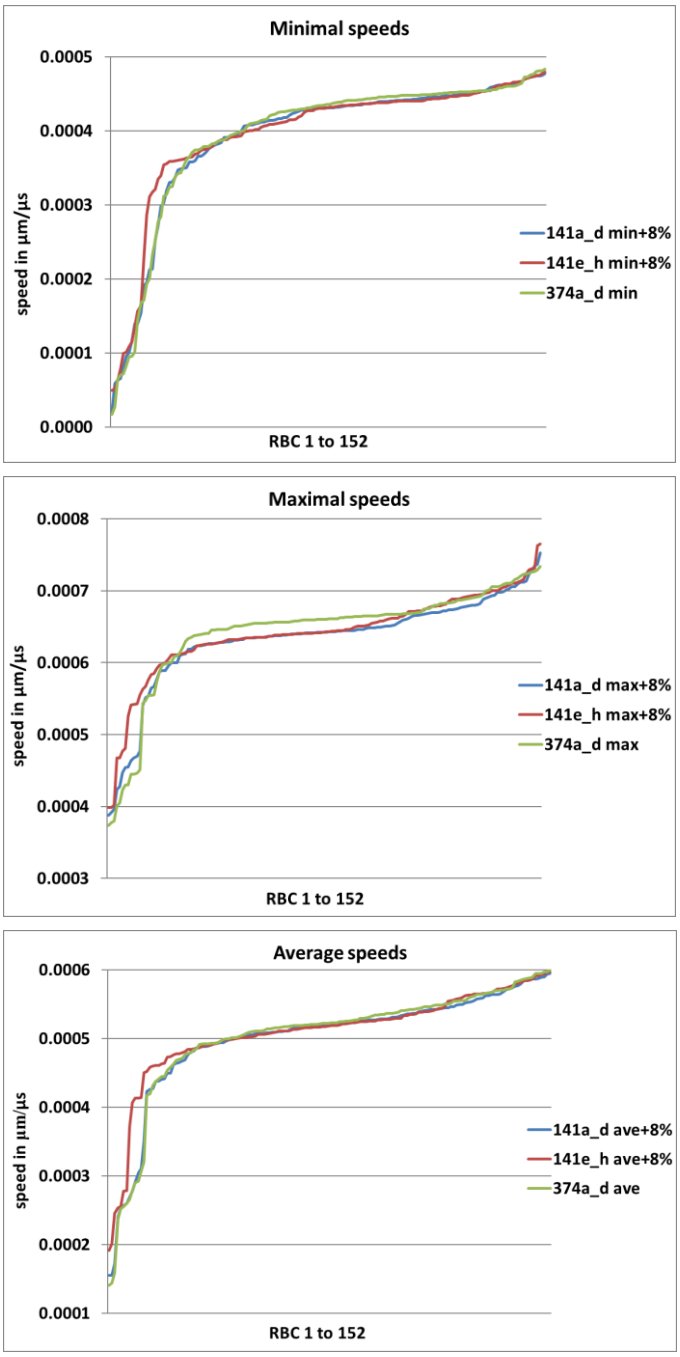


Figure 5. Illustration of datasets for minimal (maximal, average) velocities of RBCs. For each group of 50 cells 141a_d, 141e_h and 374a_d the line of ordered minimal (maximal, average) velocities is displayed. Graphs show its high similarity, confirmed by statistical tests.

The results of the statistical tests indicate that a 8% correction of the velocity values results in a high degree of data match of the individual simulations. The easily calculated coefficient 1.08 (or its inverted value) can therefore be used as a conversion for the velocity values obtained by the simulation, or, as a result of a more detailed examination, as a coefficient for changing the velocity of the fluid in the simulation experiment settings.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE V. RESULTS OF KS TEST FOR MINIMAL, MAXIMAL AND AVERAGE DEFORMATION DATASETS FOR ALL THREE GROUPS. THE NUMBERS OVER THE DIAGONAL REPRESENT THE P-VALUE OF TESTING H_0 HYPOTHESIS, THAT THE DATASETS COME FROM THE SAME DISTRIBUTION. UNDER THE DIAGONAL, WE INDICATE THE REJECTION OF THIS HYPOTHESIS ON SIGNIFICANCE LEVEL A=0.05. H_0 INDICATES THAT WE DO NOT REJECT IT, H_1 INDICATES ITS REJECTION.

Min	141*ad*	141*eh*	374*ad*
141*ad*		0.8066	0.1005
141*eh*	h_0		0.0098
374*ad*	h_0	h_1	

Max	141*ad*	141*eh*	374*ad*
141*ad*		0.8869	$1.07e^{-8}$
141*eh*	h_0		$6.25e^{-7}$
374*ad*	h_1	h_1	

Ave	141*ad*	141*eh*	374*ad*
141*ad*		0.8869	0.8066
141*eh*	h_0		0.5197
374*ad*	h_0	h_0	

TABLE VI. IN THE SAME FORM WE PRESENT THE RESULTS OF T-TEST, WHERE FOR ALL GROUPS ON SIGNIFICANCE A=0.05 WE ACCEPT H_0 HYPOTHESIS.

Min	141*ad*	141*eh*	374*ad*
141*ad*		0.7109	0.8533
141*eh*	h_0		0.8631
374*ad*	h_0	h_0	

Max	141*ad*	141*eh*	374*ad*
141*ad*		0.2594	0.2386
141*eh*	h_0		0.8628
374*ad*	h_0	h_0	

Ave	141*ad*	141*eh*	374*ad*
141*ad*		0.3437	0.7563
141*eh*	h_0		0.5545
374*ad*	h_0	h_0	

C. Comparison of velocities in channel slits

Similar observation can be done by comparing the velocity of the cells in slits. In all simulations which were run with 141-node cells, we recorded 32 passages of a RBC through the slit number 1, and 6 passages of a cell through the slit number 2. In simulations with 374-node cells, there were 9 passages of a RBC through the slit number 1, and 3 passages through the slit number 2. In all the simulations, any of the cells did not passed through the slit number 3. We observed the velocity of those cells inside of the slits. After that, we compared the average velocities. Results can be seen in TABLE VII.

We have observed as well an average velocity of the cells during entering the slits. This velocity was measured as a time needed to pass the part of the channel which contains the beginning of the obstacles, as depicted in Figure 6.

The observation was done with complete data from the simulation - in simulations with 141-node cells, there were 34 entrances to the slit 1 and 7 entrances to the slit 2. In simulations with 374-node cells, there were 15 entrances to the slit 1 and 3 entrances to the slit 2. The results of this comparison are presented in TABLE VIII.

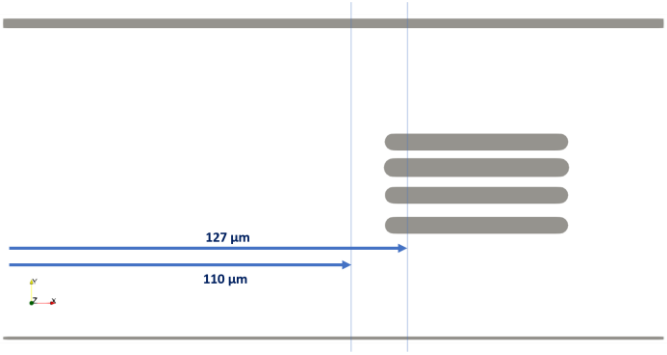


Figure 6. Measurement of the cells velocity during the entrance to the narrow slits. the velocity is calculated as the ratio of the distance depicted in the picture and the time needed to pass this distance

TABLE VII. COMPARISON OF AVERAGE X-VELOCITY OF A CELL PASSING THROUGH SLITS IN SIMULATIONS WITH 141-NODE CELLS AND IN SIMULATIONS WITH 374-NODE CELLS

	141-node simulations	374-node simulations	ratio "374"/"141"
slit 1	$0,000220 \mu m/\mu s$	$0,000239 \mu m/\mu s$	1.089
slit 2	$0,000118 \mu m/\mu s$	$0,000118 \mu m/\mu s$	1.001

TABLE VIII. COMPARISON OF AVERAGE X-VELOCITY OF A CELL ENTERING THE SLITS IN SIMULATIONS WITH 141-NODE CELLS AND IN SIMULATIONS WITH 374-NODE CELLS

	141-node simulations	374-node simulations	ratio "374"/"141"
slit 1	$0,000189 \mu m/\mu s$	$0,000197 \mu m/\mu s$	1.044
slit 2	$0,000105 \mu m/\mu s$	$0,000116 \mu m/\mu s$	1.097

D. Explication of the difference in velocities

As we can observe until now, the cells modeled with 374 node mesh are slightly faster. This can be explained by a distribution of the mesh nodes through the cells surface. As we can see in Figure 7, the majority of the points for both meshes is situated very close to the top and the bottom boundaries of the simulation channel. However, the cell model with 374 points have more nodes situated closer to the center of the channel (pro rata to the total number of nodes). That means that the 141-node cell are mainly moved by a liquid which is flowing close to the boundaries, and so it is flowing slower. The 374-node cell is moved also by the liquid which is moving inside of the channel, and so it flows faster than the liquid close to the boundaries.

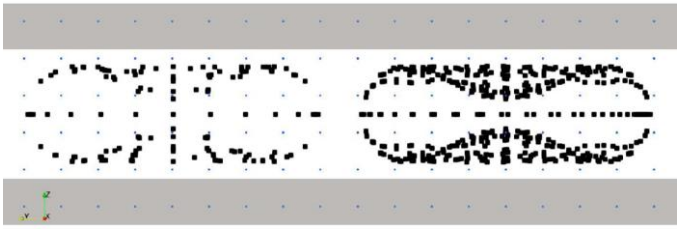


Figure 7. Distribution of the mesh nodes over the cells surface

E. Comparison of RBC deformations

In studies [2], [3] we have suggested several methods for comparing the rotation of RBCs in channels. This procedure could also be used in this study. However, due to the specific type of the channel and the observed RBC behavior, we have decided to modify this methodology to the RBC deformation monitoring.

Our aim was to use the simplest and most easily obtainable data output from the simulation again. As mentioned above, the output of simulation experiments contains also the coordinates of the circumscribed cuboid for each RBC. These values are also one of the basic outputs of real-time video processing. The chosen parameter describing the RBC deformation, taking into account the specifics of our experiment, is the width of the cuboid in y-direction. The brief verification confirmed that these values are reciprocal with the x-coordinate difference. The dimension of the cuboid in z-direction is not changing during the simulation, with respect to the relatively small height of the channel.

This values of the RBC deformation were then treated in the same way as the values of the cells velocities. For each RBC we first calculated values of the width of its cuboid during the whole simulation. From these values, we have selected minimum, maximum, and average values for each RBC. For each dataset of groups 141a_d, 141e_h and 374a_d, we obtained 152 values of minimum, maximum, and average y-width of the circumscribed cuboid. After that we've ordered their values by size.

In order to verify the match of simulation results, we again used depiction of the datasets, and statistical tests to confirm or reject the hypothesis that these datasets come from the same nonparametric distribution. The graphs of the values of the individual datasets shows considerable similarity, but the statistical results indicate the complexity of the problem. As shown in Figure 8 and in TABLE IX. and TABLE X. (with a similar structure as in the case of RBC velocities), only t-Test for mean of minimal y-width confirms our hypothesis.

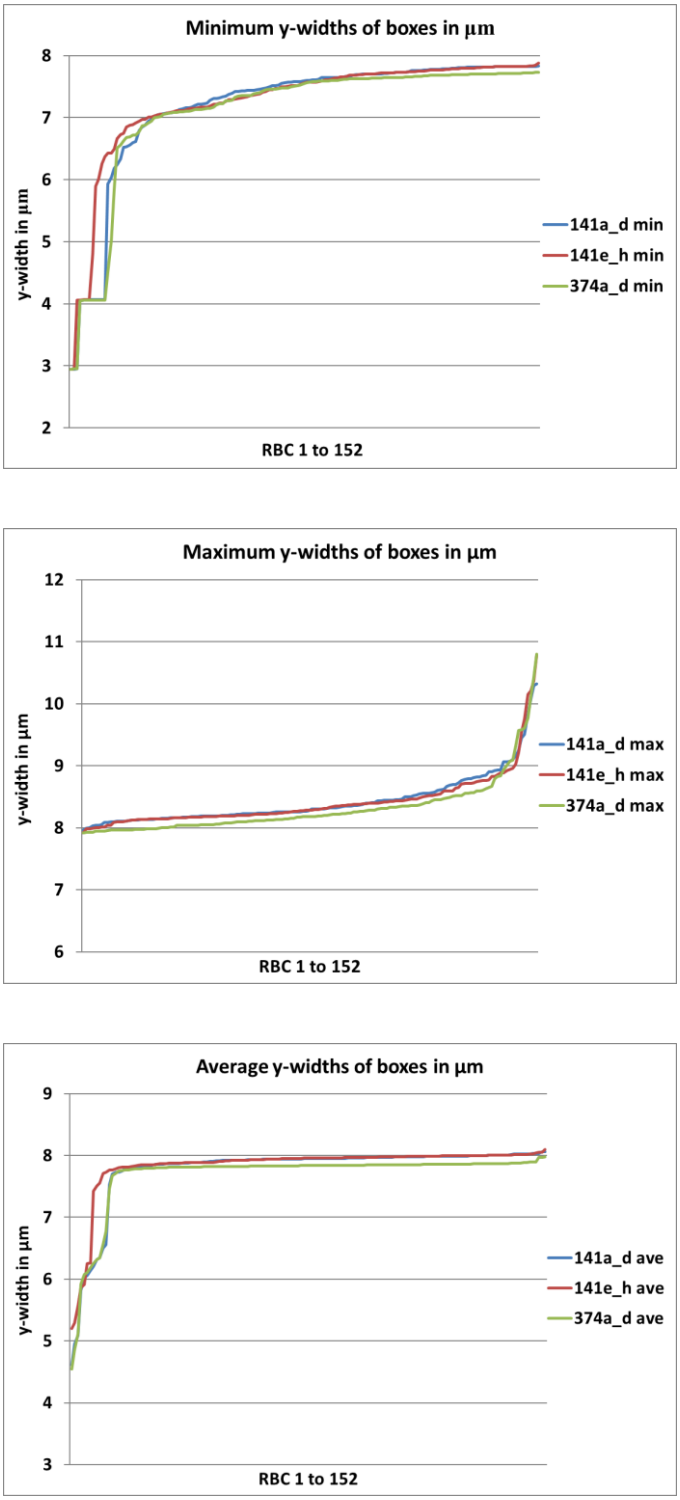


Figure 8. Graphs of datasets for minimal, maximal and average y-width of cells' circumscribed cuboids. Lines of ordered widths for groups of 50 cells 141a_d, 141e_h and 341a_d show very similar behavior of this parameters, but not sufficiently accurate for all statistical tests

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE IX. RESULTS OF KS TEST FOR MINIMAL, MAXIMAL AND AVERAGE DEFORMATION DATASETS FOR ALL THREE GROUPS. THE NUMBERS OVER THE DIAGONAL REPRESENT THE P-VALUE OF TESTING H_0 HYPOTHESIS, THAT THE DATASETS COME FROM THE SAME DISTRIBUTION. UNDER THE DIAGONAL, WE INDICATE THE REJECTION OF THIS HYPOTHESIS ON SIGNIFICANCE LEVEL A=0.05. H_0 INDICATES THAT WE DO NOT REJECT IT, H_1 INDICATES ITS REJECTION.

Min	141*ad*	141*ef*	374*ad*
141*ad*		0.8869	$1.17E-06$
141*ef*	h_0		$6.25E-07$
374*ad*	h_1	h_1	

Max	141*ad*	141*ef*	374*ad*
141*ad*		0.8869	$6.25E-07$
141*ef*	h_0		$1.17E-06$
374*ad*	h_1	h_1	

Ave	141*ad*	141*ef*	374*ad*
141*ad*		0.4305	$3.86E-34$
141*ef*	h_0		$4.67E-36$
374*ad*	h_1	h_1	

TABLE X. IN THE SAME FORM WE PRESENT THE RESULTS OF T-TEST, WHERE FOR ALL GROUPS ON SIGNIFICANCE A= 0.05 WE ACCEPT H_0 HYPOTHESIS

Min	141*ad*	141*ef*	374*ad*
141*ad*		0.5362	0.5445
141*ef*	h_0		0.202
374*ad*	h_0	h_0	

Max	141*ad*	141*ef*	374*ad*
141*ad*		0.7211	0.0182
141*ef*	h_0		0.0488
374*ad*	h_1	h_1	

Ave	141*ad*	141*ef*	374*ad*
141*ad*		0.3474	0.1649
141*ef*	h_0		0.0123
374*ad*	h_1	h_1	

VI. CONCLUSION

In our work, we made a comparison between simulations with different models of RBCs. We focused on characteristics of the movement of the cells - the velocity and the rotation. We found out that for the simulations with 374-node RBCs, the cells were moving faster than in the simulations with 141-node RBCs. The difference between the velocities of the two simulation sets was about 8%.

In order to confirm or reject the hypothesis about the similarity of the two types of simulations, we applied KS-test to the data obtained from the two simulations. In the case of simulations with 141-node RBC model, velocities were corrected by a coefficient 1.08. We think that this difference is due to different density of nodes in the cells surface. It seems that the 374-node RBC interacts better with the fluid.

The hypothesis about the compliance of the two velocity datasets is not rejected in the most of the cases. In the case of the two datasets with data about deformation of the cells, the hypothesis about the compliance of the two datasets was rejected in most of the cases.

On the other hand, we observed the behavior of the cells in simulations with equivalent initial seedings. We can state that qualitative results were not very different in models with 141 or 374-node RBC models. Every cell that enter a narrow slit in a 141-node RBC simulation, get across the same trajectory in the equivalent 374-node RBC simulation.

The model of the RBC does not have an important influence to the qualitative run of the studied simulations. The differences were important in case, where we were interested in quantitative characteristics of the simulations, as the exact values of the cell velocities, or their deformation during their life in the simulation box.

REFERENCES

[1] K. Bachratá and H. Bachratý, On modeling blood flow in microfluidic devices, ELEKTRO 2014: 10th International Conference, IEEE, ISBN 978-4799-3720-2, 2014, pp. 518-521

[2] K. Bachratá, H. Bachratý, M. Slavík, Statistics for comparison of simulations and experiments of flow of blood cells, EPJ Web of Conferences, Vol. 143, art. no. 02002, 2017

[3] H. Bachratý, K. Kovalčíková, K. Bachratá and M. Slavík, Methods of exploring the red blood cells rotation during the simulations in devices with periodic topology, 2017 International Conference on Information and Digital Technologies (IDT), 2017, pp. 36-46, doi: 10.1109/DT.2017.8024269

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

[5] I. Cimrák, M. Gusenbauer and I. Jančigová, An ESPResSo implementation of elastic objects immersed in a fluid, Computer Physics Communications, vol. 185, 2014, pp. 900-907

[6] R. Tóthová, I. Jančigová and M. Bušík, Calibration of elastic coefficients for spring-network model of red blood cell, International Conference on Information and Digital Technologies (IDT), 2015, pp. 376-380

[7] C. H. D. Tsai et all, An on-chip RBC deformability checker significantly improves velocity-deformation correlation, Micromachines, 7, 176, 2

Extending data flow coverage with redefinition analysis

Alexander Kolchin
V.M. Glushkov Institute of cybernetics
National Academy of Sciences
Kyiv, Ukraine
kolchin_av@yahoo.com

Stepan Potiyenko
V.M. Glushkov Institute of cybernetics
National Academy of Sciences
Kyiv, Ukraine

Thomas Weigert
Updraft LLC
Palatine, IL, USA
thomas.weigert@updraftworks.com

***Abstract*— Data flow-oriented coverage criteria are widely used in software testing. This paper proposes three novel def-use coverage criteria. The main objective of these criteria is to improve the specificity of test goals, to obtain more easily understood test cases, and to strengthen fault detection of derived test suites.**

Keywords— testing, coverage criteria, data flow analysis

I. INTRODUCTION

Testing is the most widely used means of assessing quality in software development. Formal test criteria aid in choosing appropriate test inputs to test with. Suitable criteria make it more likely that testers will find faults in the program and provide greater assurance that the system under test is of high quality and reliability. Selecting test cases typically relies on structural coverage criteria [1, 2]. Control flow-based criteria (such as statement and branch coverage) are too weak for defect detection, while path coverage is usually not realizable: the number of paths grows exponentially with the number of branches and may become infinite in programs containing loops. It is also critical that test cases are meaningful to a developer. If tests are completely unintuitive, testers are not likely to trust a test suite [3]. Data flow analysis appears to be a reason-able compromise [4–7]. Data flow coverage criteria examine cause-effect relations, input-output dependencies, etc., and thus achieve more meaningful test scenarios. Informally, data flow coverage focuses on how a variable is defined and used in the model. Data flow testing was introduced by Herman in 1976 [1, 4]. Widely used data flow-based criteria are direct def-use pairs [5], chains of def-use pairs connecting inputs with outputs [6], and k-definition/reference interactions [7]. These criteria can be used as a coverage metric to measure the level of thoroughness of a test suite or as a test goal or guiding heuristic in automatic tests generation [8].

In this paper we describe our experience in test suite generation using different data flow coverage criteria and highlight problems encountered, in particular multiple redefinitions before usage and insufficient examination of consequent redefinition contexts. We introduce improvements on data flow criteria: pure def-use pairs and two variants of def-redef-use triples.

II. BACKGROUND

Let G=(C, E, s, f) be a flow graph of a program, where C is a set of vertices, E is a set of edges, s the initial vertex, and f the final vertex. Each variable occurrence is classified as definition or use (in the right part of an assignment, in a parameter of an output, or in a predicate of a condition). A path on the graph G is a finite sequence of vertices c_0, c_1,…,c_k, where for all i, (0<i<k), an edge $(c_{i-1},c_i)\in E$. A complete path is a path where $c_0=s$, $c_k=f$. Let x be a variable and $c\in C$. Then defs(c) is the set of all variables defined at c (i.e., variables that are assigned a new value). A path p = (n, c_1, ..., c_k, m), k≥1, is called def-clear from vertex n to vertex m with respect to x if $x\notin defs(c_i)$ for any $1\le i\le k$.

Combinations of defs and uses form a family of coverage criteria [5]. Let P be a set of complete paths on a graph G. For example, P satisfies the *all-defs* criterion, if for every vertex c of G and every variable $x\in defs(c)$, P includes a def-clear path w.r.t. x from c to some usage; i.e., if all assignments will be used. P satisfies the *all-uses* criterion, if for every vertex c and every $x\in defs(c)$, P includes a def-clear path w.r.t. x from c to all uses, i.e., if each computation and condition affected by a definition of x will be tested. P satisfies the *all-paths* criterion, if P includes every complete path of G [5]. We describe the following extensions to these data flow coverage criteria: pure def-use pairs and def-redef-use triples, denoted as [D:U]v and [D:R:U]v, respectively, where v is a variable, and D, R and U are locations of definition, redefinition, and usage in the program flow graph.

III. RELATED WORK

Data flow testing is important because it augments control flow-based criteria by focusing on the definition and use of variables in the model, which tends to lead to more efficient and targeted test suites. For example, experiments described in [9, 10] indicate that given the same fixed amount of time for test generation, data flow testing achieves significantly higher mutation scores than test suites leveraging branch coverage.

Rapps and Weyuker [5] described a family of data flow coverage criteria distinguishing the use of variables between computations and predicates. The most practical criterion is all-uses since the number of paths required by all-du-paths or all-paths is too big or even infinite due to possible loops. Ntafos proposed the k-tuples strategy as an extension to def-use pairs [7] requiring all-k chains of direct data dependencies, i.e., for all sequences of k+1 nodes, each node must be directly dependent on the previous one. Laski and Korel [11] suggested a data dependency strategy relying on the liveness of vectors of variables treated as arguments to an instruction or program block. Ural and Hong [6] extended data flow testing with control dependence and tested data chains starting from inputs to outputs. In [12], model behavior context was introduced to refine test goal specifications. [3–15] extensively utilize data flow coverage to improve the efficiency of test suites. [3, 8, 9, 12, 14, 15] describe algorithms to derive test cases satisfying data flow coverage criteria from a formal model. [6] demonstrates that dependence-oriented coverage criteria can be characterized through temporal logic.

IV. PURE DEF-USE PAIRS

Motivation. Test cases should be simple, free of redundancy, intuitive to a developer, and goal-oriented. Relying on the

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

conventional data flow criteria in industrial projects, we encountered the following difficulties. First, tests often set up more than one scenario but then make a specific selection. The code snippet in Fig.1 shows some error processing. Def-use pairs for variable ret are: [2:9], [4:9], [6:9], [8:9]. The following test input set was generated: (1,1,1), (0,1,1), (0,0,1), (0,0,0).

```
1. in(a, b, c);
2. ret := 'ok';
3. if(a == 0)
4.    ret := 'err 1';
5. if(b == 0)
6.    ret := 'err 2';
7. if(c == 0)
8.    ret := 'err 3';
9. out ret;
```

Figure 1. Redundant redefinitions

Although this set satisfies the all-uses coverage criterion, the last two inputs indeed start more than one error scenario, but only the last one is observed. The purpose of such tests becomes unclear and they may lead to redundancies. While the last test case was technically correct and achieves the specified coverage (examines def-use pair [8:9] of variable ret), it was rejected by developers because it introduces additional scenarios beyond the last one really tested. (This would not be a problem if the test was produced fully automatically and its execution time were negligible). Secondly, when the same execution sequence is not guaranteed (due to parallelism and/or implementation peculiarity), these error codes may get rearranged and it may turn out that the test case describes the correct outcomes, but the test will nevertheless fail.

Solution. Introduce the *pure def-use pair* coverage criterion.

Definition. A sub-path $(n,c_1,...,c_k,m)$, $k\geq 1$, is *redef-clear* from vertex n to vertex m w.r.t. variable x if $x\notin defs(c_i) \vee c_i \in predom\_tree(m)$ for all $1\leq i\leq k$.

Definition. A path p covers a pure def-use pair [D:U]v if it covers def-use pair [D:U]v and there are no definitions of variable v before location D except for those belonging to the pre-dominator tree of D, i.e., if $p = (s, p_1, D, p_2, U, \ldots, f)$, where p_1 is redef-clear w.r.t. v and p_2 is def-clear w.r.t. v.

#	all-uses	outputs	def-use pairs	pure def-use pairs
1	(1,1,1)	'ok'	[2:9]ret	(1,1,1)
2	(0,1,1)	'err 1'	[4:9]ret	(0,1,1)
3	(**0**,0,**1**)	'err 2'	[6:9]ret	(**1,**0,1)
4	(**0**,**0**,0)	'err 3'	[8:9]ret	(**1**,**1**,0)

Figure 2. Redundant redefinitions: test cases

In the example in Fig.1, the assignment at line 2 belongs to the pre-dominator tree of the assignment at line 6, and thus does not prevent pair [6:9]ret to be pure. On the other hand, the assignment at line 4 does, and thus pair [6:9]ret can only be pure if the path will avoid line 4. Fig.2 describes the difference between a test suite satisfying the all-uses criterion and a test suite satisfying pure def-use pair coverage. Note that the tests examine the same set of def-use pairs, but the two last test cases, while resulting in the same outputs, start with different inputs and traverse different lines of code. For example, in test case #4, lines 4 and 6 will not be executed.

The pure def-use coverage criterion subsumes the all-uses criterion, as it only refines and strengthens the context of coverage. However, a test suite satisfying the pure def-use criterion may miss to cover def-use pairs which are not reachable without previous non-predominating redefinition.

In many practical situations, relying on pure def-use pair coverage allowed us to avoid test redundancy and led to simpler and shorter test cases and reduced the overall size of the test suites.

V. DEF-REDEF-USE TRIPLES

Motivation. During experiments assessing the mutation score of pure def-use pairs coverage, some mutations that were killed using all-uses coverage became hidden. This effect arises from the shortened test scenarios when requiring it to avoid possible redefinitions and led to undetected faults such as missed assignments or incorrect dominance frontiers.

Solution. Consider two types of usable redefinition contexts: def-redef$_v$-use and def-redef$_\Phi$-use. The former aims to check that the definition at redef$_v$ is not missed by requiring, in addition to def-use pair coverage, a change of the value of a variable at the tested definition (which thus becomes a redefinition). The latter aims to check for a dominance frontier fault requiring a change of value at locations which are incident with dominance frontiers (in static single assignment form [16]).

In Fig.3, the original program has a possible redefinition of variable ret (lines 5,6), which is omitted in the faulty program. Admissible test inputs for all-uses and pure def-use coverage are shown in Fig.4. The third test case from the all uses-based test suite *may accidentally* detect the missed assignment while the fault will *definitely* be missed by the test suite produced using the pure def-use criterion.

```
1. in(a, b, x, y);          1. in(a, b, x, y);
2. ret := x;                2. ret := x;
3. if(a == 1)               3. if(a == 1)
4.    ret := y;             4.    ret := y;
5. if(b == 1)               5.
6.    ret := x;             6.
7. out ret;                 7. out ret;
```

Figure 3. Missed assignment: original (left) and faulty (right) program

In order to strengthen fault detection, we apply def-redef$_v$-use coverage which requires the values of x and y to be different in order to change the value of variable **ret** during consequent assignment.

#	All-uses	Outputs	Pure def-use
1	(0,0,2,3)	2	(0,0,2,3)
2	(1,0,2,3)	3	(1,0,2,3)
3	(**1**,1,2,3)	2	(**0,**1,2,3)

Figure 4. Missed assignment: def-use test cases

Fig.5 shows an example with mixed dominance frontiers. The faulty program has an extra else statement at line 5 and thus the first conditional block changes its dominance frontier from line 5 to line 8. In order to detect such fault we need to check if

the assignment at line 7 is possible after line 4 has been executed. Def-redef$_\Phi$-use coverage will require such redefinition.

1. in(a, b, x, y);	1. in(a, b, x, y);
2. ret := x;	2. ret := x;
3. if(a == 1)	3. if(a == 1)
4. ret := y;	4. ret := y;
5.	**5. else**
6. if(b == 1)	6. if(b == 1)
7. ret := x;	7. ret := x;
8. out ret;	8. out ret;

Figure 5. Wrong dominance frontier: original (left) and faulty (right) program

Definition. A path p covers a def-redef$_v$-use triple [R:D:U]v if it includes location vertex R, such that [R:U]v is a valid def-use pair, and covers def-use pair [D:U]v, i.e., if p = (s, …, R, p_1, D, p_2, U, …, f), where p_1 is def-clear w.r.t. v, p_2 is def-clear w.r.t. v, and variable v changes its value at D.

Definition. A test suite is def-redef$_v$-use adequate w.r.t. def-use pair [D:U]v if it covers a def-use pair [R:U]v and includes at least one test case covering the def-redef$_v$-use triple [R:D:U]v.

Definition. Locations D and R are F-incident if there is a sub-path in the flow graph G leading from D to R and they have a common dominance frontier at location F.

Definition. Let DF(D) denote the set of vertices of dominance frontiers for vertex D. A test suite is weak def-redef$_\Phi$-use adequate w.r.t. def-use pair [D:U]v if it covers the def-use pair and it includes a test case covering a def-redef$_v$-use triple [D:R:U]v for each $F \in DF(D)$, where D and R are F-incident.

Extend the dominance frontier mapping from vertices to sets of vertices: $SDF(L) = \cup_{x \in L} DF(x)$.
The iterated dominance frontier $SDF^+(L)$ is the limit of the increasing sequence of sets of vertices:

$$SDF_1 = DF(L);$$
$$SDF_{i+1} = SDF(SDF_i).$$

Definition. A test suite is strong def-redef$_\Phi$-use adequate w.r.t. def-use pair [D:U]v if it covers the def-use pair and for each $F \in SDF^+(D)$ includes a test case such that its path covers a def-redef$_v$-use triple [D:R:U]v, where D and R are F-incident.

In many practical cases, applying weak def-redefΦ-use criterion is sufficient to detect missed assignments when using a maximal oracle approach. The strong version may result in larger test suites which can be improved by appropriate variable selection.

VI. Empirical evaluation

An empirical study aimed at evaluating the effect of the proposed criteria was performed. This study was guided by the following research questions:

RQ1: How do these criteria impact the size of the test suite compared to all-uses coverage?

RQ2: How do these criteria affect the ability of a test suite to detect faults?

RQ1 examines whether the proposed criteria result in significant changes to the size of the test suite. The proposed criteria impose additional restrictions on model behavior, and thus, covering some def-use pairs may not be feasible under these restrictions. In order to be able to compare the proposed extensions to data flow coverage, we augment the test suite based on all-uses in such cases. The size of the test suite will be assessed as the sum of all events in all tests.

RQ2 is concerned with test suite efficiency, which is not an objective of pure def-use coverage but is desirable nevertheless. For def-redef-use, we expect a significant increase in the mutation score.

Experimental procedure. Four different medium-sized models from the telecom, automotive, and finance domains were selected and are described in Table 1.

For each model, a set of mutants inducing typical faults was generated: operator reference fault (replace '∧' by '∨', or vice versa), negation fault (replace a variable of a sub-formula by its negation), associative shift fault (change the associativity of operators, such as $x_1 \wedge (x_2 \vee x_3)$ vs. $(x_1 \wedge x_2) \vee x_3$), missing variable fault (omit a condition in a formula, e.g., $x_1 \wedge x_2$ implemented as x_1), missing assignment fault (omit some assignment, leaving the previous variable value), dominance frontier fault (add or remove else statements, shifting the end points of conditional blocks) [8, 13, 17]. Data flow analysis was leveraged to prune untestable mutants: mutants which have no reachable usage are removed as "equivalent". The automatic test generation procedure described in [3, 15] was applied for all-uses, pure def-use, and pure def-use combined with weak def-redef$_\Phi$-use and missed pairs.

TABLE 1: MODELS USED

characteristics	model id			
	1	**2**	**3**	**4**
number of attributes	211	72	54	112
conditions (i.e., branches)	215	107	119	128
assignments	506	174	256	544
number of du-pairs	1013	493	541	1218
mutants generated	1233	814	758	1182

Results and analysis. Table 2 summarizes the observed results. Mutants killed refers to tests that fail to pass due to some transition having become inapplicable (therefore, the whole scenario is determined to be infeasible) or a mismatch detected in input parameters. These results indicate that def-redef-use positively affects fault detection. Remarkably, besides the quantitative improvement, tests obtained from pure def-use coverage often resemble manually written test scenarios.

Table 2 licenses the following conclusion regarding above research questions:

RQ1: On average, pure def-use coverage reduces the size of the test suite by 15-30%; the combined strategy increases its size roughly 30-35% when compared to all-uses coverage.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE 2: TEST GENERATION RESULTS

statistics	model id			
	1	**2**	**3**	**4**
test suite size (all-uses)	7624	1213	1340	3678
test suite size (pure def-use)	5955	1080	927	2831
test suite size (combined)	9976	1750	2024	5607
mutants killed (all-uses)	554	467	452	546
mutants killed (pure def-use)	479	401	375	468
mutants killed (combined)	693	651	536	743

RQ2: On average, pure def-use coverage decreases the mutation score by roughly 15% while the combined strategy increases it by 25-35% compared to all-uses coverage.

VII. CONCLUSION

Three new data flow coverage criteria are proposed. Each criterion can be used by itself or can be combined with other criteria [8] to refine tests.

The pure def-use strategy not only aims at reducing test suite size metrics, but also results in qualitative improvements of test suites with respect to readability, logical connectedness, redundancy, usefulness for debugging, and more. The pure def-use strategy produces simpler test cases that are considered more intuitive by practitioners.

The def-redef-use strategies attempt to find additional non-trivial paths resulting in a more exhaustive examination of system behavior and to increase the mutation score targeting the missing assignment and incorrect dominance frontier faults. These can easily be extended taking into account the type of variable usage: e.g., for p-use, the requirement to change the value of a variable at each redefinition location can be strengthened by requiring the result of predicate evaluation at the use-location to change.

Empirical results demonstrated positive effects on the overall quality and efficiency of test suites produced without significant impact on the size of the resulting test suite. We have developed a prototype tool implementing the criteria based on algorithms presented in [3] and [15]. Our experiments showed that the resultant test suites are more easily understood. Going forward, we plan to apply smart test suite minimization heuristics [8, 13] to improve test suite efficiency and to explore in more detail the impact of the proposed criteria on test suite size and the ability to reveal defects.

REFERENCES

[1] Su, T., et al.: A survey on data-flow testing. ACM Comput. Surv. 50, 5. (2017)

[2] Volkov, V., et al.: A survey of systematic methods for code-based test data generation. Artif. Intell. Vol.2, pp.71–85. (2017)

[3] Weigert at al. Generating Test Suites to Validate Legacy Systems. Lecture Notes in Computer Science, vol. 11753. pp. 3-23. (2019)

[4] Herman, P.M.: A data flow analysis approach to program testing. Aust. Comput. J. 8(3), pp. 92–97. (1976)

[5] Rapps S., Weyuker E. Data flow analysis techniques for test data selection. In: proc. of Int. Conf. of Softw. Eng. pp. 272-277. (1982)

[6] Hong H., Ural H. Dependence testing: extending data flow testing with control dependence. LNCS Vol.3502. pp. 23-39. (2005)

[7] Ntafos S. On required element testing. IEEE Trans. Softw.Eng. vol.10, pp.795-803. (1984)

[8] Kolchin A., Potiyenko S, Weigert T. Challenges for automated, model-based test scenario generation. Communications in Computer and Information Science, vol 1078. pp. 182-194. (2019)

[9] Vivanti M., Mis A. and oth. Search-based data-flow test generation. In IEEE Int. Symp. on Softw. Reliab. Engineering. 10p. (2013)

[10] Li N., Praphamontripong U., Offut J.: An experimental comparison of four unit test criteria: mutation, edge-pair, all-uses and prime path coverage. IEEE International Conference on Software Testing, Verification and Validation, pp. 220–229. (2009)

[11] Laski J., Korel B. A Data Flow Oriented Program Testing Strategy. In IEEE Transactions on Softw. Eng. vol. 9 (3). pp. 347–354. (1983)

[12] Kolchin, A.V. Interactive method for cumulative analysis of software formal models behavior. Proc. of the 11th Int. Conf. on Programming UkrPROG'2018, CEUR-WS vol. 2139, pp. 115–123. (2018)

[13] Kolchin A., Potiyenko S., Weigert T. Efficient increasing of the mutation score during model-based test suite generation. Proc. of the 12th Int. Conference UkrPROG'2020. CEUR-WS vol. 2866. pp. 331-341. (2020)

[14] Chaim M., Baral K. Offut J. Efficiently Finding Data Flow Subsumptions. 14th IEEE Conference on Software Testing, Verification and Validation (ICST). pp. 94-104. (2021)

[15] Kolchin A. A novel algorithm for attacking path explosion in model-based test generation for data flow coverage. Proc. of IEEE 1st Int. Conf. on System Analysis and Intelligent Computing, SAIC. pp. 226-231. (2018)

[16] Static single assignment book, 2018. [Online], Available at http://ssabook.gforge.inria.fr/latest/book.pdf

[17] Papadakis M. at al. Mutation Testing Advances: An Analysis and Survey. Advances in Computers, vol. 112. Elsevier, pp. 275 – 378. (2019)

The Benefits and Future of Remote Patient Monitoring

Kimberly Gandy, MD, PhD,[1,2] Myra Schmaderer, RN, PhD,[4] Anthony Szema, MD,[5,6] Chris March, BS,[6] Mary Topping, MBA[2], Anna Song, PhD,[7] Marcos Garcia-Ojeda, PhD,[7] Arthur Durazo, PhD[7], Jos Domen, PhD[2], Paul Barach, MD, MPH[2,8,9]

[1]Adjunct Associate Professor, Department of Biomedical and Health Informatics, University of Missouri, Kansas City.
[2] Play-it Health, Inc. 8101 College Boulevard, Overland Park, KS
[3] University Visiting Scholar, Stanford University, Stanford, CA
[4] University of Nebraska Medical Center, Omaha, NE
[5] Department of Technology and Society, College of Engineering and Applied Sciences, Stony Brook University, Stony Brook, NY
[6] Three Village Allergy and Immunology, South Setauket, NY
[7] University of California Merced School of Medicine, Merced, California
[8] Jefferson College of Population Health, Philadelphia, PA, USA
[9] Interdisciplinary Research Institute for Health Law and Science, Sigmund Freud University Vienna, Vienna, Austria
Corresponding Author: Kimberly Gandy, MD, PhD, Email: kgandy@playithealth.com

***Abstract*- The COVID-19 pandemic exposed the need to harness and leverage digital tools and technology for remote patient monitoring (RPM). RPM involves the monitoring of biometrics outside of the hospital or clinic setting with the associated transmission of data to clinicians in a manner that allows actionable insights to improve health outcomes [1]. There are many benefits of RPM for clinicians — ease of access to patient data, the ability to deliver higher-quality care to more patients with a lower risk of burnout — and for healthcare providers — lower costs and higher efficiency, to name just a couple. RPM will be limited in its impact until engagement and adherence can be maximized. We report on our ongoing implementation research with multiple academic and community partners to enhance RPM solutions to increase adherence and patient engagement in diverse patient settings. We demonstrate that RPM solution based on health behavior models and personalized coaching is technically feasible with high adherence rates. This work, however, remains in its infancy. Despite growing interest in remote patient monitoring, substantial gaps in the evidence base needed to be addressed using implementation and behavior science methods to support its ability to improve outcomes. Specifically, questions that remain unresolved include: How effective are RPM devices and associated interventions in changing important clinical outcomes of interest to patients and their clinicians? Which elements of RPM interventions lead to a higher likelihood of lasting success in affecting clinically meaningful outcomes? Which patient populations would benefit most?**

Rigorous, ongoing evaluation of RPM devices and platforms will be essential for elucidating their value and driving coverage decisions and adoption programs for the most effective solutions.

Keywords - mobile Health; patient-centered care; patient satisfaction; patient activation; patient empowerment; patient engagement; patient involvement; communication programs, remote patient monitoring; digital health

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

I. INTRODUCTION

A. Background

Improving patient safety and quality of care remain key goals of health care systems. In 2001, the Institute of Medicine identified patient-centered care as a health care quality indicator [2]. Within the past decade, top healthcare organizations have embraced mobile health (mHealth) as part of their patient-centered initiatives and their drive to achieve the quadruple aim [3] [4].

Peripheral biosensors are non-invasive devices used to acquire, transmit, process, store, and retrieve health-related data [1]. Biosensors have been integrated into a variety of hardware prototypes, including watches, wristbands, skin patches, shoes, belts, textiles, and smartphones [5] [1]. Patients have the option to share data obtained through biosensors with their providers or social networks to support clinical treatment decisions and disease self-management [6] [7].

RPM has long been integrated into focused areas of disease management, such as care of patients with pacemakers or implantable cardioverter–defibrillators. RPM for these patients can reduce costs and supplement or replace in-office care, while offering convenience and heightened surveillance for clinical events. In recent years, RPM technology has expanded into new areas, including chronic and acute care management for multiple common conditions. Devices used in patients' homes now capture physiological parameters such as weight, blood pressure, oxygen saturation, and blood glucose levels and transmit these data to clinicians for review. For example, wrist-worn pulse oximeters transmitting oxygen-saturation data may be used to monitor lung function in patients with chronic obstructive pulmonary disease, and continuous glucose monitors may wirelessly transmit to physicians information about blood-sugar control in diabetic patients at different times of day and between office visits [8].

The concept of leveraging technological innovations to enhance care delivery has many names in healthcare. The terms digital health, mobile health, mHealth, wireless health, Health 2.0, eHealth, quantified self, self-tracking, telehealth, telemedicine, precision medicine, personalized medicine, and connected health are among the terms often used synonymously [9]. A 2007 study found 104 individual definitions for the term telemedicine [10]. For the purpose of this study, we define remote patient monitoring (RPM) as the use of a non-invasive peripheral device that automatically transmits data to a web portal or mobile app for patient self-monitoring and/or health provider assessment and clinical decision-making.

It is possible, although not yet demonstrated at scale, that evidence-based RPM can improve clinical outcomes. The literature on RPM reveals enthusiasm overhype about promises to improve patient outcomes, reduce healthcare utilization, decrease costs, provide abundant data for research, and increase physician satisfaction [11]. Non-invasive biosensors allow RPM to offer patients and clinicians real-time data that has the potential to improve the timeliness of care, boost treatment adherence, and drive improved health outcomes [12]. The passive gathering of data may also permit clinicians to focus their efforts on diagnosing, educating, and treating patients, and ultimately, to improving productivity and efficiency of the care provided. Numerous studies reveal that treating blood pressure obtained in office visits does not allow clinicians to optimally improve a patient's condition. In hindsight, we may wonder why this realization has taken so long.

Virtual care is a multi-component entity [13]. Telehealth emerged as the stimulus for virtual care development and has served as the foundation upon which other critical elements are being built. One such critical element of remote RPM, is the ability to accurately and consistently monitor patient biometrics in the home or other settings where the patient spends the overwhelming majority of their time. A recent systematic review and examination of high-quality studies on RPM found that remote patient monitoring showed early promise in improving

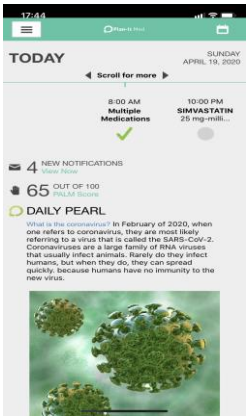


Figure 1. Opening screen of the Plan-it Med app shows both adherence and educational content data.

outcomes for patients with select conditions, including obstructive pulmonary disease, Parkinson's disease, hypertension, and low back pain [14]. Remote patient monitoring can allow determination of truly representative biometric data baselines from which to make interventions. As a result, it is possible to know if the treatments that we prescribe are effective in altering the day-to-day lives of our patients [15]. The ability to design more effective and sustainable clinical trials and treatments that can dramatically improve wellness [16].

B. The Success of Digital Health Will Continue to Be Determined by Patient and Staff Trust & Engagement

We have known for decades that the lack of adherence to the regimens prescribed by clinicians contributes to well-documented deleterious effects of chronic illnesses in spite of optimal prescribed treatments [17]. A combination of technology and applied behavioral science innovation is necessary, however, to assure RPM and virtual care reach significant levels of adherence and their resultant full potential [18]. Without high engagement in RPM, the biometric insights gained are limited, regardless of the capacity of technology to actually capture and reflect back to patients and providers real time measurements.

Virtual care is effective when consideration is given to the implementation challenges associated with the innovations. Key implementation elements include validated health behavior models, care pathways, and tailored coaching programs. The implementation cycle must begin with an understanding of the patient's true unmet needs; continue with monitoring the patient in the settings of their daily lives; and, conclude by contributing to a better understanding by the patient of their health and wellness needs, and supporting the goals of the patient, clinician, and health system.

The journey of Play-It-Health (PIH) began in behavioral science and transplantation research. We noted that children were dying after transplantation because they were not receiving their transplant medications properly, thus compromising the health of their transplant organs and their lives [19], [20]. Though the story of these children was top of our minds it was clear that this scenario extended far beyond the confines of transplantation or pediatrics [21]. The majority of patients do not adhere to their prescribed regimens for a wide variety of reasons, and the results have dire medical consequences [22]. Patients will continue to suffer from poor outcomes until adherence is improved, and clinicians and patients work together to co-design treatment goals cooperatively. The reasons for nonadherence are numerous and understandable. They include a lack of understanding of the importance of the prescribed drug regimens, a lack of resources for obtaining or following through on therapies, forgetfulness, lack of coordination with a caregiver, a perceived mismatch or lack of resonance between the patient and the clinician of clinical realities or treatment priorities, and rebelliousness [23]. Contrary to popular opinion, this latter reason represents a rare contributing cause. Often underestimated is the perceived mismatch between the realities as perceived by the clinicians and what is driving patient behaviors. It is this mismatch in mental models that remote patient monitoring may have a powerful role in addressing, thereby bridging a huge existent gap in clinical care.

The dominant thinking ten years ago was that the problems in patient adherence were the fault of the medical community and its inability to listen to patients and rapidly incorporate the latest technology solutions. A flood of apps and digital health solutions hit the market, many of which were intentionally designed to exclude clinicians from treatment algorithms. It took about 5 years for the digital health community to realize that patients still want their clinicians to be involved in their care, but significant damage had been done to the uptake and trust of digital health efforts. Ineffective studies coupled with over-promised outcomes dampened the enthusiasm for many promising technologies.

From the outset, we designed our RPM solution with evidenced-based data as a guide, incorporating the deep insights and input from clinicians into all treatment algorithms, and have remained committed to this course.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

II. METHODS

We report on data from two groups of patients.

The first group is part of an IRB-approved prospective study of 72 patients admitted to the hospital in 2018 with heart failure from a predominantly rural area. Patients were asked to engage in biometric monitoring on weight scales, track their medications, education, and engage in a monthly telehealth visit. Patients were randomly enrolled at the time of discharge into three cohorts: (i) usual care (UC), n = 26 (UC: mobile application (Plan-it Med®, PiM) with medication, appointment and remote monitoring regimen entered into the mobile application); (ii) mHealth, n = 23: mobile application as in group 1 plus app-mediated reminders and incentive verbiage; and (iii) mHealth+, n = 23: everything in group 2 and a monthly telehealth meeting through PiM with a nurse practitioner or community health worker.

The second group included 102 patients enrolled in Plan-it Med® from a pulmonary clinic over the course of a year, beginning in February of 2020. These patients were enrolled for acute intervention and management during the course of the COVID-19 pandemic. These patients used the PIM app predominantly for remote patient monitoring and education regarding asthma and COVID-19. They were asked to check their pulse oximetry daily with an iHealth pulse oximeter that interfaced with the PiM system.

A. The Plan-it Med® Mobile Technology Platform

The Plan-it Med® platform tracks, scores, and incentivizes adherence to the critical components of a clinical regimen. The system provides reminders when medications, appointments, labs, or biometric measurements are due; tracks when patients are engaged; and provides incentive messaging to encourage medication, appointment, and RPM adherence. The system also provides daily education with associated questions in reference to why medications or treatment regimens are needed (Figure 1). Educational modules address twenty different clinical conditions and can be modified in real time. This latter functionality was utilized frequently during the pandemic to provide real time updates. The system is device-agnostic and interfaces with a variety of peripheral devices. The platform tracks patient and clinical interactions throughout their daily routines and in all settings.

B. Outcome Measures

In both groups, patient engagement with the PiM application and adherence to the requested regimens for remote patient monitoring were monitored. In both groups, hospital admissions, emergency room and urgent care visits, and complications were tracked. In the first study, patient satisfaction surveys were performed, and medication adherence was also tracked. In the second group, the number of abnormal biometric parameters and the associated interventions and diagnoses were tracked.

C. Data Analysis

In the first study, descriptive statistics were reported as counts (%) and means. All variables were examined for meeting the assumptions of the statistical tests used. Enrollment (recruitment efficiency and attrition), intervention delivery, usability and acceptability counts (%) were calculated.

III. RESULTS

A. Medication adherence

Figure 2 shows the amount of improvement of medication adherence in Groups 2 and 3 above the adherence level of UC patients. The level of adherence for UC patients is the baseline, or 0, in this graph. In other words, this graph is designed to demonstrate the adherence improvement in the mHealth+ groups above the adherence and engagement of those patients given the app without its full functionality. After a year of iterations,

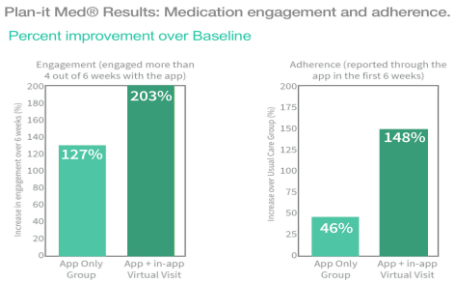


Figure 2. Medication engagement and adherence.

engagement, and adherence 127 patients demonstrated improved measures (46%), respectively, in the mHealth group and 203 and 148%, respectively, in the mHealth+ group. These boosts in engagement were achieved strictly with reminders, positive reinforcement cues, point accumulation, and follow up messaging. The rewards platform was not deployed.

B. Adherence to RPM in Plan-it Med® is High. (Fig 3)

Pulse oximetry data is shown from 38 patients onboarded early in the COVID-19 pandemic. This period of monitoring shown is representative of the critical period for monitoring in the containment of SARS-CoV-2 after a presumptive exposure event. As shown in Figure 3, adherence to daily pulse oximetry monitoring is high and is 96% at 9 weeks after enrollment.

C. High Long-Term Adherence in Plan-it Med®.

Adherence in this population was also high months after onboarding. In patients that were still on the platform an average of 11.5 months after onboarding, over 65% of the patients were still using the platform every day, and over 88% of the patients were using the platform every other day (Figure 4).

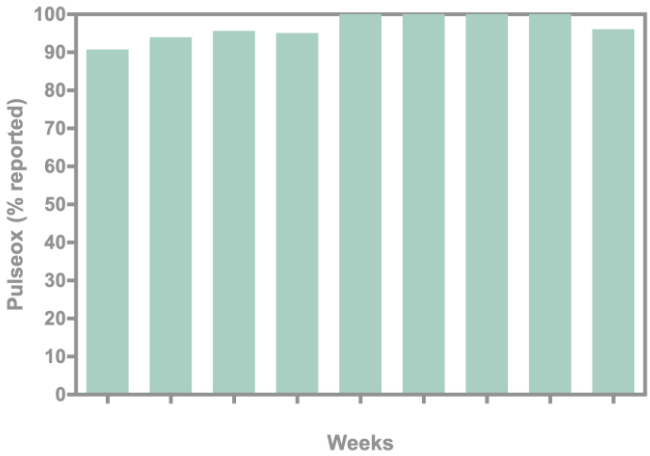


Figure 3. Weekly RPM (pulse oximeter data) reporting the first nine weeks the patients are participating

In the second study, there is a definite attrition of patients using RPM. Of the patients that stopped using RPM, forty percent stopped when they felt their conditions had improved or the threat of pulmonary complications from COVID had passed; forty-three percent stopped using PiM secondary to associated costs incurred with copays or deductibles. Digital health technologies will need to carefully consider patient attrition as patients improve and the impacts on long term wellness.

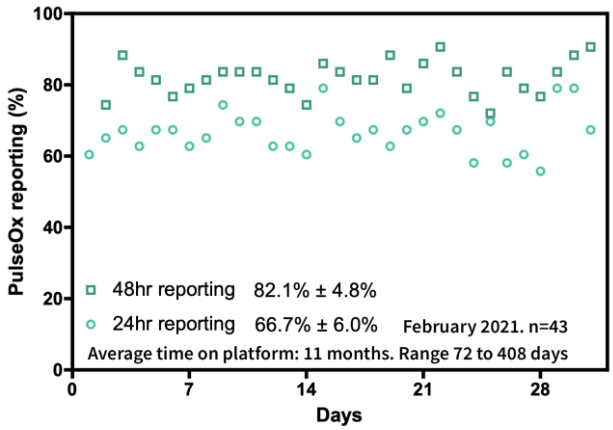


Figure 4. Reporting rate for pulse oximeter data for patients 11 months after starting the protocol. A majority reports daily, and a large majority reports at least every other day.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

IV. DISCUSSION

A. *There are numerous technical models for effective collection of RPM data*

For several decades, digital health solutions have allowed biometrics to be monitored outside of the hospital. Within the last decade, however, these solutions were coupled with the capacity to transmit this data digitally and easily to central collection systems where patients and clinicians could view the data more easily and continuously. Cloud based technologies expanded these capacities and the associated data sharing.

Data can now be reliably transmitted via a variety of methods to clinicians. Data can be collected through devices that harbor cellular chips for direct transmission of data from the device. Data can also be collected on devices that have blue-tooth capacity to transmit this data to a smart device such as a phone or tablet or even a cellular hub. Finally, phones can be turned into cellular hubs. There are privacy and security benefits and drawbacks to each model system, and the data transmission must be matched to the client and/or patient resources and preferences.

B. *Nonadherence is a problem in many RPM systems, limiting outcome improvement*

Interestingly and not unexpectedly, the same problems with nonadherence that plagued the health industry in relation to other health behaviors have continued to plague the remote patient monitoring industry. Lack of adherence or engagement with remote patient monitoring is one of the biggest barriers to both effective implementation and achievement of meaningful results.

The reasons for RPM nonadherence can be divided into four distinct areas:

- False expectations. Patients with relatively normal vitals cannot be expected to monitor those vitals each morning in a vacuum. Unless the patient is a regimented individual that likes the repetitiveness of routines, it will be hard to engage a patient in this type of unchanging regimen if they are not receiving feedback. Indeed, it has long been appreciated in clinical care that if a patient feels they are doing well they will not engage in a repetitive, time consuming behavior that yields little data that would change their care.
- Cost. Patients are incurring unexpected costs with RPM. Before the pandemic, Medicare patients were often seeing copays associated with RPM. This was a clear deterrent to usage. Additionally, some patients insured with private payers were encountering RPM associated payments due to deductibles on their associated plans.
- Inconvenience. In some early models of RPM tracking, capturing biometrics was just inconvenient, difficult, and/or time-consuming. Ease of use is improving as the technology matures.

There is already a backlash in the community in reference to this lack of engagement with existing digital health systems. Many organizations offered RPM solutions with little associated communication or engagement tools. As a result, adherence was low. Some organizations increased FTEs to engage in RPM and were unable to bill when they were not receiving the necessary data per month. This has dampened the enthusiasm for RPM in some circles.

As Plan-it Med® was designed with adherence and engagement as a primary goal, RPM existed in a platform positioned to render high levels of adherence. The incentives and positive messaging have been optimized over years to improve adherence. The system also offers five flexible methods of communication with patients: in-app messaging, texting, telehealth, secure e-mail, and phone calls. This allows communication to be tailored to a given patient's wishes and to have multiple ways to reach a patient. Though our levels of adherence are high, they can continue to be optimized as we better understand the variable responses and uptake in our preliminary studies.

C. *Lessons Learned from the RPM Work*

- Patients will trust RPM and need to truly see the benefit of RPM for their care at the time of enrollment. This should be the primary focus of the enrollment visit from the clinician, whether virtual or in person.

- Patients need to understand the associated costs, if any. Every effort should be made to clarify these issues prior to implementation in a transparent and truthful manner.
- For some patients, the benefits of RPM will be in the recuperation from acute events. Goals for this use case should be established up front as a design feature not a bug to design out.
- The technology should be matched to the patient. It is best to offer a variety of interface options. Different patients have different engagement expectations, connectivity and device needs, and options need to be available.
- Patients should be allowed to disengage and re-engage with RPM, based on the evolution of their clinical conditions without penalty or censure.
- Multiple communication avenues should be offered to patients and their understanding of these methods should be assured and verified. Some patients will prefer phone calls. Some will prefer telehealth visits. Some will prefer in-app messaging or texts. Regardless of the method preferred, maintaining communication and alignment of needs should be the key objectives.

D. Moving the Population Health Needle Will Require RPM to Reach Diverse Populations

To truthfully move the needle in outcomes and population health, however, RPM will have to reach the patients that need it most, the underserved and those in rural communities. As often happens in healthcare, the financial models up to this point, have not benefited these populations [30]. In fact, rural health centers (RHCs) and Federally Qualified Health Centers (FQHCs) have heretofore been unable to receive additional reimbursement for RPM services. RPM solutions had to be bundled into existing programs and or already allocated funding allotments.

With the anticipated initiatives for home-care and the emphasis on the underserved in the new infrastructure legislation, this could change. There is now optimism that the work to optimize RPM for rural and underserved populations will be revitalized and better supported. This new trend could have tremendous implications for population health. As markets are pushed to develop systems that reach broader populations, the extrapolations to international markets will be facilitated, and entirely new frontiers of collaboration and innovation could be opened.

Conclusions: Future of Digital Health and RPM

As the use of remote patient monitoring services grows — driven by health care limitations imposed by the Covid-19 pandemic — clinicians, payers, and patients face important questions regarding the volume, value, and appropriate use of this care model.

The next frontier for remote patient monitoring will involve using RPM data to develop predictive algorithms for care [24]. Monitoring current clinical scenarios at the start will lead to predicting clinical scenarios, facilitate targeted intervention, and potentiate preventative care using large datasets and machine learning tools. As RPM datasets are generated from intermittent biometric data to continuous biometric data and as proteomic and genetic data assays become more affordable, new predictive algorithms will emerge. It will be critical to determine exactly what predictions are desired and require an understanding of the predictions that will potentiate effective interventions.

Future research should identify and remedy potential barriers to RPM effectiveness and impact on clinical outcomes. For example, factorial design trials could evaluate variants of an RPM intervention in terms of frequency, duration, intensity, and timing. We also found that there are few large-scale clinical trials demonstrating a clinically meaningful impact on patient outcomes. Most studies have relatively short follow-up periods. Given that many of these studies were described as pilot studies, it is clear that the field of RPM is relatively new and evolving. Larger studies with multiple intervention groups will better distinguish which components are most effective and whether behavior changes can be sustained over time using RPM.

Future studies would also highly benefit from a mixed-methods approach in which both patients and clinicians are interviewed [25]. Adding a qualitative component will give researchers

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

insights into which RPM elements best engage and motivate patients, nurses, allied health workers, and physicians. Behavior change is complex; understanding how and if specific devices and device-related interventions and incentives motivate health behavior change is an important area that is still not well understood. For example, previous studies have found that most devices result in only short-term changes in behavior and motivation [26]. Activity trackers have been found to change behavior for only approximately three to six months [27]. Studies have found that cash incentives performed worse than charity incentives, illustrating that incentivizing individuals is complex and nuanced. Gaining a better understanding of how individual users interface with these health-related technologies will assist in developing evidence-based devices that have the potential to change behavior over longer periods of time.

An inherent shortcoming of most peripheral device studies is difficulty in following double-blind procedures; the intervention arms necessarily include patient engagement or, at minimum, placement of the device on the patient's body, which can be difficult to blind. Some studies have used devices that were turned off or were non-functional to reduce a potential placebo or Hawthorne effect [28], but given the data feedback loop integrated into many of these devices, it is extremely difficult to blind the provider receiving the data, which may impact results. Nonetheless, this shortcoming would tend to benefit the active intervention, making it more likely to show a difference in an unblinded study.

For RPM interventions to impact healthcare, they will need to impact outcomes that matter to patients [29]. Examples include patient-reported health related quality of life, symptom severity, satisfaction with care, resource utilization, hospitalizations, readmissions, and survival. There is little data investigating the impact of RPM on these outcome measures. The interventions are likely to succeed if they are developed directly and cooperatively in partnership with end-users—i.e. patients and front line clinicians.

Rigorous, ongoing evaluation of RPM devices and platforms will be essential for elucidating their value and driving coverage decisions and adoption programs for the most effective solutions.

V. References

1. Majumder S, Mondal T, Deen MJ. Wearable Sensors for Remote Health Monitoring. Sensor. 2017;17:130.
2. Barach P. Crossing the Quality Chasm: A New Health System For the 21St Century. Washington DC: Institute of Medicine; 2001.
3. Gafur S, Schneider EC. Engaging Patients Using Digital Technology — Learning from Other Industries. N Engl J Med [Internet]. 2019 [cited 2021 May 21]; Available from: https://catalyst.nejm.org/patients-digital-consumer-focused-industries/
4. Wang A, Ahmed R, Ray J, Hughes P, McCoy E, Auerbach MA, et al. Supporting the Quadruple Aim Using Simulation and Human Factors During COVID-19 Care. Am J Qual. 2021;36:73–83.
5. Andreu-Perez J, Leff DR, Ip HM, Yang GZ. From wearable sensors to smart implants—toward pervasive and personalized healthcare. IEEE Trans Biomed Eng. 2015;62:2750–62.
6. Ajami S, Teimouri F. Features and application of wearable biosensors in medical care. J Res Med Sci. 2015;20:1208–15.
7. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med. 2015;7:283.
8. Dunn P, Hazzard E. Technology approaches to digital health literacy. Int J Cardiol. 2019;293(October 15):294–6.
9. Atallah L, Lo B, Yang GZ. Can pervasive sensing address current challenges in global healthcare? J Epidemiol Glob Health. 2012;2:1–13.
10. Dobkin BH, Dorsch A. The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. Neurorehabil Neural Repair. 2011;25:788–98.
11. Oh H, Rizo C, Enkin M, Jadad A. What is eHealth (3): a systematic review of published definitions. J Med Internet Res. 2005;7:e1.
12. Pevnick JM, Fuller G, Duncan R, Spiegel BMR. A large-scale initiative inviting patients to share personal fitness tracker data with their providers: initial results. PLoS ONE. 2016;11:e0165908.
13. Omboni S, McManus RJ, Bosworth HB, Chappell LC, Green BB, Kario K, et al. Evidence and Recommendations on the Use of Telemedicine for the Management of Arterial Hypertension: An International Expert Position Paper. Hypertension. 2020;76(5):1368–83.
14. Noah B, Keller MS, Mosadeghi S, Stein L, Johl S, Delshad S, et al. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. NPJ Digit Med. 1:20172.
15. Vegesna A, Tran M, Angelaccio M, Arcona S. Remote patient monitoring via non-invasive digital technologies: a systematic review. Telemed J E Health. 2017;23:3–17.
16. Chau JP-C, Lee DT-F, Yu DS-F, Chow AY-M, Yu WC, Chair S-Y, et al. A feasibility study to investigate the acceptability and potential effectiveness of a telecare service for older people with chronic obstructive pulmonary disease. Int J Med Inf. 2012;81:674–82.
17. Martin LR, Williams SL, Haskard KB, Dimatteo MR. The challenge of patient adherence. Ther Clin Risk Manag. 2005;1(3):189–98.
18. Hilty DM, Armstrong CM, Edwards-Stuart A, Gentry MT, Luxton DD, Krupinski EA. Sensor, Wearable, and Remote Patient Monitoring Competencies for Clinical Care and Training: Scoping Review. J Technol Behav Sci [Internet]. 2021 [cited 2021 May 20]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7819828/pdf/41347_2020_Article_190.pdf
19. Dobbels F, Van Damme-Lombaert R, Vanhaecke J, De Geest S. Growing pains: non-adherence with the immunosuppressive regimen in adolescent transplant recipients. Pediatr Transpl. 2005 Jun;9(3):381–90.
20. Kaufman M, Shemesh E, Benton T. The adolescent transplant recipient. Pediatr Clin North Am. 2010 Apr;57(2):575–92, table of contents.
21. Kaplan A, Price D. Treatment Adherence in Adolescents with Asthma. J Asthma Allergy. 2020;13:39–49.
22. Lin C-S, Khan H, Chang R-Y, Liao W-C, Chen Y-H, Siao S-Y, et al. A study on the impact of poor medication adherence on health status and medical expense for diabetes mellitus patients in Taiwan. Med Baltim. 2020;99:e20800.
23. Jin J, Sklar GE, Min Sen Oh V, Chuen Li S. Factors affecting therapeutic compliance: A review from the patient's perspective. Ther Clin Risk Manag. 2008;4(1):269–86.
24. El-Rashidy N, El-Sappagh S, Islam SMR, El-Bakry HM, Abdelrazek S. Mobile Health in Remote Patient Monitoring for Chronic Diseases: Principles, Trends, and Challenges. Diagnostics. 2021;11:607.
25. Bruce CR, Harrison P, Nisar T, Giammattei C, Tan NM, Bliven C, et al. Assessing the Impact of

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Patient-Facing Mobile Health Technology on Patient Outcomes: Retrospective Observational Cohort Study. JMIR Mhealth Uhealth. 2020;8(6).
26. Klasnja P, Consolvo S, Pratt W. In Proc. SIGCHI Conference on Human Factors Computing Systems. In Vancouver, BC Canada; 2011.
27. Shih PC, Han K, Poole ES, Rosson MB, Carroll JM. Use and adoption challenges of wearable activity trackers. In: iConf Proc. Newport Beach, CA; 2015.
28. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. J Clin Epidemiol. 2014;67:267–77.
29. Subbe C, Barach P. Impact of Electronic Health Records on Pre-defined Safety Outcomes in Patients Admitted to Hospital. A Scoping Review. BMJ Open. 2011;11:e047446.
30. Tsao LR, Villanueva SA, Pines DA, Pham MN, Choo EM, Tang MC, et al. Impact of Rapid Transition to Telemedicine-Based Delivery on Allergy/Immunology Care During COVID-19. J Allergy Clin Immunol Pract.

Multi-UAV Routing for Critical Iinfrastructure Monitoring Considering Failures of UAVs:

Reliability Models, Rerouting Algorithms, Industrial Case

Ihor Kliushnikov, Vyacheslav Kharchenko, Herman Fesenko
Department of Computer Systems, Networks and Cyber Security
National Aerospace University "KhAI"
Kharkiv, Ukraine
{i.kliushnikov, v.kharchenko, h.fesenko }@csn.khai.edu

Elena Zaitseva
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
elena.zaitseva@fri.uniza.sk

***Abstract*—Route-based reliability models of the unmanned aerial vehicle (UAV) fleet carrying out monitoring of critical infrastructure facilities and comprising main and redundant UAVs are developed. A nuclear power plant (NPP) and monitoring stations (MSs) are considered as a critical infrastructure and critical infrastructure facilities, respectively. These models are used when multi-UAV routing for NPP monitoring and allow calculating the probability of the successful fulfilment of the plan (SFP) for the UAV fleet to cover the whole target MSs of the NPP. The dependencies showing the relationship of the probability of the SFP to both the UAV reliability function and used route-based reliability models are obtained and explored. An example of the proposed models application for routing main and redundant UAVs of the fleet to cover the whole target MSs of the Zaporizhzhia NPP is given.**

Keywords— unmanned aerial vehicle; routing; route-based reliability model; monitoring station; nuclear power plant

I. INTRODUCTION

A. Motivation

The Fukushima NPP accident showed that wired networks, connecting MSs of the automated radiation monitoring system to the crisis centre (CrS), are vulnerable to both natural and man-made disasters. To cope with the similar problems, UAV-based wireless subsystems ((UAV)-enabled wireless networks) can be deployed.

Importantly, the use of UAVs for monitoring of critical infrastructure facilities has some features [1-3]:

- environmental conditions may differ from the normal conditions of the UAV utilization (smoke, high temperature, radiation);
- external conditions may very according to the section features of the UAV route;
- UAV reliability indicators may deteriorate due to the negative impact of the environment;
- when deploying monitoring systems, ready-made commercial UAVs are often used as integral parts of monitoring systems without fully meeting the requirements for their reliability characteristics.

Therefore, when designing a UAV-based monitoring system, it is necessary to ensure its required level of reliability, considering the features listed above, failures of UAVs, the number of the target MSs, and other parameters of the system.

B. State of the Art

The problems of using UAV-based monitoring systems are considered in many studies. Monitoring of critical infrastructure facilities via UAVs may cover:

- radiation dose rate measurement, air sampling, surveying or mapping [4-6];
- detection of nuclear sources [7];
- location of lost radioactive sources [8];
- characterizing remediation effectiveness [9];
- deployment of an Internet of UAV-based accident monitoring system [10], [11].

Study [12] deals with a multirobot scheduling problem in which UAVs have to be recharged during a long-term mission. The study introduces a separate team of dedicated charging robots that the UAVs can dock with in order to recharge. The goal of the study is to schedule and plan minimum cost paths for charging robots such that they rendezvous with and replenish the UAVs during the mission.

Work [13] presents the autonomous battery exchange operation for small scale UAVs, using a mobile ground base that carries a robotic arm and a service station containing the battery exchange mechanism. The goal of this work is to demonstrate the means to increase the autonomy and persistence of robotic systems without requiring human intervention.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Shin et al. [14] design an auction-based mechanism to control the charging schedule in multi-UAV setting in the presence of mobile charging stations. Charging time slots are auctioned, and their assignment is determined by a bidding process. The main challenge in developing this framework is the lack of prior knowledge on the distribution of the number of UAVs participating in the auction.

Paper [15] proposes the use of a ground-based refuelling vehicle (RV) to increase the operational range of a UAV in both spatial and temporal domains. A two-stage strategy for coupled route planning for UAV and RV to perform a coverage mission is developed. The first stage computes a minimal set of refuelling sites that permit a feasible UAV route. In the second stage, multiple Mixed-Integer Linear Programming (MILP) formulations are developed to plan optimal routes for the UAV and the refuelling vehicle taking into account the feasible set of refuelling sites generated in stage one.

Yu et al. [16] present an algorithm for planning a tour for an energy-limited UAV and tours for the UGVs with determining the best locations to place stationary charging stations. The authors study three variants for charging: multiple stationary charging stations, single mobile charging station, and multiple mobile charging stations.

Seyedi et al. [17] address the problem of achieving persistent surveillance over an environment by using energy-constrained UAVs which are supported by mobile charging stations. Specifically, the trajectories of all vehicles and the charging schedule of UAVs are planned by the authors for minimizing the time between two consecutive visits to regions of interest in a partitioned environment.

Work [18] considers route optimization problems for UAVs, which act as a team when inspecting or supporting a given set of objects in the presence of alternative and dynamic depots (starting and/or landing sites) and resource constraints.

Gordan et al. [19] describe a general overview of unmanned aerial vehicles (UAVs) and their potentiality in several engineering applications.

Paper [20] presents a way to design and fabricate a UAV-based air monitoring system to monitor air pollutant emissions over an oil field.

Tmušić et al. [21] describe the environmental conditions, constraints, and variables that could possibly be explored from UAV platforms and offer protocols that can be applied under all scenarios.

Work [22] considers the reliability of UAVs in a delivery network to minimize expected loss of demand.

Work [23] considers route optimization problems for UAVs, which act as a team when inspecting or supporting a given set of objects in the presence of alternative and dynamic depots (starting and/or landing sites) and resource constraints.

Hence, most of the considered works propose approaches to designing UAV-based monitoring systems without taking into account failures of UAVs.

Works [24, 25] describe the reliability models of UAV fleets, but the models don't take into account the specialties of designing monitoring systems and planning their use to ensure the specified requirements.

C. Aim and Objectives

The aim of the work is to develop route-based reliability models to improve multi-UAV routing for NPP monitoring missions.

The objectives of the paper are:

- to form the main requirements for a UAV fleet used as a mobile part of the NPP monitoring system;
- to develop a set of route-based reliability models of the UAV fleet carrying out monitoring of NPP and comprising main and redundant UAVs;
- to give an example of the proposed models application for UAV routing to cover the whole target MSs of the Zaporizhzhia NPP.

II. DEVELOPMENT OF TASKS OF UAV FLEET MISSION PLANNING CONSIDERING FAILURES OF UAVS

A. Ways to use UAV fleet for critical infrustucture monitoring

There are two main ways to use UAV fleet for critical infrastructure monitoring:

- the first way is to use UAVs as a mobile parts of a monitoring system responsible for collecting data from MSs (Fig.1);
- the second way is to use UAVs for forming a UAV-enable wireless network of the monitoring system for transmitting data from an MS to the CrS (Fig. 2).

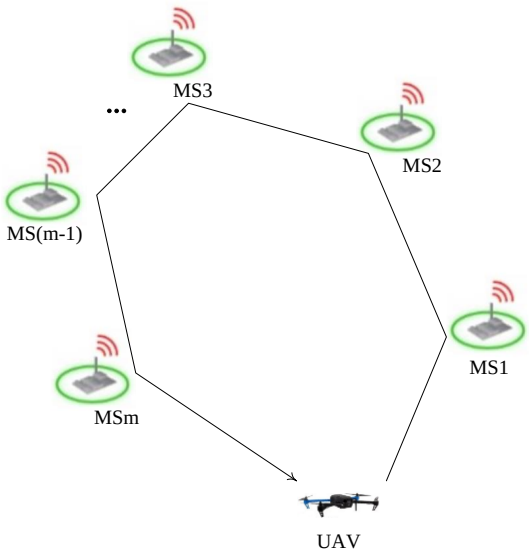


Figure 1. An UAV as a mobile part of the monitoring system responsible for collecting data from MSs

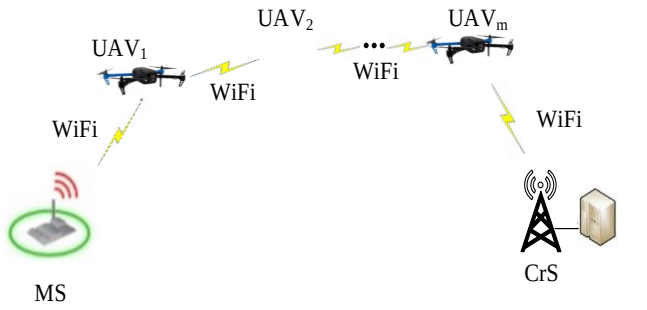


Figure 2. An example of a UAV-enable wireless network allowing the monitoring system to transmit data from the MS to the CrS

In this work, the first way of the UAV utilization is considered.

B. *The main requirements for UAV fleet used as a mobile part of the NPP monitoring system*

Using UAV fleet as a mobile part of the NPP monitoring system it is necessary:

- to ensure the maximum effectiveness in carrying out monitoring of the NPP with the given resources;
- to determine the minimum resources to ensure the required effectiveness in carrying out monitoring of the NPP.

In our case, resources are the number of UAVs, and an indicator is the probability of the SFP for the UAV fleet to cover the whole target MSs of the NPP.

A number of additional constraints can be given when UAV fleet mission planning. For example, the required time for collecting the data from an MS when visiting it.

Assume that UAV fleet monitoring mission planning involves distributing MSs into groups and giving a dedicated main UAV of the fleet to each of these groups to visit their MSs.

As a UAV of the fleet, moving along a route section, can fail, it is necessary to have redundant UAVs, the number of them is determined by the required value of the probability of SFP.

In this paper, two options for deployment of the redundant UAVs (RDs) are considered:

- the first one envisages deployment of RDs with the main UAVs (MDs);
- the second one envisages deployment of RDs on the depot which is located separately from the MDs.

A general route-based model of the UAV fleet used as a mobile part of the NPP monitoring system is shown in Fig. 3 where:

- MS_{i_ki} is MS k of MS group i where $i = 1,\ldots,n$, $fi = 1,\ldots,mi$;
- $L_{i\_(fi-1),fi}$ is the length of the route section from point $(fi\text{-}1)$ to point fi. For example, $L_{i\_0,1}$ is the route section from the depot of the UAV of MSs group i to MS_{i_1}, and $L_{i\_1,2}$ is the length of the route section from MS_{i_1} to MS_{i_2}.

It is vital to note that the number of groups of MSs (flight routes), is equal to the number of the main UAVs (N_{MD}) of the UAV fleet.

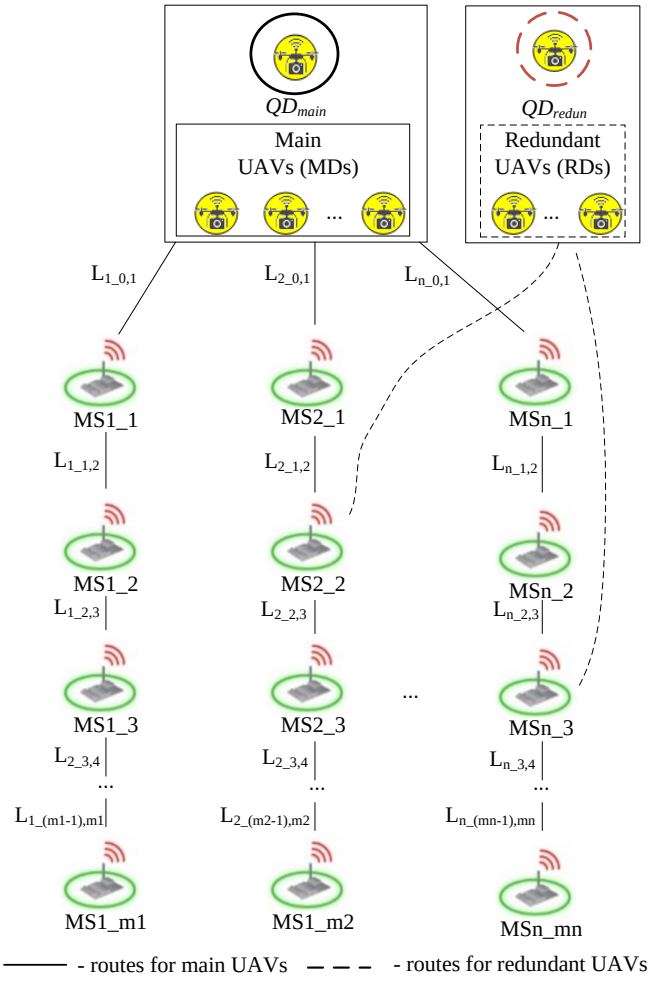


Figure 3. A general route-based reliability model of the UAV fleet used as a mobile part of the NPP monitoring system

Depending on the option to deploy the RDs, various approaches are used to calculate the probability of the SFP, which will be discussed below.

III. ROUTE-BASED RELIABILITY MODELS OF THE UAV FLEET

A. *Classification of route-based reliability models of the UAV fleet*

Classification of route-based reliability models, taking into account various scenarios for multi-MD/RD routing, the number of the MDs/RDs, and reliability of the MD/RD, is shown in Fig. 4. As we can see from Fig.4, the models can be described by data tuple S(n[m],k) where n is the number of

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

main routes (the number of MDs (N_{MD}) as well), *m* is the number of route section, and *k* is the number of redundant UAVs (N_{RD}).

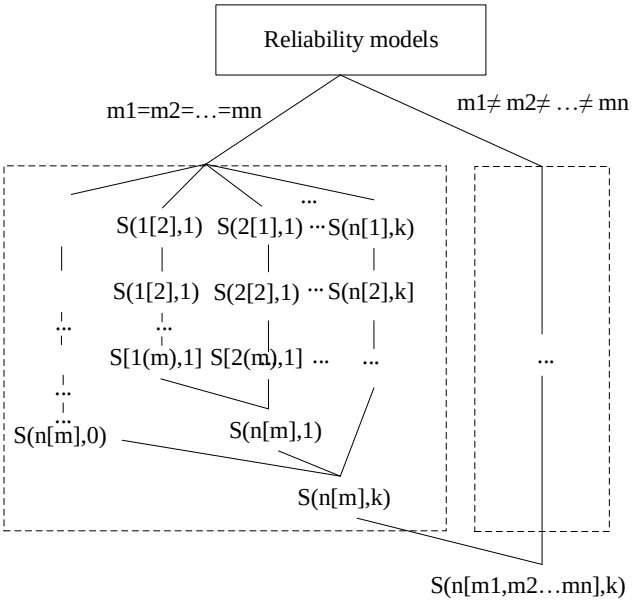


Figure 4. Classification of route-based reliability models of the UAV fleet

B. *Model S (2[2],1)*

A model describing utilization of 2 MDs and 1 RD for 2 routes with 2 MSs in each one is shown in Fig. 5.

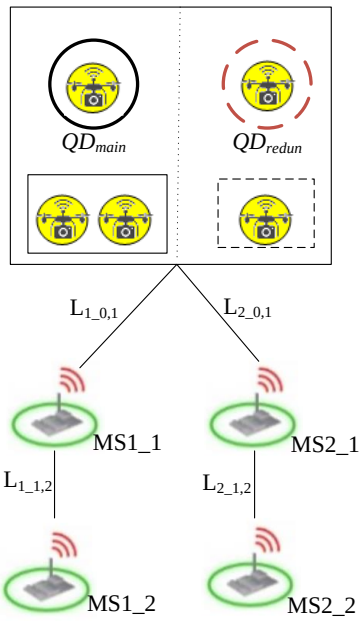


Figure 5. Graphical presentation of model S(2[2],1)

Let us p_{MD} and p_{RD} is the the reliability function of an MD and RD, respectively, and the reliability function is constant during the flight.

Assume that $L_{1\_0,1} = L_{1\_1,2} = L_{2\_0,1} = L_{2\_1,2}$, and $p_{RD}=1$.

In this case, the probability of the SFP is defined as:

$$P_{SPF(S(2[2],1))} = \left(p_{MD}^{\,2} + 2p_{MD}(1-p_{MD})p_{MD}\right)p_{MD}^{\,2} + \\ + p_{MD}^{\,2}\,2p_{MD}(1-p_{MD})\,p_{MD}^{\,2} = \\ = p_{MD}^{\,4} + 2p_{MD}^{\,4} + 2p_{MD}^{\,5} - 2p_{MD}^{\,6} = 3p_{MD}^{\,4} - 2p_{MD}^{\,6}. \quad (5)$$

If a UAV should return to its depot, the probability of the SFP is defined as:

$$P_{SPF}\left(S\left(2[2],1\right)\right) = \left(p_{MD}^{\,2} + 2p_{MD}(1-p_{MD})p_{MD}\right)p_{MD}^{\,2} + \\ + p_{MD}^{\,2}\,2p_{MD}(1-p_{MD})\,p_{MD}^{\,2}p_{ret}^{4} = \left(3p_{MD}^{\,4} - 2p_{MD}^{\,6}\right)p_{ret}^{4} \quad (6)$$

where p_{ret} is the probability of the UAV successful return to the depot.

For $p_{MD} = 0.9$ $P_{SPF}\left(S\left(2[2],1\right)\right) = 0.9054.$

If redundant UAVs deploy separately from main UAVs and $p_{RD} < 1$, the probability of the SFP is defined as:

$$P_{SPF}\left(S\left(2[2],1\right)\right) = \left(p_{MD}^{\,2} + 2p_{MD}(1-p_{MD})p_{MD}p_{RD}\right)\times \\ \times p_{MD}^{\,2} + p_{MD}^{\,2}\,2p_{MD}(1-p_{MD})\,p_{RD}\,p_{MD}^{\,2} = \\ = p_{MD}^{\,4}\left(3 - 2p_{MD}^{\,2}p_{RD}\right). \quad (7)$$

C. *Model S (2[2],2)*

A model describing utilization of 2 MDs and 2 RDs for 2 routes with 2 MSs in each one is shown in Fig. 6.

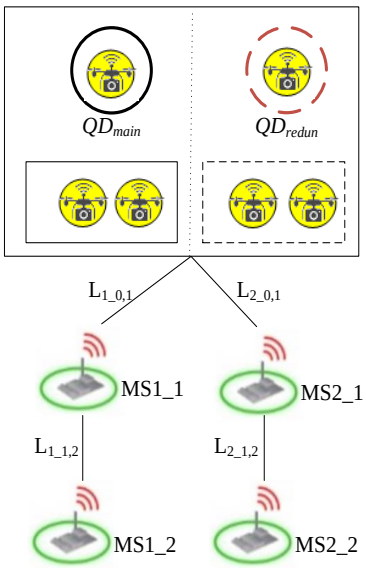


Figure 6. Graphical presentation of model S(2[2],2)

If $p_{RD}=1$, the probability of successful plan fulfilment is defined as:

$$P_{SPF}(S(2[2],2) = p_{MD}{}^4 + 2p_{MD}(1-p_{MD})p_{MD}p_{MD}{}^2 + *p_{MD}{}^2 2p_{MD}(1-p_{MD})p_{MD}p_{MD} + +2p_{MD}(1-p_{MD})p_{MD}p_{MD}(1-p_{MD})p_{MD}{}^2 = p_{MD}{}^4 + 2p_{MD}{}^4(1-p_{MD}) + 2p_{MD}{}^5(1-p_{MD}) + +2p_{MD}{}^5(1-p_{MD})^2 = = p_{MD}4\left(1+2(1-p_{MD})(1+p_{MD}+p_{MD}(1-p_{MD})\right). \quad (8)$$

If $p_{RD} < 1$, the probability of the SFP is defined as:

$$P_{SPF}\left(S(2[2],2)\right) = p_{MD}{}^4 + 2p_{MD}(1-p_{MD})p_{MD}p_{MD}{}^2p_{RD} + +p_{MD}{}^2 2p_{MD}(1-p_{MD})p_{MD}p_{MD}p_{RD} + +2p_{MD}(1-p_{MD})p_{MD}p_{MD}(1-p_{MD})p_{MD}{}^2p_{RD}{}^2 = = p_{MD}{}^4 + 2p_{MD}{}^4(1-p_{MD})p_{RD} + 2p_{MD}{}^5(1-p_{MD})p_{RD} + +2p_{MD}{}^5(1-p_{MD})^2 p_{RD}{}^2 = = p_{MD}4\left(1+2p_{RD}(1-p_{MD})(1+p_{MD}+p_{MD}p_{RD}(1-p_{MD})\right). \quad (9)$$

D. Model S (n[m],2)

The probability of the SFP for a model of utilization of *n* MDs and 1 RD for *n* routes with *m* MSs in each one can be defined as ($p_{RD} = 1$):

$$P_{SPF}\left(S(n[m],1)\right) = p_{MD}{}^{nm} + np_{MD}{}^{(n-1)}(1-p_{MD})p_{MD}p_{MD}{}^{n(m-1)} + +np_{MD}{}^n p_{MD}{}^{(n-1)}(1-p_{MD})p_{MD}{}^2 p_{MD}{}^{n(m-2)} + +np_{MD}{}^{n(m-2)}p_{MD}{}^{(n-1)}(1-p_{MD})p_{MD}{}^n p_{MD}{}^{(m-2)} + \ldots + +np_{MD}{}^{n(m-1)}p_{MD}{}^{(n-1)}(1-p_{MD})p_{MD}p_{MD}{}^{(m-1)}. \quad (10)$$

E. Model S(2(m1,m2),1)

A model describing utilization of 2 MDs and 1 RD for 2 routes with m1/m2 MSs (m1 ≠ m2) in the first/second route is shown in Fig. 7.

The probability of SPF when using 2 MDs and 1 RD for 2 routes with m_1/m_2 MSs (m1 ≠ m2) in the first/second route is defined as ($p_{RD} = 1$):

$$P_{SPF}(2[m1,m2],1) = P_{SPF}(2[m1],1) + p_{MD}{}^{2m1} + +p_{MD}{}^{2m1}\left(p_{MD}{}^{(m2-m1)} + (m2-m1)(1-p_{MD})p_{MD}{}^{m2}\right). \quad (11)$$

F. Research of the models

Let us calculate the probability of the SFP for a UAV fleet covering 4 target MSs to collect information from them using one of the following models: (S(1[4],1, S(2[2],1), and S(2[2],2). Note that models S(2[2],1) and S(2[2],2) can be used when the route is divided into two parts (Fig. 8).

Assume that $L_{1\_0,1} = L_{1\_1,2} = L_{2\_0,1} = L_{2\_1,2}$, p_{MD} =0.9, and p_{RD} = 1. Results obtained are shown in Fig. 9.

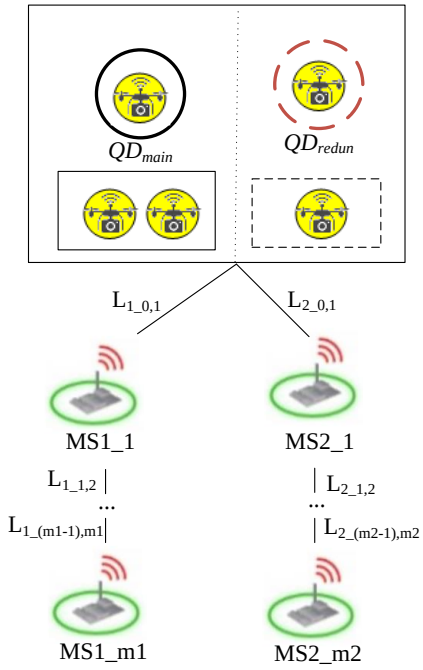


Figure 7. Graphical presentation of model S(2[m_1,m_2],1)

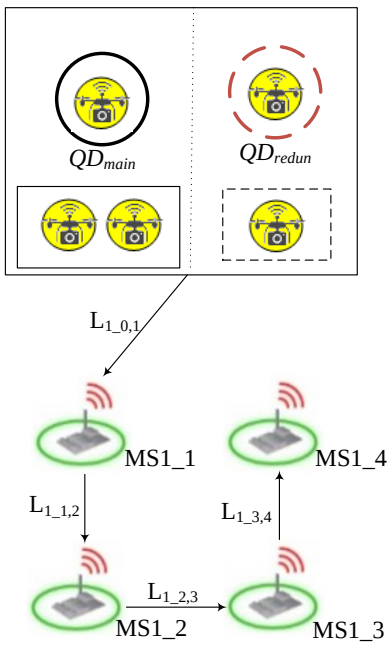


Figure 8. Graphical presentation of model S(1[4],1)

The results show that model S(2[2],2) should be chosen among the presented models as it provides the best value of the probability of the SFP. According to this model, 2 RDs should be used instead of 1 RD in models S(1[4],1) and S(2[2],1)).

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

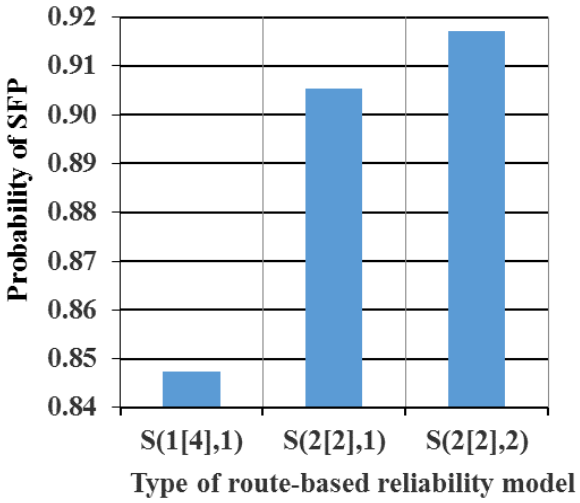


Figure 9. Diagram of the probability of SPF when using models S(1[4],1), S(2[2],1), and S(2[2],2)

IV. CASE STUDY

A. Mission statement

Let us have the following scenario. As a result of Zaporizhzhia NPP accident, the wired networks connecting 4 MSs (MS10, MS15, MS17, MS18) with the CrS were damaged (Fig. 10).

It is necessary to carry out the monitoring mission on visiting these MSs via UAVs to collect radiation monitoring data from them using the Wi-Fi equipment placed both at a MS and the on-board.

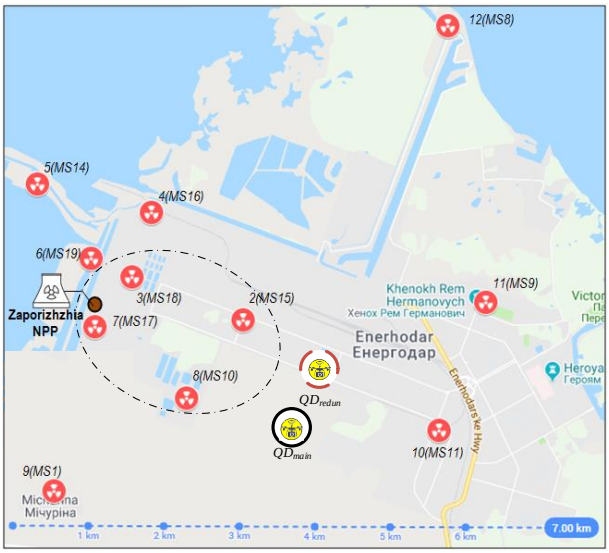


Figure 10. Target MSs and the depot for UAVs

For example, the WiFi equipment of the DJI Mavic 2 Enterprise Dual quadcopter (hereinafter the quadcopter (QD)) allows performing this mission due to communications between the MS and the QD hovering on the MS during the needed time. An MS is a point of the route for the QD to visit. The main (QD_{main}) and redundant (QD_{red}) quadcopters are deployed in the same depot.

B. Solution

Assume that the reliability function of the MD when using model S(n[m],k) can be calculated as:

$$p_{S(n[m],k)} = e^{-\lambda t_{S(S(n[m],k)}} \quad (11)$$

where λ is the MD failure rate.

Let us have the following initial data: $L_{0\_15} = 1.00$; $L_{15\_18} = 1.59$; $L_{18\_17} = 0.85$; $L_{17\_10} = 1.56$; $L_{18\_0} = 2.34$; $L_{17\_0} = 2.31$; $L_{0\_10} = 1.00$; $\lambda = 0.5$ 1/h; $V_{MD} = V_{RD} = 40$ km/h; $p_{MDS(n[m],k)} = 1$.

Let us have the following assumptions.

- The UAV visits each target MS during 2 minutes.
- The RD has to use the route assigned to the MD if the last fails;
- Only one RD is used in each scenario on UAV route planning;
- It is necessary to provide the probability of the SFP ≤ 0.92.

Let us determine the values of the probability of the SFP for UAV fleet consisting of 1 MD and 1 RD (Fig. 11).

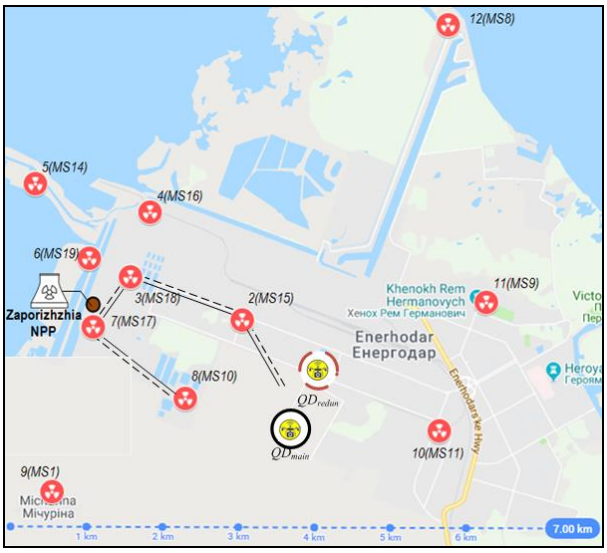


Figure 11. Utilization of model S(1[4],1) for UAV routing to cover 4 target MSs of the Zaporizhzhia NPP

In this case, the route comprises four sections (model S(1[4],1)) and its total length is

$$L_{S(1[4],1)} = L_{0\_15} + L_{15\_18} + L_{18\_17} + L_{17\_10} = 5.00 \text{ km}.$$

The route flight time can be defined as:

$$t_{S(1(4),1)} = \frac{L_{S(1[4],1)}}{V} = \frac{5.00}{40} = 0.125 \text{ h.}$$

The reliability function of the main UAV is

$$p_{MD_{S(1(5),1)}} = e^{-\lambda t_{S(1[4],1)}} = e^{-0.5 \cdot 0.125} = 0.9277 .$$

The probability of SFP is

$$P_{SPF}(1[4],1) = {p_{MD}}^4 + 4 p_{MD} (1 - p_{MD}) {p_{MD}}^3 = 0.911.$$

Let us divide the one route into two parts:

$$L_{0\_15} \to L_{15\_18} \text{ and } L_{0\_10} \to L_{10\_17}.$$

In this case, we can use model (2[2],1) (Fig. 12).

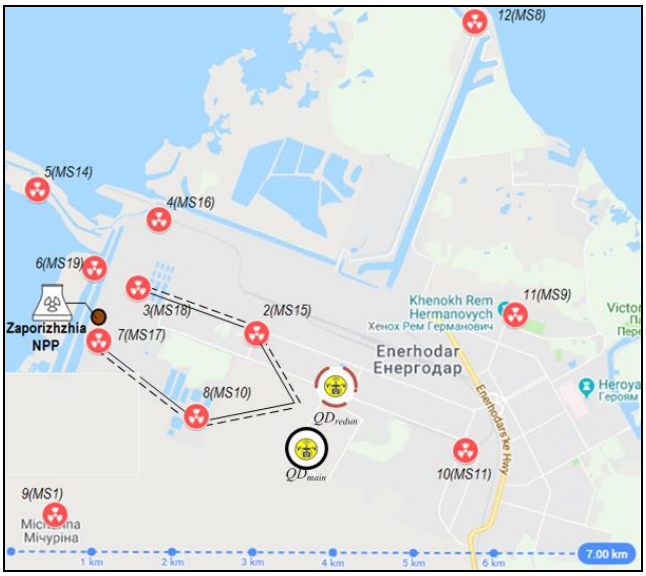


Figure 12. Utilization of model S(2[2],1) for the UAV routing to cover 4 target MSs of the Zaporizhzhia NPP

Let us introduce the assumption that we have two equal route sections.

In this case, the route length is

$$L_{S(2[2],1)} = L_{0\_15} + L_{15\_18} = 2{,}59 \text{ km.}$$

The route flight time is

$$t_{S(2[2],1)} = \frac{L_{S(2[2],1)}}{V} = \frac{2.59}{40} = 0.065 \text{ h.}$$

The reliability function of the MD is

$$p_{MD_{S(2[2],1)}} = e^{-\lambda t_{S(2[2],1)}} = e^{-0.6 \cdot 0.065} = 0.9617 .$$

The probability of the SFP is

$$P_{SPF}(2[2],1) = {p_{MD}}^4 \left({p_{MD}}^2 + 2(1 - p_{MD}) \left(1 + p_{MD} + {p_{MD}}^2\right)\right) = 0.9601.$$

Thus, model S(1[4],1) does not provide the required probability of SFP. Using model S(2[2],1)), which supposes to divide the route into two parts, allows meeting the requirements. 3 UAVs (2 MD and 1 RD) should be utilized instead of 2 UAVs (1 MD and 1 RD) when using model S(1[4],1).

V. Conclusion

The paper proposes a set of route-based reliability models of the UAV fleet carrying out monitoring of NPP and comprising main and redundant UAVs. Depending on the model used, approaches to calculation of the probability of the SFP for UAV fleet to cover the whole target MSs of the NPP taking into account failures of UAVs and utilization features of redundant UAVs are proposed.

An example of the proposed models application for routing of UAVs responsible for covering 4 target MSs of the Zaporizhzhia NPP is shown and examined.

Further research can be devoted to investigation of monitoring station elements (UAV fleet, CrS, communications etc.), considering them as multi-state systems [26].

Acknowledgment

This work is partly supported by the grant APVV SK-SRB-18-0002 of the Slovak Research and Development Agency

References

[1] R. Bielawski, W. Rządkowski, R. Perz. “Unmanned Aerial Vehicles in the protection of the elements of a country’s critical infrastructure – selected directions of development”. *Security and Defence Quarterly*, no. 5, pp. 3–19, 2018.

[2] A.L. Moore. “UAV surveillance of critical infrastructure, capability and constraints”. *Proc. of the 49th Session of the International Seminars on Nuclear War and Planetary Emergencies held in Erice*, Sicily, 2019.

[3] A. Sachenko, V. Kochan, V. Kharchenko, H. Roth, V. Yatskiv, M. Chernyshov, V. Bykovyy, O. Roshchupkin, V. Koval, H. Fesenko et. al Mobile post-emergency monitoring system for nuclear power plants. *CEUR Workshop. In Proc. 12th International Conference on ICT in Education, Research and Industrial Applications*, ICTERI’2016, Kyiv, Ukraine, vol. 1614. pp. 384-398.

[4] D. Connora, P. G. Martin, and T. B. Scott. “Airborne radiation mapping: overview and application of current and future aerial systems”. Int. J. of Remote Sensing, vol. 37, no. 24, pp. 5953–5987, Nov. 2016.

[5] V. Burtniak, Y. Zabulonov, M. Stokolos, L. Bulavin, V. Krasnoholovets. “Application of a territorial remote radiation monitoring system at the Chornobyl nuclear accident site”. *J. of Applied Remote Sensing*, no. 12(4), pp. 1-13, 2018.

[6] J. Lüley, B.Vrban, Š. Čerba, F. Osuský, V Nečas. “Unmanned radiation monitoring system”. EPJ Web of Conferences, 2020, p. 225.

[7] J. Aleotti, G. Miccon, S. Caselli, G. Benassi, N. Zambelli. “Detection of nuclear sources by UAV teleoperation using a visuo-haptic augmented reality interface”. *Sensors*, no. 17(10), 2017, p. 22-34.

[8] P. Xiao, B. Tang, X. Huang, P. Wang, L. Sheng, W. Xiao, X. Zhu, C. Zhou. “Locating lost radioactive sources using a UAV radiation monitoring system”. *Applied Radiation and Isotopes*, vol. 150, 2019, pp. 1–13.

[9] P.G. Martin, O.D. Payton, J.S. Fardoulis, D.A Richards, Y. Yamashiki, T.B. Scott. Low altitude unmanned aerial vehicle for characterising remediation effectiveness following the FDNPP accident”. *Environmental Radioactivity*, vol. 151, pp. 58–63, 2017.

[10] H. Fesenko, V. Kharchenko, A. Sachenko, R. Hiromoto, and V. Kochan, “An Internet of Drone-based Multi-version Post-severe Accident Monitoring System: Structures and Reliability,” in *Dependable IoT for*

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Human and Industry Modeling, Architecting, Implementation, V. Kharchenko, A. Kor, A. Rucinski, Eds. Denmark, The Netherlands: River Publishers, 2018, pp. 197-217.

[11] H. Fesenko, I. Kliushnikov, V. Kharchenko, S. Rudakov, E. Odarushchenko. "Routing an Unmanned Aerial Vehicle During NPP Monitoring in the Presence of an Automatic Battery Replacement Aerial System". *In Proc. of the 11th IEEE Int. Conf. Dependable Systems, Services and Technologies*, Kyiv, Ukraine, pp. 34-39, 2020.

[12] N. Mathew, S. L. Smith, S. L. Waslander. "Multirobot rendezvous planning for recharging in persistent tasks". *IEEE Trans. Robot*, vol. 31, no. 1, 2015, pp. 128–142.

[13] E. Barrett, M. Reiling, S. Mirhassani, R. Meijering, J. Jager, N. Mimmo, Callegati F., Marconi L., Carloni R., Stramigioli S. "Autonomous battery exchange of UAVs with a mobile ground base". *In Proc. IEEE Int. Conf. Robotics and Automation*, Brisbane, Australia, 2018, pp. 699–705.

[14] M. Shin, J. Kim, M. Levorato. Auction-based charging scheduling with deep learning framework for multi-UAV networks. *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4235–4248, 2019.

[15] P. Maini, K. Sundar, M. Singh, S. Rathinam, , P.B. Sujit. "Cooperative aerial-ground vehicle route planning with fuel constraints for coverage applications". *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, 2019, pp. 3016–3028.

[16] K. Yu, A.K. Budhiraja, S. Buebel, P. Tokekar. "Algorithms and experiments on routing of unmanned aerial vehicles with mobile recharging stations". *J. Field Robotics*, vol. 36, 2019, pp. 602–616.

[17] S. Seyedi, Y. Yazicioğlu, D. Aksaray. "Persistent surveillance with energy-constrained UAVs and mobile charging stations". *IFAC-PapersOnLine*, vol. 52, 2019, pp. 193–198.

[18] V.P. Horbulin, L.F. Hulianytskyi, I.V. Sergienko. "Optimization of UAV Team Routes in the Presence of Alternative and Dynamic Depots". *Cybern Syst Analysis*, no. 56, pp. 195–203, 2020.

[19] D. Giordan, M.S. Adams, I. Aicardi. et al. "The use of unmanned aerial vehicles for engineering geology applications". *Bull Eng Geol Environ*, 79, 2020, pp. 3437–3481.

[20] L. Siwen. "Development of a UAV-Based System to Monitor Air Quality over an Oil Field". *Graduate Theses & Non-Theses*, 2018, p. 187-198.

[21] G. Tmušić, S. Manfreda, H. Aasen, M.R. James. at al. "Current Practices in UAS-based Environmental Monitoring". *Remote Sens*, vol. 12, 2020, p. 1001-1018.

[22] M. Torabbeigi, G.J. Lim, S.J. Kim. "UAV delivery schedule optimization considering the reliability of UAVs". *In Proc. Of the International Conference on Unmanned Aircraft Systems* (ICUAS), Dallas, TX, USA, 2018, pp. 1048-1053.

[23] I. Kliushnikov, H. Fesenko, V. Kharchenko. "Scheduling UAV fleets for the persistent operation of UAV-enabled wireless networks during NPP monitoring". *Radioelectronic and Computer Systems*, no. 1, 2020, pp. 29-36.

[24] H. Fesenko, V Kharchenko, E. Zaitseva. "Evaluating reliability of a multi-fleet with a redundant UAV fleet: an approach and basic model". *In Proc. of the IEEE International Conference Information and Digital Technologies*, Zilina, Slovakia, 2019, pp. 128-132.

[25] H. Fesenko, V. Kharchenko. "Reliability models for a multi-fleet of UAVs with two-level hot standby redundancy considering a control system structure". *In Proc. of the 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technologies and Application*, Metz, France, 2019, pp. 1030-1035.

[26] M. Kvassay, P. Rusnak, E. Zaitseva, J. Kostolny. "Minimal Cut Vectors of Multi-State Systems Identified Using Logic Differential Calculus and Multi-Valued Decision Diagrams". *In Proc. of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, Venice, Italy. 2020, pp. 3053-3060.

Describing rumours: a comparative evaluation of two handcrafted representations for rumour detection

Luisa Francini, Paolo Soda, Rosa Sicilia
Unit of Computer Systems and Bioinformatics, Department of Engineering
Università Campus Bio-Medico di Roma, Italy
{l.francini,p.soda,r.sicilia,}@unicampus.it

***Abstract*—Nowadays, people use more and more social media as a source of information, leading to an increased and uncontrolled spread of misinformation. For this reason, tools to detect unverified and instrumentally relevant news, named as *rumours*, are necessary. In this work we compare two state-of-the-art handcrafted representations, namely User-Network and Social-Content, designed for developing machine learning-based rumour detection systems, in order to analyse which descriptors best capture the information hidden in unknown rumours. To this end we set up an experimental assessment implementing a Leave-One-Topic-Out evaluation on 8 different topics retrieved from Twitter social microblog. The results obtained for both representations are low as we designed a simple and non optimised pipeline for a fair comparison. Besides this, we were able to find out that the User-Network set of feature results more stable to topic changes. As a further contribution, we introduce two new datasets labelled for rumour detection task on Twitter.**

***Index Terms*—Rumour detection; Feature Representation; Social-Microblog; Twitter**

I. Introduction

These days, an increasing number of users inform themselves via social media instead of traditional news sources [1]. Social media can be accessed by everyone, so even ordinary citizens can report events as well as their own feelings and experiences, but the absence of systematic control of posts on these platforms may lead to spread misinformation, especially in the context of breaking news, which often appear first on social media and only after on traditional media systems [2]. Among the different kinds of misinformation, a rumour is defined as an unverified news in circulation with an instrumental value and likely to be dangerous [3]. Twitter is a platform that, due to how it allows posts to be created, facilitates the real-time dissemination of news, which causes rumour to spread rapidly with many consequences.

For this reason, the interest in the study of machine learning and deep learning methods for rumour detection has been growing in recent years. Most of this work is related to *macro-level* rumour detection, which means that the detection system considers as rumour news carried by a set of microblog posts, aggregated according to the same conversational thread [4], a specific topic [5], [6], a particular event [7], [8] or their content and enquiry level [9]. In contrast, a *micro-level* approach could be more useful and efficient in many applications: it aims to discriminate among the single posts which one are rumours from those that are not, even if they belong to the same conversation thread, topic, etc. This type of study is more effective since in many cases it is possible to find that the subject of a conversation may not be a rumour, but actually several of the posts in the conversation are. Hence, we focus on this second rumour detection task, which is more challenging and less tackled in the literature: indeed there are few models and approaches that cope with this problem, which exploit both machine learning [10], [11], [12], [13], and deep learning [14], [15].

Although deep learning approaches are more and more gaining the scene, some of them do not rely only on a deep representation but still employ handcrafted descriptors to represent some characteristics of this multi-modal problem, e.g. related to text or the user [14].

In this context, our work aims at analysing two well-established handcrafted representations [10], [11] that model different properties of the multi-modal problem of rumour detection on Twitter. In detail we offer two main contributions: (i) we compare the two state-of-the-art representations considering different machine learning classifiers over a pool of 8 Twitter datasets for micro-level rumour detection in a binary task, rumours vs. non-rumours; (ii) we present two novel datasets retrieved from Twitter and manually labelled with a robust methodology.

The rest of the manuscript is organized as follows: next section describes the literature on machine and deep learning methods for micro-level rumour detection, highlighting the type of feature used, section III presents the datasets used in this study, section IV describes the two handcrafted representations and section V how the experiments were conducted. Finally, section VI analyses the results and VII provides concluding remarks.

II. Background

As mentioned in section I, macro-level rumour detection considers as rumour news carried by a set of microblog posts, while micro-level rumour detection focuses on identifying rumours among single posts and, for this reason, it could be more useful and challenging. We report below recent work in this second field, both using machine learning [11], [12], [13], and deep learning [14], [15]. All of this work tackles the

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE I: Number of tweet contained in the various datasets, showing also how many tweet originated the conversation and how many reactions they generated.

Topic	# of Source Tweets			# of Reactions	# of Tweets
	# of Rumours	*# of Non-Rumours*	*Tot*		
Charlie Hebdo	458	1621	2079	36187	38266
Ferguson	284	859	1143	22260	23403
Germanwings	238	231	469	4020	4489
Ottawa Shooting	470	420	890	11394	12284
Sydney Siege	522	699	1221	22775	23996
RoyalWedding	53	134	187	277	464
TrumpRussia	122	96	218	370	588
Vaccine	91	168	259	699	958

binary task classifying posts in rumour and non-rumour, but using different features. In our previous work [10], we also studied the problem of representing social microblog data for micro-level rumour detection, introducing newly designed user and network level features that allow to correctly identify 90% of rumours. The test was made on two health topic datasets collected from Twitter and, differently, from the rest of the following literature we addressed a 3-class problem (rumour, non-rumour and unknown). Zubiaga et al. [11] exploited two types of features, i.e. content-based features and social features, and tested them both individually and in combination. They tested different classifiers on the public dataset PHEME [16], finding that the best one is the Conditional Random Fields (CRF), which recognises 56% of rumours. PHEME is a public collection of tweets divided in 5 topics and it will be fully described in section III. Also in [12] the authors perform all experiments with PHEME, extracting a new type of social context features, i.e. *Rumor spread*, *Average rumor speed*, *Which day*, *Which month* and *Working days*. They combine this new set of descriptors with common features suggested by [17] and they adopt an ensemble-voting classifier technique to identify rumours and non-rumours. The outcomes show that this approach leads to better results compared to known state-of-the-art systems, with an accuracy up to 78%. In [13], Pratiwi et al. use a combination of user and tweet based feature with TF-IDF (term frequency-inverse document frequency) as a method of feature selection. They collect 47,449 records from the Indonesian-language Twitter. They determine that a Support Vector Machine (SVM) classifier combined with TF-IDF feature selection method improves accuracy up to 76%.

With reference to the work employing deep learning techniques, in [14] the authors propose an architecture that leverages 27 handcrafted features, describing both tweet and user information. With a multiple-layer stack attention mechanism they filter noise or unnecessary information. Their main objective is to address the early rumour detection problem, so they propose an hybrid neural network combining a task-specific character-based bidirectional language model and stacked Long Short-Term Memory networks (LSTM). The experiments performed on PHEME dataset result in an accuracy of 86%. In [15] Ashgar et al. use only text-based features and propose a system which combines a Bidirectional LSTM (BiLSTM) with a Convolutional Neural Network. Their experiments on PHEME reach an accuracy of 86%.

The work presented so far on micro-level rumour detection shows the interest of the research community on developing effective handcrafted descriptors catching multi-modal aspects of this problem, e.g. contents, users, network, etc. Although the deep learning approaches are more and more gaining the scene, these studies are relatively new, so there is no well established deep representation method designed for this specific task. Moreover, the work exploiting deep learning still uses handcrafted features to describe text or user characteristics [14], for these reasons, in the following we compare, as described above, the two well-established handcrafted representations presented in [10] and [11], which model different properties of the multi-modal problem of rumour detection on Twitter, leaving the deep representation analysis to future work.

III. Materials

In this work we use two different collections of data: the first one is named as *PHEME*, which contains Twitter posts from 5 breaking news, the second one is named as *UCBM* collection, containing 3 topics.

The five topics in PHEME are [11]:

- Charlie Hebdo shooting (CH): two brothers forced their way into the offices of the French satirical weekly newspaper Charlie Hebdo, killing 11 people and wounding 11 more, on January 7, 2015. In this dataset 22% of samples are rumours.
- Ferguson unrest (FU): citizens of Ferguson in Missouri, USA, protested after the fatal shooting of an 18-year-old African American, Michael Brown, by a white police officer on August 9, 2014. The final samples account for 25% of rumours and 75% of non-rumours.
- Germanwings plane crash (GW): a passenger plane from Barcelona to Düsseldorf crashed in the French Alps on March 24, 2015, killing all passengers and crew. The plane was ultimately found to have been deliberately crashed by the co-pilot. About 51% of samples are rumours.
- Ottawa shooting (OS): shootings occurred in Ottawa's Parliament Hill, resulting in the death of a Canadian soldier on October 22, 2014. In this dataset there are 470 rumours, which means that about 53% of the total set is assigned to that class.

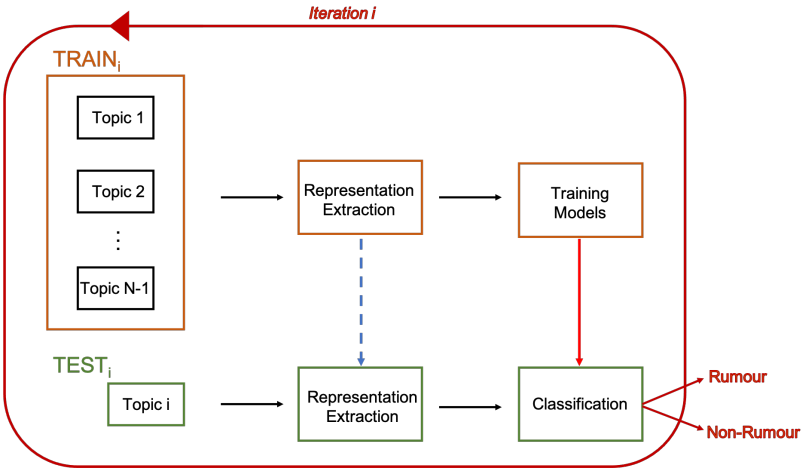


Fig. 1: Pipeline of the representation evaluation process. The blue dashed arrow refers to eventual information extracted on the training set that is needed to represent the test set, such as Word2Vec models learned on the training set.

- Sydney siege (SS): a gunman held hostage ten customers and eight employees of a Lindt chocolate cafe located at Martin Place in Sydney on December 15, 2014. This dataset is composed of 43% of rumours.

The three datasets in UCBM collection were acquired in June 2018. The labelling process was carried out by three human annotators and their labels were then used to build a gold standard, computed as the majority voting on the complete set of labels of annotators involved in the study. The included topics are:

- Vaccine (V): the tweets in this dataset are correlated to the health domain, in particular to the theme of vaccines. The final dataset accounts for 958 samples with 29% of rumours and 71% of non-rumours.
- RoyalWedding (RW): in May 2018 Prince Harry and Meghan Markle celebrated their marriage. The 464 tweets in this topic are related to this event and 29% of them are rumours.
- TrumpRussia (TR): tweets in this dataset refer to the events after Trump's election in 2016 and the judicial enquiry arising from suspected Russian interference in the election campaign (Russiagate). It contains 588 tweets and 65% of them are rumours.

The Vaccine dataset has already been presented in a previous work [10] as a topic belonging to the health domain, while the last two are introduced here for the first time.

TABLE I summarizes the characteristics of each dataset including the number of rumour and non-rumours present in each topic. It is worth noting that in the case of UCBM collection both the tweet that generated the conversation (the so called source tweet) and the related reactions were labelled. With the term "reaction" we refer to users' interactions (retweets and replies) with a specific tweet, so together source tweet and reactions compose a conversation. Since we focus on a micro-level analysis, rumours should be discriminated also within a conversation, hence reactions should be labelled too for a supervised task. In PHEME collection this information is not provided and, although the source label could be extended to all its retweets since they have the same content, this does not hold for replies. Indeed, tagging all reactions in a thread with the same label may generate noise. Hence, to carry out a micro-level analysis on datasets with an homogeneous structure, in the following we consider only the source tweets excluding all the reactions, even when they are labelled.

IV. Methods

In this section we describe the methodology we adopt to compare the two representations [10], [11], represented in Fig.1. We can distinguish two main phases: training (train) and test. In each iteration *i*, one of the datasets is used as test set, whereas all the others are employed for training. In the training phase, first, the two representations are extracted from the *n-1* topics, then this information is used to learn the chosen classifiers. Likewise in the test phase, the features are extracted from the topic excluded in the previous step and then the samples are classified in rumour or non-rumour.

As mentioned in section I, our aim is to offer a comprehensive comparison of two handcrafted representations which have proven to be effective for rumour detection on Twitter. In the following, we briefly offer the reader a complete description of the two representations considered.

A. Data Representation and Feature extraction

The first feature set was designed in our previous work [10], and it captures information at user and network levels, so we refer to it as "User-Network" representation. The second one was proposed in [11] by Zubiaga et al. and it catches aspects related to content of the posts and social context of the user, being referred to as "Social-Content" representation.

1) **User-Network Representation** [10]: this representation consists of 24 descriptors. The user-level features identify the characteristics of the user with his/her statuses, whilst, the network-level descriptors identify the interactions between

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

users in the network. These two levels can be further divided into three different groups:

- Influence potential: it describes the capacity of causing an effect in indirect or intangible ways and it contains features of a user, as the number of followers and of followings, properties of a tweet, as the presence of question marks, and features which represent the probability that a tweet is retweeted and the probability of a URL to be shared.
- Personal interest: it is useful to understand the reaction of a user about a specific news and it is expressed by a sentiment score computation.
- Network characteristics: it catches the propagation structure of the retweets and replies conversation graphs, considering measures related to graph theory, such as PageRank, closeness and betweenness centrality, and other structural descriptors related to graphs depth (conversation size) and properties of the users and tweets in a conversation (the fraction of users followers of the root, fraction of tweet with an URL).

It is possible to find the formal definition of all the 24 descriptors in [10].

2) **Social-Content Representation** [11]: it is composed of 12 descriptors. The content-based features model the dissimilarities between the different contents generated by rumours and non-rumours. They include measures that catch emphasis through grammar and punctuation in the posts (e.g. part-of-speech and tags) and a representation of the words with vectors extracted using the Word2Vec network [18]. The social-based features capture the user's behaviour and are indicative of his/her experience and reputation in the network (e.g. number of statuses posted and number of lists a user belongs to).

It was possible to extract all the feature of both representation methods from the json files of the tweets. For social-content features, we built word vectors representation as reported in the original manuscript [11]. In detail we trained the Word2Vec model with 300 dimensions to generate the vocabulary on the training set and then we employ such model to extract the word vectors on the test topic (blue dashed arrow in Fig.1) so that the event (and the vocabulary) in the test set was unknown.

V. Experimental setup

Similar to Zubiaga's experiments [11], we divided the samples in train and test set by topic, i.e. all the datasets, except the test topic, are joined to train the models (as indicated in Fig.1). This is a kind of validation called *Leave-one-topic-out* and, to evaluate the two representations, we computed the average of performance scores over all test topics. We used as measures Accuracy, Precision, Recall and F_1-score which are formally defined in the following equations. The notations TP, TN, FP and FN indicate True Positive, True Negative, False Positive and False Negative, respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 - score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4}$$

We do not employ a classical k-fold cross-validation, since this implies a random split of samples, so that the training set might include samples that belong to topics in the test set. The resulting mixture of topics does not allow to model a real world environment where a classifier trained on a known set of topics might be challenged to discover rumours regarding unknown new topics.

We implement 3 different experiments in Leave-one-topic-out, one with the two collections unified and the other two considering only the datasets belonging to the same collection. These are shortly referred to as *Leave-One-Topic-Out merging the collections* and *Leave-One-Topic-Out within the collections*, as now detailed.

Leave-One-Topic-Out merging the collections: Firstly, we consider all 8 datasets unifying the two collections, so the training set contains all the samples belonging to 7 out of 8 topics, while the eighth topic is used as test set. With this experiment we intend to find out what happens to the discriminating power of a representation when we enlarge the training set with new topics.

Leave-One-Topic-Out within the collections: We also implement two tests, one using only the 5 news presented in PHEME and the other with the 3 topics in UCBM collection. Comparing these two with the first leave-one-topic-out experiment, we want to determine whether adding new topics causes performance improvements with respect to smaller collections. Moreover, looking at the differences between the two tests within the collections, we want to understand what happens when we test the same representation on data that carry different information: indeed the first collection contains breaking news, whilst the second one reports on political, medical and gossip events.

We select 5 classifiers to evaluate the two representations belong to different learning paradigms: a Decision Tree (DT), as a binary tree, a Support Vector Machine (SVM) with linear kernel, as a kernel machine, a Random Forest (RF) as an ensemble of trees, a Gaussian Bayesian (GB) as statistical paradigm and an XGBoost (XGB), which is based on a gradient boosting framework.

VI. Results

In this section, we present the analysis of the results obtained comparing the two representations. First, we compare the results of the Leave-one-topic-out merging the collections, reported in TABLE II, with the experiments in Leave-one-topic-out within the collections, reported in TABLE III. In each table the right panel reports the performance obtained with the social-content representation, whereas the left panel reports the performance with the user-network representation.

TABLE II: Results attained using the experimental procedure named as *Leave-One-Topic-Out merging the collections*.

Classifier	*User-Network representation*				*Social-Content representation*			
	Accuracy	**Precision**	**Recall**	**F_1-score**	**Accuracy**	**Precision**	**Recall**	**F_1-score**
DT	0.56±0.05	0.45±0.12	0.30±0.12	0.33±0.05	0.48±0.03	0.37±0.08	0.48±0.15	0.40±0.04
SVM	0.61±0.011	0.00±0.00	0,00±0.00	0.00±0.00	0.53±0.08	0.39±0.13	0.33±0.33	0.28±0.24
RF	0.59±0.05	0.51±0.21	0.28±0.15	0.32±0.08	0.49±0.07	0.39±0.10	0.58±0.20	0.45±0.12
GB	0.46±0.06	0.40±0.11	0.69±0.27	0.46±0.14	0.60±0.10	0.17±0.06	0.01±0.01	0.01±0.01
XGB	0.59±0.07	0.51±0.25	0.16±0.12	0.19±0.10	0.54±0.08	0.40±0.13	0.30±0.13	0.30±0.14

TABLE III: Results attained using the experimental procedure named as *Leave-One-Topic-Out within the collections*.

Collection	Classifier	*User-Network Representation*				*Social-Content Representation*			
		Accuracy	**Precision**	**Recall**	**F_1-score**	**Accuracy**	**Precision**	**Recall**	**F_1-score**
PHEME	**DT**	0.59±0.09	0.46±0.13	0.28±0.15	0.32±0.07	0.56±0.06	0.39±0.16	0.37±0.10	0.38±0.13
	SVM	0.61±0.14	0.00±0.00	0.00±0.00	0.00±0.00	0.58±0.11	0.38±0.15	0.1±0.09	0.16±0.07
	RF	0.60±0.09	0.47±0.17	0.23±0.10	0.39±0.09	0.57±0.07	0.38±0.16	0.36±0.18	0.36±0.17
	GB	0.51±0.06	0.38±0.12	0.30±0.25	0.25±0.16	0.40±0.14	0.38±0.16	0.90±0.06	0.52±0.16
	XGB	0.60±0.11	0.47±0.19	0.16±0.18	0.19±0.13	0.59±0.12	0.36±0.17	0.06±0.07	0,08±0.07
UCBM	**DT**	0.57±0.07	0.30±0.15	0.27±0.17	0.26±0.08	0.44±0.06	0.39±0.11	0.67±0.14	0.47±0.03
	SVM	0.59±0.14	0.15±0.00	0,25±0.00	0.19±0.00	0.54±0.13	0.36±0.13	0.33±0.44	0.27±0.23
	RF	0.54±0.06	0.54±0.25	0.55±0.29	0.43±0.13	0.58±0.03	0.30±0.16	0.40±0.19	0.34±0.15
	GB	0.47±0.03	0.41±0.14	0.43±0.41	0.34±0.14	0.59±0.13	0.20±0.08	0.07±0.01	0.10±0.02
	XGB	0.56±0.09	0.68±0.37	0.16±0.14	0.20±0.13	0.56±0.14	0.42±0.04	0.15±0.03	0.22±0.04

Results are summarized showing each performance measure computed averaging it across all leave-one-topic-out folders, so the mean on the test topics, with the related standard deviations. Comparing the two experiments (TABLE II against TABLE III), we can make the following considerations: on the one hand, increasing the dataset sample size, moving from Leave-One-Topic-Out within the collection to Leave-One-Topic-Out merging the collections, does not vary the performance of the User-Network representation, which seems stable across the experiments. On the other hand, the Social-Content representation shows an opposite behaviour with a negative effect on the accuracy: indeed augmenting the dataset sample size with new topics decreases the discriminative power of this feature set. This could be due to the difference between the kind of information conveyed by PHEME and UCBM topics; for example, among UCBM events there is the Vaccine dataset, which is about a medical field and naturally has a different vocabulary than the others. Hence, from this first comparison the User-Network feature set seems more stable to topic changes.

To get a deeper insight on this matter, Fig. 2 shows how the accuracy and F_1-score vary according to the "classifier"-"test topic" combination in each experiment. In these figures we use red for the experiments with the 8 datasets, blue for those only with PHEME collection and green for those with UCBM collection. The value of the performance is expressed by the magnitude of the circles radius: for example, in Fig. 2a considering the "classifier"-"test topic" pair Decision Tree-Vaccine, we draw a green circle to refer to the experiments performed only the UCBM collection, and a red circle to represent those will both the collections. The fact that red circle has a radius larger than the green one implies that the accuracy is reduced when the experiments are carried out according to the Leave-one-Topic-Out within UCBM. Looking at the results with User-Network representation we can observe that considering the performance on each single test topic, there is no great difference between the experiments with the Leave-one-topic-out merging the collections and within the collections, indeed the red circles overlay blue ones (purple circles) and green ones (darker-green) in most of the cases (Fig. 2a and Fig. 2c). On the contrary, with Social-Content representation there is an evident change in circle sizes when comparing the two Leave-one-topic-out experiments using these descriptors (Fig. 2b and Fig. 2d). This confirms the considerations that we deduced looking at the averaged results reported in TABLE II and TABLE III, were the metrics, Accuracy, Precision, Recall and F_1-score, are topic-averaged.

Focusing now on TABLE II, there are no clear differences between User-Network and Social-Content representations applied on the merged collection. For this reason we conducted a statistical analysis to assess which representation could be considered as better describing the task. The comparisons were made in the following way:

- Fixing the classifier, we assigned a score equal to 1 to the winning representation, i.e. with the highest performance on a dataset, and a score equal to 0 to the losing one. Then we summed up the scores among all datasets. For example if a representation won over all datasets with a specific classifier, it would have a final score equal to 8.
- Fixing the test dataset, we assigned a score equal to 1 to the winning representation, i.e. with the highest performance on a test topic, and a score equal to 0 to the losing one. Then we summed up the scores among all the classifiers. In this case, if a representation won over all the classifiers, it would have a final score equal to 5.

Both the studies were validated calculating the p-value score with the Wilcoxon ranksum test for each pairwise comparison. Representative graphs of this analysis are shown in Fig. 3, where the fractions above each bar represent the number of

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

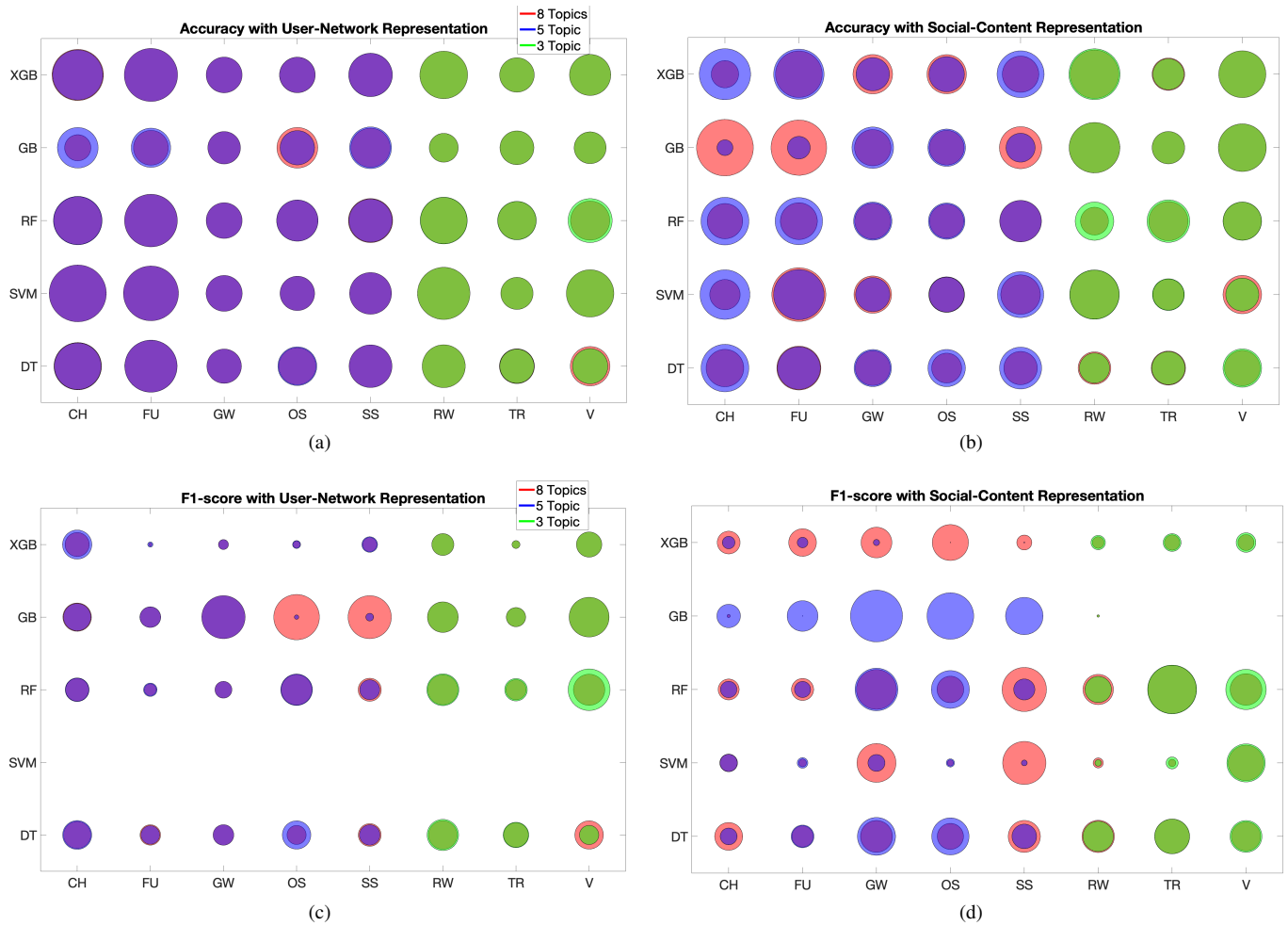


Fig. 2: Graphical representation of how accuracy and F_1 vary for each "classifier"-"test topic" pair. In each plot the classifiers (DT, SVM, RF, GB and XGB) lay on the y-axis, whereas the test topics (CH, FU, GW, OS, SS, RW, TR and V) lay on the x-axis. The panels in the first row refer to the accuracy reported by User-Network (a) and Social-Content (b) descriptors. The panels in the second row show the F_1-score reported by User-Network (c) and Social-Content (d) descriptors.

times that the differences among performances were found statistically significant ($p < 0.05$). For the sake of brevity, we do not report precision and recall as the F_1-score is a measure that depends on them. The height of the blue histograms indicates how many times the User-Network representation wins with that classifier (Fig. 3a and 3b) or on that topic (Fig. 3c and 3d), whereas the height of the orange bars counts the Social-Content representation wins. In most of the cases, the accuracy resulted higher employing User-Network features than using Social-Content features. Moreover, referring to Fig. 3a, the best accuracy results are achieved by User-Network using the Decision Tree and Random Forest classifiers, as in our previous work [10]. A different behaviour is shown in the F_1-score histograms of Fig. 3b: with most of the classifiers User-Network representation generates a score lower than the Social-Content one. This could be due to the different dimensions of the feature spaces: Social-Content has an higher number of descriptors (>300 features) with respect to the User-Network (24 features), so the configuration position of the the samples in feature space would be sparser. This allows the success of classifiers such as the Linear SVM, which does not suffer the course of dimensionality, but it is even favoured by the high dimensional space. Straightforwardly the incompatibility of the User-Network representation with such classifier tears down the winning score of this feature set. This consideration also hold for Fig. 3d, where the User-Network representation generates a F_1-score lower than the Social-Content one. Another possible reason for these results may be the imbalance in the datasets: in fact, as can be seen from the TABLE I, there are more non-rumours than rumours, suggesting that future work may focus on this issue.

Finally, we would like to point out that, despite the low results obtained for both representations, our aim was to compare two feature groups in order to see which one was more informative, and for this reason no feature selection operation was carried out and we didn't use all the classifiers cited in [10], as the Nearest Neighbour, and in [11] as the

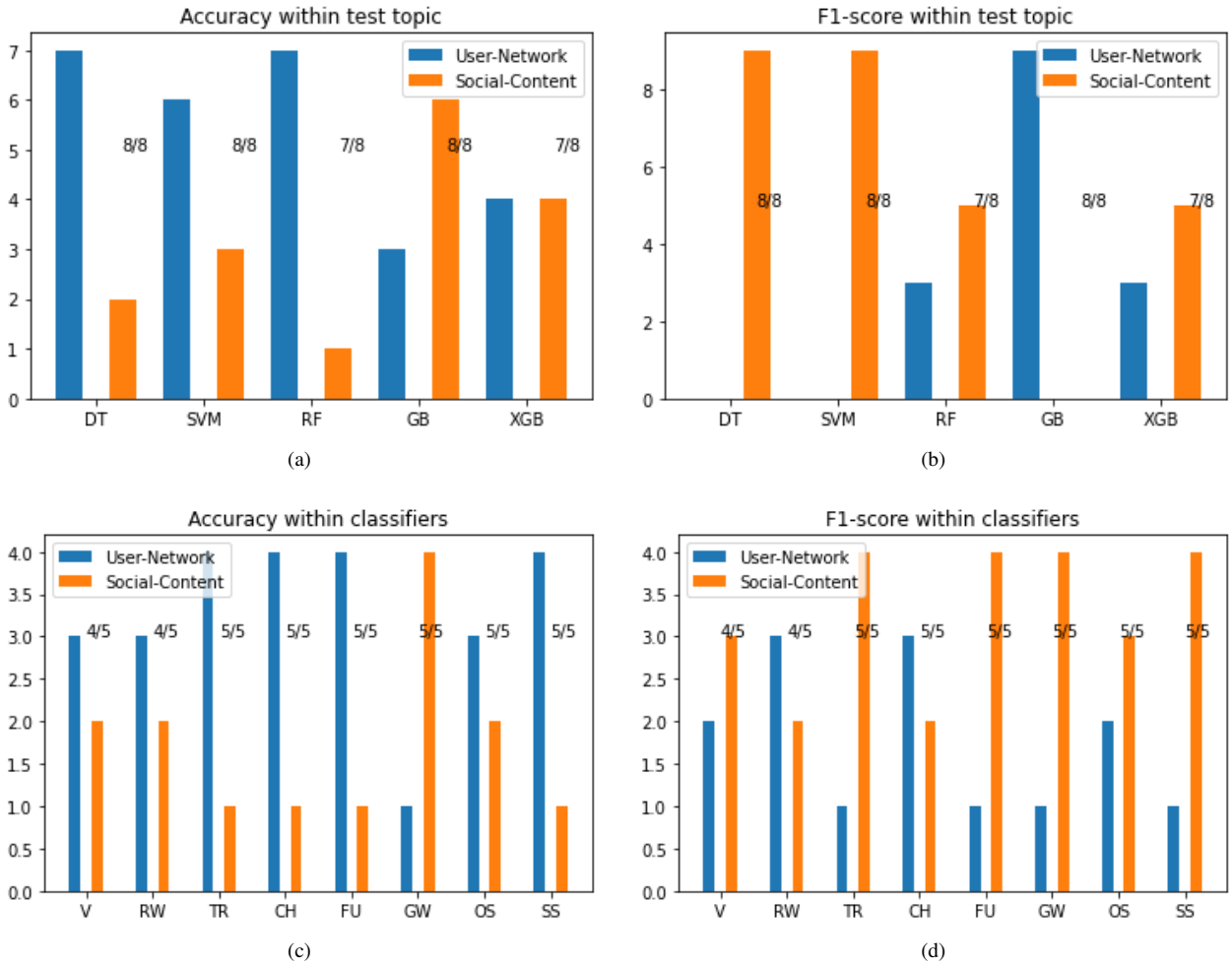


Fig. 3: Bar-based representation of the win-tie-loss comparison between the classifiers and the topics. Each bar denotes the number of wins and the fraction above stands for the number of times the comparison is statistically significant according to the Wilcoxon ranksum test ($p < 0.05$). In the first row the histograms are obtained fixing the classifier: plot (a) refers to the comparison over accuracy, whereas plot (b) to the comparison over F_1-score. In the second row the histograms are obtained fixing the topic: plot (c) refers to the comparison over accuracy, whereas plot (d) to the comparison over F_1-score. Blue bars indicate User-Network wins, whilst orange bars represent Social-Content wins.

Conditional Random Field (CRF).

VII. Conclusions

This work explored the comparison of two state-of-the-art handcrafted representations for the micro-level rumour detection task on Twitter. We carried out an extensive analysis on two datasets collections, the PHEME collection and the UCBM collection, presenting also two new manually labelled datasets.

The tests were performed in a leave-one-topic-out setup, so that the representations would be evaluated for their discriminative power in recognizing rumours in tweets of a topic not present in the training set. These experiments revealed that a feature set with information about the user and his/her interactions with other users is more stable to topic changes.

Future work could be directed toward five main directions: first, to investigate the User-Network representation more thoroughly by carrying out a feature selection or even by combining the two sets of descriptors; second, to identify a well-established deep learning approach to extend the comparative analysis also to deep features; third, it would be worth considering to exploit the best representation found by this comparative analysis in combination with a deep learning approach, as done in the literature with other handcrafted descriptors; fourth, to investigate methods solving the problem of imbalance between classes; fifth, test the User-Network representation also with the Conditional Random Fields classifier used in [11].

References

[1] J. B. L.Wong, "Motivation for sharing news on social media." *Proc. 8th Int. Conf. Social Media Soc.*, pp. 1–5, 2017.

[2] S. Hamidian and M. T. Diab, "Rumor detection and classification for twitter data," *ArXiv*, vol. abs/1912.08926, 2019.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

[3] N. DiFonzo and P. Bordia, *Rumor psychology: social and organizational approaches.* American Psychological Association, 2007.

[4] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," in *IEEE 31st International Conference on Data Engineering*. IEEE, 2015, pp. 651–662.

[5] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proceedings of the 20th international conference on World Wide Web*. ACM, 2011, pp. 675–684.

[6] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 1751–1754.

[7] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks." in *IJCAI*, 2016, pp. 3818–3824.

[8] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PloS one*, vol. 12, no. 1, p. e0168344, 2017.

[9] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1395–1405.

[10] R. Sicilia, S. L. Giudice, Y. Pei, M. Pechenizkiy, and P. Soda, "Twitter rumour detection in the health domain," *Expert Systems with Applications*, 2018.

[11] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *International Conference on Social Informatics*. Springer, 2017, pp. 109–123.

[12] H. M. Jabir, M. A. Naser, and S. O. Al-mamory, "Rumor detection on twitter using features extraction method," in *2020 1st. Information Technology To Enhance e-learning and Other Application (IT-ELA*, 2020, pp. 115–120.

[13] A. Pratiwi and E. Setiawan, "Implementation of rumor detection on twitter using the svm classification method," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, pp. 782–789, 10 2020.

[14] J. Gao, S. Han, X. Song, and F. Ciravegna, "Rp-dnn: A tweet level propagation context based deep neural networks for early rumor detection in social media," in *LREC*, 2020.

[15] D. M. Asghar, A. Habib, A. Habib, A. Khan, R. Ali, and A. Khattak, "Exploring deep neural networks for rumor detection," *Journal of Ambient Intelligence and Humanized Computing*, 10 2019.

[16] E. Kochkina, M. Liakata, and A. Zubiaga, "Pheme dataset for rumour detection and veracity classification," Jun 2018. [Online]. Available: https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_ and_Veracity_Classification/6392078/1

[17] J. Cao, J. Guo, X. Li, Z. Jin, H. Guo, and J. Li, "Automatic rumor detection on microblogs: A survey," *ArXiv*, vol. abs/1807.03505, 2018.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

Application of the A* Algorithm for Navigation of Workers in Simulation Models of Railway Yards

Terézia Sliacka[1], Michal Varga[1] and Norbert Adamko[2]
[1]Department of Informatics
[2]Department of Mathematical Methods and Operations Research
Faculty of Management Science and Informatics, University of Žilina
Univerzitná 8215/1, 010 26, Žilina, Slovak Republic
Teridlo.Sliacke@gmail.com
{Michal.Varga, Norbert.Adamko}@fri.uniza.sk

Abstract— **Modelling the movement of workers in detailed railway yard simulation models can be accomplished by adaptation of universal pedestrian simulation models. While the modelling of walking behaviour remains almost identical, specific approach to pathfinding is desirable – mostly due to rules of movement in the environment and its dynamic changes. The paper presents modification of typical A* pathfinding algorithm, which enables safe navigation of railyard worker agents in specific environment of railway yards. Proposed algorithm utilizes the concept of walkability and takes into account areas with different level of hazard. Such approach can be expanded to consider dynamic moving obstacles typical for railway yards (i.e. moving trains) as well. We present the construction of navigation structures and the algorithm of routing of railyard worker agents utilizing such structure. Proposed solution has been successfully tested in simulation model of marshalling yard.**

Keywords— A* pathfinding algorithm; simulation; rail transport; railyard worker; navigation

I. INTRODUCTION

The year 2021 is the European year of rail [1]. Shifting the cargo and passenger transport to rails brings challenges in maximizing the efficiency of railway infrastructure. Often it is not possible to transport single wagon cargo from its origin to the destination directly, it is rather transported using several carefully composed trains transporting cargo with similar destinations. These trains are composed in special railway stations – marshalling yards. Efficient operation of these vital transportation nodes can help maximize the utilization of railway infrastructure and lower the travel time of the cargo. Simulation tools that model the operation of marshalling yards at a sufficient level of detail, e.g. Villon simulation tool [2] [3], are useful tools, which allow evaluation of various proposed operational scenarios without disturbance of real operation of the yard.

Resources play a key role in service systems, such as railway nodes (marshalling yards are one kind of these nodes). One can utilize the resources to serve the customers, in the case of railway nodes, trains are the customer of the system. Shunting or train locomotives, tracks, as well as personnel belong to resources that are employed in such systems. The personnel executes tasks such as technical and traffic inspection, brake check or un/coupling of wagons. While several simulation models, e. g. Villon or Railsys [4] provide the possibility to model the activities of personnel, none of them, as far as we know, considers the detailed movement of personnel in the railway node environment. By lowering the time personnel spends on transportation from its standpoint position to the place where the service must be executed, we can finish the service sooner and therefore make the whole operation of such system more efficient. There exist operations where detailed modelling of these activities plays a crucial role – consider for example internal locomotive drivers that must be transported from one side of the depot to the other. If one wants to optimize the work of such a system, detailed modelling of personnel including the movement of personnel should not be omitted or simply approximated.

During human personnel movement, specific movement restrictions are applied. These restrictions take into consideration the infrastructure layout (e.g. where it is safe to cross the tracks using bridges, footpaths or underpasses), positions of trains as well as current train movements. The model of personnel movement must be flexible enough to consider not only the current state of the infrastructure but also its dynamic changes (e.g. train movement and tracks occupation) in the future.

There are several decisions that personnel (or any moving human) performs during movement. These can be organized into three levels [5]:

1. strategic – defines tasks and the order of their execution. One tries to fulfil the tasks in the most straightforward order. In the simulation models, this level of deliberation can be performed by the model designer.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

2. tactical – specifies the action choice; an area where the action is performed and the routing into that area. Our focus in this paper is devoted to this step. We propose a route construction respecting specific rules of the environment. To construct a navigation route one should choose a good model of infrastructure representation that is capable of (i) executing some of the path-finding algorithms and (ii) respect specified restrictions.
3. operational – defines the behaviour during the movement. The movement can be performed at different levels of details, for the sake of this paper microscopic movement level is considered. Several approaches tackle operational decision making on a microscopic level in different ways. Probably the most common is the social forces model [6] other approaches are based on magnetic fields [7] or discretization of local space of moving entities [8].

The need for proper movement in a dynamic environment can be found in several areas. In the real world, these problems typically occur in robot navigation. Several approaches exist to tackle this problem [9] [10] [11] [12]. Focusing on pedestrian we can distinguish the solutions on microscopic [13] [14] [15] as well as mesoscopic [16] levels. In general, navigation models used in simulation models utilize graphs [17], matrices [14], fields [18] or a combination of beforementioned [16] that are adapted to dynamic conditions and are capable to reproduce realistic movement and emergent phenomena. The principle of maps splitting the area into smaller cells can be used to adapt the concept of walkability [19]. Walkability defines the will of a pedestrian to move in a specific area based on different indicators [20] [21].

In this paper, we propose an approach combining walkability with navigational structures to find a route through the environment respecting its specific rules. Two navigational structures are composed – (i) a map of heterogenous cells [16] that are weighed respecting the walkability score in a particular part of the environment; (ii) a graph with vertices located in the centres of cells and edges with weight determined from the walkability score. Then the A* algorithm is used to find the shortest path on a constructed graph. This way, it will be later possible to modify the heuristic of the A* algorithm to take into consideration custom pedestrian information about the infrastructure. We also present the application of the proposed navigation model to the navigation of personnel in the marshalling yard.

II. Navigational Structures

Navigation using the walkability principle is not tied to a specific type of infrastructure or moving entities. It can be used to navigate pedestrians on roads (where one would prefer sidewalks and pedestrian crossings) as well as cars in terrain or workers at railway yards. The navigation and movement itself have to be performed on a suitable infrastructure model. In our simulation tools (e.g. PedSim or Villon) we utilise hierarchically organised infrastructure models to model the movement of pedestrians or workers. At the top of the hierarchy, there are levels that are then divided into zones. Besides defining the granularity of the movement model, i.e. microscopic, mesoscopic or macroscopic, each zone also contains different types of maps consisting of rectangular hierarchically arranged cells with distinct properties. The properties of the cells are specific to the particular type of map to which they belong, e.g. gradient map, direct visibility map, mesoscopic map [16] or in our case the walkability map. In order to support detailed movement modelling, the cells do not have a single size but can be divided into smaller sub-cells. When a cell collides with the so-called dividing entity (e.g., an obstacle), it is divided into 4 smaller cells of equal size. The division continues recursively down to the specified depth. The principle of this cell division is discussed in more detail in [16].

A. Walkability map

The walkability map is a structure that is utilized to model the movement of pedestrians in the infrastructure using the walkability principle. The cells of the walkability map define two distinct properties: *passability* and a *walkability score*.

The *passability* value indicates whether it is possible for a moving entity (e.g. worker), to step on a given cell or not. This boolean property is set to false for cells that collide with obstacles that can be never crossed, such as walls or represent prohibited area. Let us take a railway station as an example. In the case of navigation of passengers who wants to board a train, one would probably consider only cells that collide with platforms and sidewalks as passable cells. Passing outside these reserved areas is not permitted for passengers and would therefore be prohibited in the simulation model by cell's passability value. However, railway staff may also enter other areas of the station, but individual areas cannot be considered to be equally preferable for movement – railway workers should primarily use the sidewalk for the crossing, but may also enter the track area if necessary. Therefore, we introduce another property of the walkability map's cells, the *walkability score*. The walkability score indicates how preferable a given cell is for the movement of entities. Besides the physical properties of the infrastructure, this value also depends on the type of moving entities.

In order to support detailed modelling, the walkability map utilizes mentioned cell division principle with specified cell dividing entities.

1) Dividing entities

Entities found in our infrastructure model can be divided into two groups – physically present entities and logical entities. Zones, statistical lines, etc. belong to the group of logical

entities that do not cause division of underlying cells. On the other hand, cells are divided into sub-cells if they contain any physically present entity, which are hence called dividing entities. The division of cells guarantees that the existence of such entities influences the smallest possible parts of the infrastructure models.

Fig. 1 shows a part of the track with a walkway crossing the tracks. The walkability map created above this section of the infrastructure is shown in Fig. 2, where we can identify cells that were divided into four smaller cells due to the presence of dividing entities, i.e. tracks and footpaths.

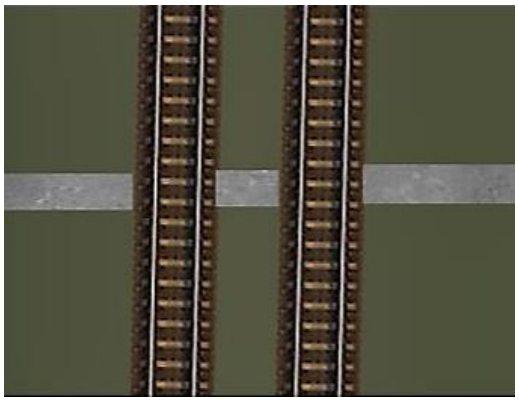


Figure 1. Tracks with intersecting footpath

2) Evaluation of walkability score

To determine the walkability score w_S of cells of the walkability map, we first need to define the weights of the individual dividing entities. These weights describe the pedestrian's preferences to pass through a given dividing entity. As is the standard when implementing priority data structures, in this case, a lower value means a higher priority and thus a greater preference to step on a given entity. The weight is assigned only to entities that are passable. The next step is to determine the *default walkability score* $w_{default}$. The cell's walkability score is then determined as follows:

$$w_S = w_{default} + \sum (\textit{weights of colliding entitites})$$

Fig. 2 shows a map whose cells are coloured according to the walkability score. A cell labelled A and all cells of the same colour do not collide with any entity, thus they have the default walkability score. B-labelled cells collide with the track. The cells marked C contain the footpath and thus have a different walkability score value. Lastly in the cells marked D the track intersects with the footpath and their walkability score is therefore determined by the weights of both entities.

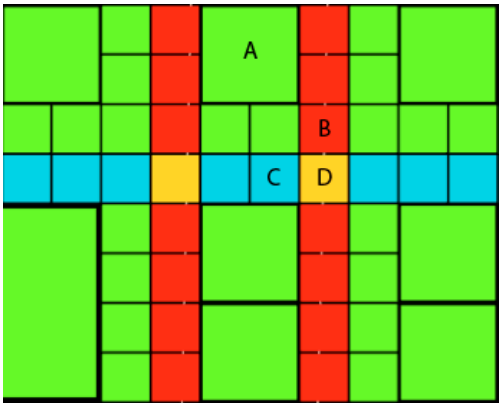


Figure 2. Walkability map of tracks with intersecting footpath

The moving entity is navigated using the map by moving from the current cell to the next suitable one. To determine the suitable cell to move to, a special walkability graph is utilised. Thanks to the way the map is constructed, it is easy to build the graph over it. The path that the moving entity will follow can be then found on the graph using an algorithm for finding the shortest path.

B. Walkability graph

The walkability graph represents properties of the walkability map in a way suitable for a path-finding algorithm. The walkability graph must be constructed in such a way that the path sought on it does not lead through impassable cells. In addition, the path must lead primarily through cells that have a lower walkability score, so that the moving entity is navigated along the most walkable path.

1) Construction of the walkability graph

The walkability graph is created based on the walkability map. The centres of the leaf passable cells (cells that no longer divide and that do not collide with an obstacle) represent the vertices of the graph. We follow 2 rules when creating the graph edges:

- Neighbourhood – An edge is created between each pair of adjacent cells in which there is a vertex.
- Orientation – The edges of the graph are oriented. For correct navigation, it is necessary to distinguish the direction of movement, so this must be reflected in the graph.

The graph created above the map from Fig. 2 is shown in Fig. 3. The edges of the graph are oriented, but overlap in the image.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

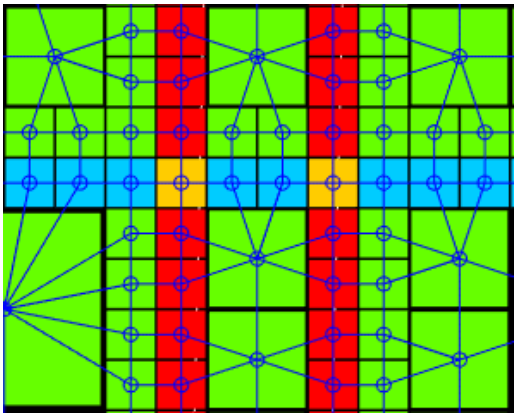


Figure 3. Walkability graph of tracks with intersecting footpath

2) Evaluation of walkability graph

To find the most suitable path in the graph, it is first necessary to define the mechanism by which we will evaluate the edges of the graph. The costs of the edges of the walkability graph are important for finding suitable paths. We are looking not only for the most walkable (e.g. safest) way to the destination but also for the shortest possible one. The shortest route can for example lead through several tracks, which cannot be considered as suitable in terms of walkability. The typical costs of edges based only on their length *d* are therefore not sufficient in this situation. We also need to consider the cell's walkability score, i.e .its suitability to be included in a walking path. Given the above, we determine the cost of the edge *c(e)*, which leads from the vertex *u* to the vertex *v*, according to the following formula

$$c(e) = d_{u,v} \,.\, w_S ,$$

where w_S is the walkability score of the cell with the vertex *v* (the end vertex of the edge) and $d_{u,v}$ is the distance between *u* and *v*.

III. Pathfinding on the Walkability Graph

We use a modified A* algorithm to find the best (most walkable and shortest possible) path. The main modification is the implementation of a mechanism that, if necessary, re-evaluates the edges of the graph based on the current state of the risk cells. Risk cells are cells with walkability score lower than a given threshold and also cells that contain entities identified as risk sources. For example, cells colliding with tracks can be considered as risk cells. If the track is occupied by a train – the edge leading to the cell colliding with this track will be invalidated. At present, we reevaluate the edges at the beginning of the path calculation. However, this approach allows for a dynamic reevaluation of edges even during the movement, leading to a possible change of the path for the moving entity. This dynamic behaviour will be addressed in our future research.

In general, the A* algorithm uses, in addition to edge costs, a heuristic function that estimates the distance of the searched vertex from the target [22]. For each visited vertex *x*, the cost of the best path *f(x)*, which contains vertex *x* is calculated using this equation:

$$f(x) = p(x) + h(x) + q(x),$$

where *p(x)* is the cost of the path from the initial vertex to the vertex *x*. The function *h(x)* represents an estimate of the costs of the path from the vertex *x* to the destination, for which it uses various heuristics, e.g. Manhattan or Euclidean distance. In our implementation, the value of *f(x)* is slightly modified by adding the dynamic component *q(x)*. This component represents the state of the cell, which is determined by the state of the entity located on the risk cell. If the state of the cell allows us to enter the respective place, the value of *q(x)* is equal to 0 (e.g. the track is empty). Conversely, if it is not possible to enter the risk cell (e.g. there is a train standing on the track, therefore crossing of the track is not possible), the value of the dynamic component *q(x)* will be a relatively high number derived from the size of the infrastructure. The value must be high enough to make it worthwhile for the pedestrian to prefer another path that does not contain the vertex *x*.

IV. Utilization of Walkability Map and Walkability Graph in Railway Node Environment

To utilize the walkability principle in a specific environment, one has to first specify the entities causing the impassability of cells (IV.A). Then the splitting entities are identified (IV.B) and the default walkability score (IV.C) is specified. Lastly, the entities' weight is set which will be used to compute the walkability score of respective cells (IV.D).

The walkability graph construction is performed always in the same way. Edges are created between the pair of neighbour passable cells. The edge's price is dependent on the walkability score of the cell the edge is oriented toward.

To be able to find the shortest path it is necessary to specify a proper value that will be used as a result of the heuristic function of the A* algorithm when it is not possible to enter the cell.

A. Passability of cells in the case of navigation of personnel in railway node

When navigating railway personnel, we consider only cells containing obstacles (e.g. walls) as impassable. If we would consider other moving persons (e.g. passengers), the impassibility could be defined differently. In that case, cells containing tracks would be likely to be impassable as well. The final decision related to cell passability is summarized in Table 1.

TABLE I. CELLS PASSABILITY

Cell passability	Colliding entities
Impassable	Obstacles (e.g.. wall)
Passable	No splitting entity, sidewalk, track

B. Splitting entities in the case of navigation of personnel in railway node

Splitting entities in the trackage infrastructure are tracks, obstacles and sidewalks.

- Track – it is not considered to be an obstacle for railway personnel. However, the movement beside or completely out of the track area should be preferred from the passing through the track. This is the reason we use it to split the cells. This way the space with and without the track is separated. If one large cell would contain both the track and the space in between the tracks, that free space would be considered as passable with a penalty, which is not wanted.
- Obstacle – typical obstacle. We want to separate its impassable area from the rest of the free space that can be part of the cell.
- Sidewalk – an entity that can be attractive (walker wants to utilize it for movement) not only for an ordinary passenger but for railway personnel as well. When navigating to the place where a service should be performed, we want the personnel to use the sidewalk if it is possible or desirable. Similarly to the track, we consider the sidewalk as a splitting entity because we want it to affect only that part of the cell, that it collides with (not the large cell of the higher levels).

C. Walkability scores

First, we established the default walkability score. Cells‘ walkability score will be related to the default walkability score and the entities‘ weights. We set the default walkability score to 10, however, it is fully customizable to the model designer (it represents if one will work in the order of tens, hundreds or millions). Entities‘ weights (and therefore the resulting walkability scores) were estimated in the experiments. In the experiments, we worked with a marshalling yard model, which we ran with different settings of entities’ weight. We used a qualified estimate to determine when the workers were navigating as expected. In cases when the expectations were not met, the entities’ weights were adjusted. We used the following approach to find suitable values.

We consider the sidewalk to be an entity, that should be utilized by the personnel for movement in the trackage if possible. For this purpose, we set the value of the sidewalk in a way, that colliding cells would have half of the default value – 5 in our scenario. We performed the experiments and found out, that the weight of the sidewalk must be lower to be attractive enough for the personnel. Using the bisection, we proceed to establish a new value. We continued following these steps (adjusting the weight and perform validating experiment) until the value of the sidewalk fulfilled expectations. Finally, based on the proper results of validating the experiment, we established the value of the cell colliding with the sidewalk to 1. Since the default value is 10 and the weight of the cell with the sidewalk is 1, the weight of the entity sidewalk itself is set to -9.

When setting the weight of the track, we began with the 5 times higher value than the default value (50 in our case). The value was appropriate for the track, however, if the cell contained both track and the sidewalk, the value was too high. The sidewalk had too low weight in such a case. We split the track weight into two intervals <10; 30> and <30; 50>. We performed sets of weight adaptation and validation experiments. We were more focused on the values from the lower interval so we could find the value when the weight of the track is too low. The resulting value was set to 15. This value still prevents personnel to pass through the tracks but is not too high when both sidewalk and track is present in the cell, so the personnel will be willing to use it.

Now we can more clearly explain why it is necessary to consider the entities colliding with the cell to establish cells’ walkability scores, and not try to define the walkability scores of the cells in advance without such assumption. Consider we would set the value of the cell colliding with the sidewalk directly to 1. Then we would set the walkability score of the cell with a track to 15. Then it would be necessary to determine the value of the cell that contains both entities. In our case, it would be simple since there exists only one such combination. However, it would be difficult to define all cells’ walkability scores that cover all combinations in infrastructure with many splitting entities that could collide with each other. We also remark that the sidewalk has a positive impact on the cell’s walkability contrary to the negative impact of the track. To establish the attractiveness of the cell for the purpose of personnel navigation, we find it to be a good way to summarize weights of all colliding entities with the cell together with the cell’s default walkability score.

We present resulting entities weights and cell walkability scores in tables II and III. Recall that these values are proper to use in the environment of the marshalling yard. For other infrastructures, it is necessary to perform calibration.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

TABLE II. WEIGHT OF ENTITIES IN THE MARSHALLING YARD

Splitting entity	Weight of entity
Sidewalk	-9
Track	15

TABLE III. WALKABILITY SCORES OF CELLS IN THE MARSHALLING YARD

Colliding entity	Walkability score
Default	10
Sidewalk	1
Track	25
Track + Sidewalk	16

D. Finding route for the railway personnel

To properly navigate personnel in the railway node it is very important to take into account dynamic changes in the environment. These changes can occur at any time. Therefore, it is not sufficient to set the prices of the edges of the navigation graph in advance since incoming train (or other similar events) will invalidate them. Recall that despite the edges heading towards the tracks are penalized, it is possible to pass through them since it is allowed by walkability scores. If the train is moving on the track, prices of respective edges must be recalculated in such a way, that these edges will be forbidden, or will become the last possible choice for personnel routing algorithm.

Heuristic, which we use in our A* algorithm implementation, penalizes the edges heading towards the occupied tracks. If the track is free at the time of the navigation route finding, the edge price $p(x)$ with the estimate of the distance of the processed vertex $h(x)$ to the destination is considered. If the track is occupied, the dynamic component $q(x)$ is added to the previous two values. In the experiments conducted on the marshalling yard, we simply used a value of 20000. The value is set to be higher than the length of the whole longest possible route in the infrastructure of our model.

V. CONCLUSION AND FUTURE WORK

The presented paper introduced a way to use the A* algorithm in finding the most suitable ways for workers in the railway yards. Navigation structures that we implemented define places that are convenient and safe for the transition, as well as those that are not suitable in terms of walkability. The suitability of stepping on a given part of the track is described by the walkability scores of walkability cells. Above the walkability map, we created a graph whose vertices are located in the centres of the leaf cells. Vertices of the neighbouring passable cells are connected by directional edges. We determined the prices of the edges of the walkability graph based on the walkability score of the cell to which the edge points. We scaled this value by the length of the edge because we do not want the path found in the graph to be only the safest, but also the shortest possible. This principle applies to any infrastructure. We used it in the marshalling yard model where the provided values were tested. After their recalibration, it is possible to navigate pedestrians in other models, even outside the track area.

We have implemented a heuristic function for the A* algorithm, which ensures that the found path respects the limitations of the infrastructure and also takes into account the preferences of the walkers. Due to the trains standing on the tracks, it was necessary to create a mechanism that penalizes the edges leading to the occupied tracks.

We would like to use the presented walkability principle in the issue of pathfinding in models used for decision-making in the construction of the new railway nodes. The user or the designer of the simulation model could draw several alternative routes for the railway staff in the railyard model. A walkability map of a given yard would be used to determine the prices of these roads. Subsequently, we would be able to evaluate which of the proposed routes is the most suitable, respectively following the regulations of the infrastructure. In addition to the dividing entities, the walkability score of the cells could also be affected by the frequency of train movement by a given cell. By identifying the most suitable path, we would be able to determine where to build a footpath. We have not yet encountered the requirement for such functionality of the simulation tool but if it was beneficial for the customer, it would not be difficult to implement.

In future, we would like to further expand the scheme with dynamic reactions of agents to situations occurring in the railyard. We are planning to modify the behaviour of the dynamic agent based on the BDI (Belief-Desire-Intention) paradigm of the ABAsim architecture [23]. By creating sensors that will monitor events in the yard at regular intervals, we will be able to make new beliefs and process information about the arrival of trains on the tracks or the passage of trains across the track. The information (beliefs) obtained from sensors will assist the actor to have an updated view of the yard, which will allow him to react quickly. The dynamic reaction could be implemented by changing the currently performed intention. This could work so that the agent smoothly switches from executing the current plan to execution of another plan after finding an interesting information (e.g. train arrival). By implementing dynamic behaviour, the agent should be able to estimate how long he can continue on his journey so that he is not hit by an incoming train and also he should stand at a safe distance from that train. Another situation that we would like to

solve with this mechanism is to bypass the train, which will come to the track only after finding the path on which the agent is currently walking. We expect that with the help of the BDI paradigm and security sensors, we will be able to achieve these goals.

REFERENCES

[1] "EUROPEAN YEAR OF RAIL," [Online]. Available: https://europa.eu/year-of-rail/index_en. [Accessed June 2021].

[2] N. Adamko and V. Klima, "Optimisation of railway terminal design and operations using villon generic simulation model," Transport, vol. 23, no. 4, pp. 335-340, 2008.

[3] A. Kavička, V. Klima and N. Adamko, "Simulations of transportation logistic systems utilising agent-based architecture," International Journal of Simulation Modelling, vol. 6, no. 1, p. 13–24, 2007.

[4] "RMCon International," rmcon Rail Management Consultants International GmbH, [Online]. Available: view-source:https://www.rmcon-int.de/home-en/. [Accessed June 2021].

[5] S. P. Hoogendoorn and P. H. L. Bovy, "Pedestrian route-choice and activity scheduling theory and models," Transportation Research Part B: Methodological, vol. 38, no. 2, pp. 169-190, February 2004.

[6] D. Helbing, L. Buzna, A. Johansson and T. Werner, "Self-Organized Pedestrian Crowd Dynamics: Experiments, Simulations, and Design Solutions," Transportation Science, vol. 39, no. 1, 1 February 2005.

[7] S. Okazaki and S. Matsushita, "A Study of Simulation Model for Pedestrian Movement with Evacuation and Queuing," in Proceedings of the International Conference on Engineering for Crowd Safety, London, 17-18 March 1993, London, 1993.

[8] M. Varga and M. Mintál, "Microscopic pedestrian movement model utilizing parallel computations," in 2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI), 2014.

[9] M. de Berg, O. Cheong, M. van Kreveld and M. Overmars, Computational Geometry, Algorithms and Applications, 3 ed., Springer, Berlin, Heidelberg, 2008, p. 386.

[10] J. Guzzi, A. Giusti, L. M. Gambardella, G. Theraulaz and G. A. Di Caro, "Human-friendly robot navigation in dynamic environments," in 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 2013.

[11] M. K. Habib, "Real Time Mapping and Dynamic Navigation for Mobile Robots," International Journal of Advanced Robotic Systems, vol. 4, no. 3, 2007.

[12] M. Kocifaj and N. Adamko, "Infrastructure representation for container terminal simulation," Central European Journal of Computer Science, vol. 4, no. 4, 2014.

[13] K. Teknomo and A. Millonig, "A Navigation Algorithm for Pedestrian Simulation in Dynamic Environments," in 11th World Conference on Transport Research, Berkeley CA, United States, 2007.

[14] A. Kormanová, M. Varga and N. Adamko, "Hybrid model for pedestrian movement simulation," in The 10th International Conference on Digital Technologies 2014, Zilina, Slovakia, 2014.

[15] J. van den Berg, M. Lin and D. Manocha, "Reciprocal Velocity Obstacles for real-time multi-agent navigation," in 2008 IEEE International Conference on Robotics and Automation, Pasadena, CA, USA, 2008.

[16] M. Čadecký, M. Varga and N. Adamko, "Mesoscopic movement model of deliberative pedestrian agents," International Conference on Information and Digital Technologies, pp. 61-67, 2015.

[17] A. Sud, E. Andersen, S. Curtis, M. C. Lin and D. Manocha, "Real-Time Path Planning in Dynamic Virtual Environments Using Multiagent Navigation Graphs," IEEE Transactions on Visualization and Computer Graphics, vol. 14, pp. 526-538, 2008.

[18] Y. Jiang, B. Chen, X. Li and Z. Ding, "Dynamic navigation field in the social force model for pedestrian evacuation," Applied Mathematical Modelling, vol. 80, pp. 815-826, April 2020.

[19] W. D. Lee, W. Ectors, T. Bellemans, B. Kochan, D. Janssens and G. Wets, "Investigating pedestrian walkability using a multitude of Seoul data sources," Transportmetrica B: Transport Dynamics, vol. 6, no. 1, pp. 54-73, 2018.

[20] G. D'Orso and M. Migliore, "A GIS-based method for evaluating the walkability of a pedestrian environment and prioritised investments," Journal of Transport Geography, vol. 82, 2020.

[21] A. Galanis and N. Eliou, "Evaluation of the pedestrian infrastructure using walkability indicators," WSEAS Transactions on Environment and Development, vol. 7, no. 12, pp. 385-394, 2011.

[22] Hart, Peter E.; Nillson, Nils; Raphael, Bertram;, "A Formal Basis for the Heuristic Determination of Minimum Costs Paths," IEEE Transaction on Systems Science and Cybernetics, pp. 100-107, 1968.

[23] Varga, Michal; Adamko, Norbert;, "Integration of BDI paradigm into ABAsim archietcture," Journal of Information, Control and Management Systems, 2014.

 978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Securing Constrained Edges in a Triangulation

Ján Kučera[1], Norbert Adamko[2], Michal Varga[1]
[1]Department of Informatics
[2]Department of Mathematical Methods and Operations Research
Faculty of Management Science and Informatics, University of Žilina
Univerzitná 8215/1, 010 26, Žilina, Slovak Republic
janislavkucera@gmail.com
{Michal.Varga, Norbert.Adamko}@fri.uniza.sk

Abstract – **Computing triangular meshes with constrained edges requires both advanced mathematics and thoughtful logic of the implemented algorithm. Our approach to this is modifying an already created triangulation (e.g., Delaunay triangulation), which can be constructed by a simple algorithm. By swapping edges between triangles and splitting them in needed scenarios, we achieve a mathematically simple algorithm, that supports the construction of wanted edges. This algorithm has been tested in a scenario where it was used to automatically generate terrain based on an infrastructure of a railway station model.**

Keywords – triangulation, constrained triangulation, securing edges, terrain generation.

I. Introduction

Triangles, being the simplest polygons, are commonly used to express more complex polygons (monotone polygon triangulation) [1]. Similarly, a point-set triangulation is a task of dividing the area/plane into triangles. The triangulation itself can be steered to maximize/minimize internal angles, edges count etc., as required in various situations.

A common subtask for triangulation algorithms is to guarantee the existence of certain edges defined by two input points. There is no simple way to insert an edge into the triangulation after it is computed (triangulation, as a maximal planar subdivision, does not allow the existence of an edge collisions outside the input set of points [1]). As there is at least 2.631^n possible triangulations of n points [2], looking through them for the one that contains all the wanted edges is not a viable option. To construct a triangulation that features the desired edges, one must triangulate the point set with a constraint for them [3].

Delaunay triangulation [4], being the most common type of triangulation, is the backbone of many computer-generated triangular meshes and objects. The advantage of constructing triangles with maximized internal angles makes it a perfect adept for generating a terrain with non-specific positions of leading vertexes [1]. Many authors have already come up with algorithms to construct constrained triangulations with, whether in a single run [5] or multiple steps, using polygon triangulations [6] [7] or edge flips [8]. These algorithms consist of advanced mathematics and provide a complex look at the problem – delivering a solid all-in-one solution for the problem.

In this paper we propose a way of modifying an already constructed triangulation (can be Delaunay), shaping it to contain the wanted edges. The basic triangulation can be constructed with ease by any known algorithm [1], as we will modify it after it is calculated. A note to take is that our algorithm does not need the input triangulation to fit the Delaunay definition – this is just a suggestion for working with good-looking triangles.

The presented algorithm deals with multiple constrained edges intersection during execution. The intersection points do not need to be a part of the input set. If there are any colliding edges, the algorithm will find their intersection points and add them into triangulation with no additional complexity.

We implemented the algorithm in simulation software Villon, where we generate a terrain that is sculpted by railways, roads, houses, and contour lines. Experiments regarding the speed of the sculpting process and the results of our algorithm will be presented as well.

II. Initial Triangulation

Fig. 1 shows an example of a simple railway infrastructure that can be an input for a triangulation algorithm. The infrastructure consists of railway tracks and a single building, which overlaps with one of the tracks. The tracks are represented by a polyline with a defined width. To describe the track's geometry using points we create an outline by extending the centre polyline to both perpendicular directions. To express the building's geometry, we can just use the corner points of the polygon.

Figure 1 - Input infrastructure for the algorithm

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

With the construction of a Delaunay triangulation of the entry set of points (in our case gathered from simple railway infrastructure), we get the starting solution for our algorithm (Fig. 2). The entry set of points consists of outline points of railways and additional points aligned in a grid, to fill the empty places in the infrastructure. Resulting triangulation does not guarantee the existence of all required edges (e.g., track outlines). However, as shown in Fig. 2, the Delaunay triangulation did naturally secure the existence of many desired edges.

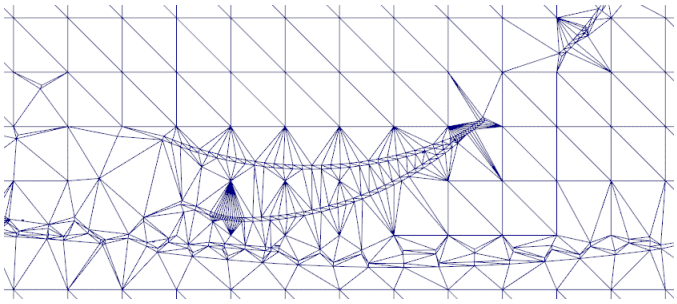


Figure 2 - Delaunay triangulation of input set of points

This property makes the Delaunay triangulation a suitable method of initial triangulation as it already secured the existence of many desired edges. The triangulation will be further processed to secure the existence of all additional edges.

To secure additional edges, we will also need to find a way to deal with parts of the infrastructure, where multiple outlines of entities collide with each other (e.g., railway switches) – at such places some desired edges will intersect with each other, forcing us to add the intersection point into the set of triangulated points. The presented algorithm solves such a situation on the go, calculating the intersection point of object outlines during its execution. The black (unlabelled) points in Fig. 3 are outline points, which are part of the initial triangulated set of points, whereas the red points (labelled a and b) are computed during securing a chain of black points, as their intersections.

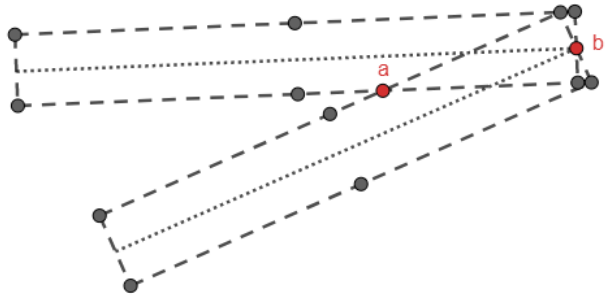


Figure 3 – Intersection points of track outlines (points a and b)

The basic idea of the algorithm is to identify all triangles that *obstruct* the desired edge and modify them in neighbouring pairs. By iteration through the list of desired edges of triangulation and securing their existence one by one, we achieve their appearance in the triangulation. The identification of not wanted triangles and the handling of specific cases is presented in the following chapters.

III. Identifying the Interfering Triangles

When securing an edge (*u*, *v*), we need to identify all the triangles that obstruct this edge. Let us define the triangle that *obstructs* the edge (*u*, *v*) as follows:

- If the triangle is constructed by one of the vertexes {*u*, *v*}, then the imaginary edge (*u*, *v*) must collide with all the triangle's edges.
- If the triangle is not constructed by one of the vertexes {*u*, *v*}, then the imaginary edge (*u*, *v*) must collide with exactly two triangle's edges.

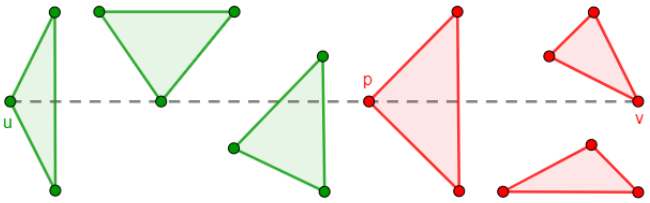


Figure 4 - Triangles obstructing the existence of the desired edge

Note that the situation in which the desired edge collides with just a single edge of a triangle can never occur, as that will mean that the point *u*/*v* do not belong to the input set of triangulated points.

In Fig. 4, the first three triangles from the left fulfil the conditions of obstructing the edge (*u*, *v*) and the last three triangles do not match either of the given conditions. Let us highlight a special case of the fourth triangle (one with a vertex labelled *p*). In such a case, we can split the problem of securing an edge (*u*, *v*) into two parts – securing an edge (*u*, *p*) and (*p*, *v*) – and continue the algorithm using recursive calls.

An edge that has already been secured during the execution of the proposed algorithm is called a *fixed edge*. A fixed edge cannot be removed during the process of securing other edges. Marking edges as fixed allows us to identify the intersection points of constrained edges without any extra computing.

To identify all the triangles that obstruct the edge (*u*, *v*), we can use a simple straight walk [9] [10] across the triangulation or traverse through the not visited neighbours of a triangle that obstruct the desired edge. The proposed algorithm relies on the identified triangles being neighbours, as it iteratively modifies them in neighbouring pairs as you can see in Fig. 5. The modifications are discussed in chapter IV.

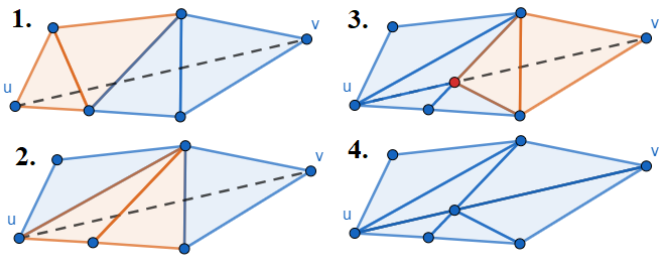


Figure 5 - Securing an edge (*u*, *v*) using promoted algorithm

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

IV. Modifying Triangle Pairs

The way of modifying triangle pairs will depend on the position of two points, discussed later as p and q. Based on their position relative to the triangle pair's common edge, we identify four possible alignments to solve. Let (e_1, e_2) be the common edge of the triangle pair, the vertexes u and w the third vertexes of the triangles. The points p and q are determined as follows:

- p – intersection point of (e_1, e_2) and (u, v)
- q – intersection point of (e_1, e_2) and (u, w)

1. $p \notin \{e_1, e_2\}$ and $q \notin \{e_1, e_2\}$

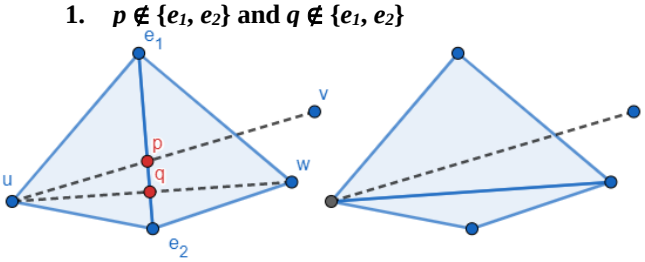


Figure 6 - Triangle pair modification (case 1 – before / after)

The most common outcome of the points is when they both lie on the edge (e_1, e_2) and are neither e_1 or e_2. In this case, we can flip the edge between the triangle pair, i.e. replace the edge (e_1, e_2) with new edge (u, w), or split the two triangles into four triangles (at the point p), see Fig. 6.

2. $p \notin \{e_1, e_2\}$ and $q \in \{e_1, e_2\}$

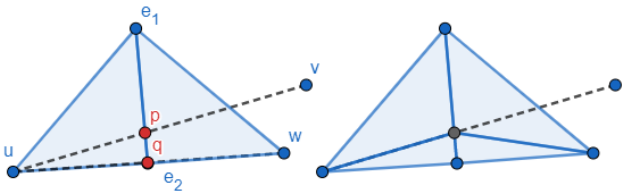


Figure 7 - Triangle pair modification (case 2 – before / after)

Another possible outcome is when the point q is one of the $\{e_1, e_2\}$, in this case we cannot flip the edge (e_1, e_2). The flip will lead to the creation of a triangle that does not fit the triangle inequality rule. The only viable option is to split the triangles into four at the point p, see Fig. 7.

3. $p \in \{e_1, e_2\}$ and $q \notin \{e_1, e_2\}$

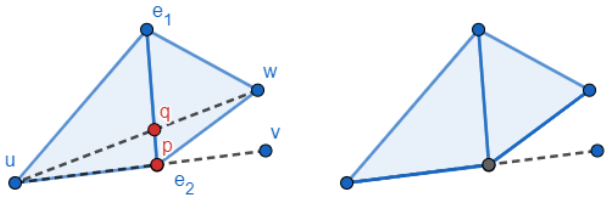


Figure 8 - Triangle pair modification (case 3 – before / after)

The exact opposite of the previous case – where the points p and q are swapped, flipping the common edge would lead to no advance and splitting into four triangles at p would lead to the creation of incorrect triangles. However, the segment (u, p) is a part of the desired edge (u, v). In the shown case we continue the algorithm by a recursive call to secure the rest of the desired edge – (p, v), see Fig. 8.

4. q does not exist

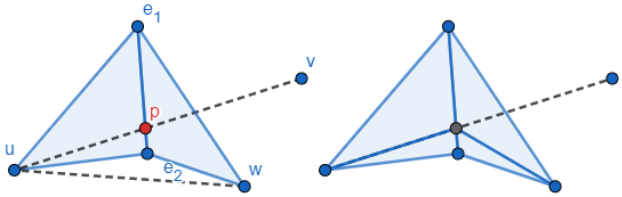


Figure 9 - Triangle pair modification (case 4 – before / after)

In the last possible situation, the intersection point q does not exist. Flipping edge, in this case, would lead to the creation of overlapping triangles in the triangulation, which would make the triangulation invalid. The only correct solution for this case is to split the triangles into four at the point p, see Fig. 9.

Note that splitting triangles into four at the point p is always a correct solution for the triangle pair – except for when the point p is one of $\{e_1, e_2\}$. Flipping the edge when possible, however, saves memory and lowers the number of final triangles. It is for the programmer to decide whether to save the coding time or the triangles count.

Pseudocode 1. SecureEdge(*u*, *v*)

```
1.  Find a triangle with a vertex u.
2.  From the triangle fan of vertex u, identify a triangle t that obstructs the edge (u, v).
3.  Let T be an empty list of triangles.
4.  Insert t into T.
5.  From the last triangle in T identify a neighbouring triangle r, such that r ∉ T
    and r obstructs the edge (u, v).
6.  Add r into T.
7.  If r does not have a vertex v, go to step 5.
8.  Let t := T[0], w := u and i := 1.
9.  If t consists of both u and v, mark this edge as fixed and exit.
10. While t != null or v is a vertex of t, do:
11.    If t does not have a neighbour T[i], set t := T[i-1]
12.    Execute SecureEdgeStep(t, T[i], w, v) – see Pseudocode 2
13.    Increase i by 1.
14. If t != null, mark the edge (w, v) as fixed.
```

Pseudocode 2. SecureEdgeStep(var t_1, t_2, var u, v)

```
1.  Let e1 and e2 be the common vertexes of t1 and t2.
2.  Let w be the third vertex of t2.
3.  Calculate p as an intersection point of (e1, e2) and (u, v)
4.  Calculate q as an intersection point of (e1, e2) and (u, w)
5.  If the edge (e1, e2) is fixed, q exists, and q ∉ {e1, e2} then:
6.     Flip the edge between t1 and t2 and exit.
7.  If p = e1(2) then:
8.     Mark edge (u, e1(2)) as fixed.
9.     SecureEdge(e1(2), v) – Pseudocode 1
10.    Set t1 := null and exit.
11. Split triangles t1 and t2 into four at the point p.
12. Mark edge (u, p) as fixed.
13. If w = v then:
14.    Mark edge (p, v) as fixed.
15.    Set t1 := null and exit.
16. Set u := p.
17. Set t1 as one of 4 created triangles, which obstructs the edge (u, v).
```

V. Experimental Results

In the process of discovering each object's outline points, we received an enumeration of the desired edges of the triangulation. By executing the *SecureEdge* algorithm on each pair of the outline points we achieve the desired outcome, shown in Fig. 10. Processing each outline's desired edges in a continuous chain minimizes the time it takes to execute step 1 in Pseudocode 1 – when securing (*b*, *c*) after (*a*, *b*), we do not need to look for a triangle with vertex *b* as it was just processed.

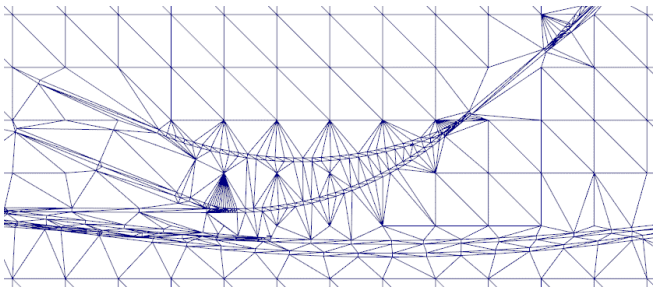


Figure 10 - Triangulation of input infrastructure with secured constrained edges

Removing code steps 5 and 6 in Pseudocode 2 leads to implementation without triangle flips. The implementation without triangle flips provides favoured triangulations, especially when the constrained edges are long.

The algorithm solved the track intersection in track switches as well as the situation with a building that collides with a track (using the same logic and code, the algorithm deals with collisions of various outlines), see Fig. 10 and 11. Without computing the intersection points, the rail switches will be deformed after the triangulation, as the intersection point would not exist.

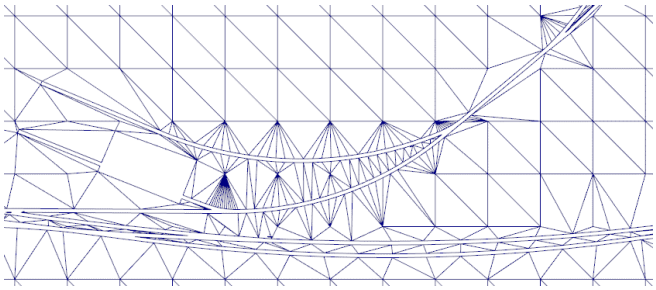


Figure 11 - Triangulation of input infrastructure with only terrain forming triangles

Filtering the triangles inside the infrastructure entities we get a set of triangles, that form the terrain of the scene. The differentiation of the triangles can also be inverted, and the single triangulation can be used for multiple purposes. In our simulation software, we used this possibility and use a single triangulation to create both the terrain and the railway meshes.

We tested the overall performance of the algorithm on multiple railway infrastructures of real and virtual railway yards. We have measured the execution time of the code for securing the existence of desired edges. The results of the experiments are summarized in the following table.

TABLE I. Execution Times on Railway Infrastructures

Triangles	Constrained edges	Time (ms)
55412	4816	63
403624	6459	248
65398	9158	193
64874	19115	460
383196	28984	562

As expected, the execution time depends both on the number of edges to secure, and in a minor way also on the number of triangles in the triangulation.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The time complexity of the algorithm depends heavily on the specific triangulation and the edges we desire. The time it takes for a single edge to secure depends on the number of obstructing triangles. In the worst-case scenario the time complexity of securing a single edge is $O(n)$, n being the total count of triangles in the triangulation. However, this is unlikely to happen in real-world situations (imagine the triangulation to be a single triangle fan, with a constrained edge across it). The algorithm modifies only triangles that obstruct the desired edge.

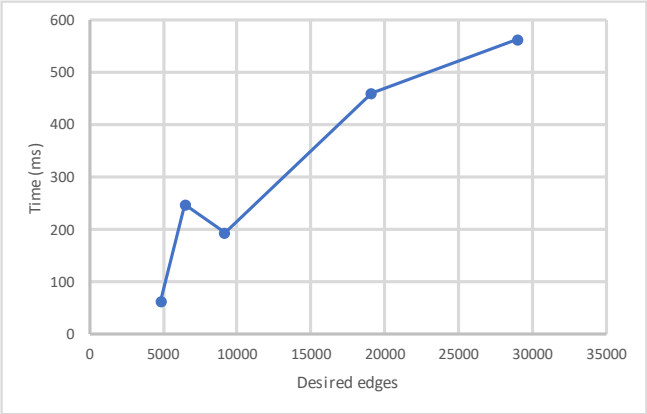


Figure 12 - Time complexity graph of the algorithm

From the graph in Fig. 12 we deduce, the complexity does not rise exponentially in any of the tested infrastructures and scales with the desired number of constrained edges in a linear ratio.

VI. Conclusion

We have provided an alternative approach to the problem of constructing a constrained triangulation. Although there are already algorithms that perform well [5] [6] [7] [8], their implementation may not be simple enough. The proposed algorithm takes any triangulation as a base and by iterative processing secures all of the desired edges. This is achieved only by simple line intersection calculations and progressive modifications of triangle pairs.

An important feature that comes with this approach is its efficiency in dealing with intersections of multiple constrained edges. Without any additional computations, the algorithm finds intersection points and includes them in the triangulation with no processing delays.

The overall performance of the algorithm has been tested on several large real-world railway yard infrastructures. The results show the algorithm's computation time scales linearly with the number of triangles and edges to secure.

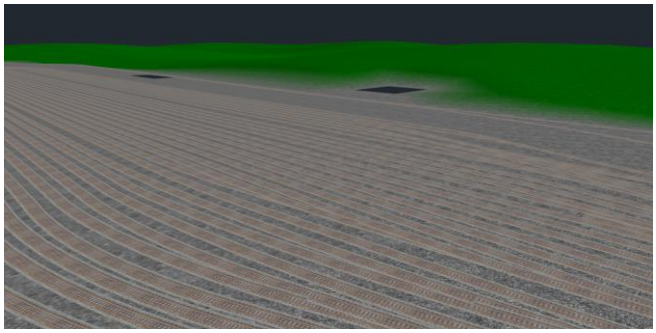


Figure 13 – Terrain meshes generated by the algorithm

If one is to implement this algorithm, we suggest considering at what purpose does he or she need the constrained edges to exist. If the purpose is to provide real-time triangulations with low response times to the user, our algorithm is not the right choice. It may work well if the triangulations are not vast, but generally, it is not a good practice. The algorithm is supposed to be simple to implement and do the needed job, even for a price of a relatively small time delay.

References

[1] M. d. Berg, O. Cheong, M. v. Kreveld and M. Overmars, Computational Geometry, Berlin Heidelberg: Springer, 2008.

[2] O. Aichholzer, V. Alvarez, T. Hackl, A. Pilz, B. Speckmann and B. Vogtenhuber, "An Improved Lower Bound on the Minimum Number of Triangulations," International Symposium on Computational Geometry, no. 7, pp. 7:1-7:16, 2016.

[3] J. R. Shewchuk, "General-Dimensional Constrained Delaunay and Constrained Regular Triangulations, I: Combinatorial Properties," Discrete & Computational Geometry, no. 39, pp. 580-637, 2008.

[4] B. Delaunay, "Sur la sphère vide," Bulletin de l'Académie des Sciences de l'URSS, Classe des Sciences Mathématiques et Naturelles, no. 6, pp. 793-800, 1934.

[5] L. P. Chew, "Constrained Delaunay Triangulations," Proceedings of the Third Annual Symposium on Computational Geometry, p. 215–222, 1987.

[6] J. De Loera and W. Komornicki, "Computing constrained Delaunay triangulations," [Online]. Available: http://www.geom.uiuc.edu/~samuelp/del_project.html.

[7] V. Domiter, "Constrained Delaunay Triangulation using Plane Subdivision," [Online]. Available: https://old.cescg.org/CESCG-2004/web/Domiter-Vid/.

[8] S. W. Sloan, "A fast algorithm for generating constrained Delaunay triangulations," Computers & Structures, pp. 441-150, 1993.

[9] O. Devillers, S. Pion and M. Teillaud, "Walking in triangulation," 2006. [Online]. Available: https://hal.inria.fr/inria-00072509/document.

[10] E. P. Mücke, I. Saias and B. Zhu, "Fast randomized point location without preprocessing in two- and three-dimensional Delaunay triangulations," in Computational Geometry 12 , 1999, pp. 63-83.

Horizontal scaling method for a hyperconverged network

Andriy Kovalenko
Faculty of Computer Engineering and Control
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
andriy_kovalenko@yahoo.com

Heorhii Kuchuk
Faculty of Computer and Information Technologies
National Technical University «KhPI»
Kharkiv, Ukraine
kuchuk56@ukr.net

Nina Kuchuk
Faculty of Computer and Information Technologies
National Technical University «KhPI»
Kharkiv, Ukraine
nina_kucnuk@ukr.net

Jozef Kostolny
Faculty of Management Science and Informatics
Univeraity of Zilina
Zilina, Slovakia
jozef.kostolny@fri.uniza.sk

***Abstract*—Today cloud platforms are gradually being replaced by hyperconverged. With a hyperconverged infrastructure, the servers, networks, storage and computing power are combined. This is done through specific software. The main advantage of hyperconverged network is to reduce operating costs. However, system performance can be reduced due to the standardization of hardware modules and centralization of control, and, consequently, QoS deteriorates. At the same time, companies that operate computer systems are attracted by the ability to quick horizontal scaling and lower operating costs. Hyperconverged networks often use addition of a new hardware modules to meet QoS requirements using a simple horizontal scaling implementation. The purpose of the paper is to develop such a horizontal scaling method, which allows decreasing transaction processing time while total cost of data transmission does not exceed the specified values. The method also allows determining the optimal characteristics of additional hardware.**

Keywords—horizontal scaling; resource allocation; hyperconverged network, delay

I. INTRODUCTION

A. Motivation

Today cloud platforms are gradually being replaced by hyperconverged [1]. With a hyperconverged infrastructure, the servers, networks, storage and computing power are combined. This is done through specific software [2]. The main advantage of hyperconverged network is to reduce operating costs. However, system performance can be reduced due to the standardization of hardware modules and centralization of control, and, consequently, QoS deteriorates [3]. At the same time, companies that operate computer systems are attracted by the ability to quick horizontal scaling and lower operating costs.

As the load on the hyperconverged network increases, the QoS requirements can be violated in terms of temporal and probabilistic factors. To recover the desired values the following options can be used:

- reconfigure the system;
- reallocate system resources;
- add new hardware.

In a hyperconverged network, addition of a new hardware is often the most attractive option due to a simple horizontal scaling implementation. However, this raises the problem of choosing the optimal characteristics of such hardware, taking into account the condition to meet QoS requirements under the lowest possible costs.

B. Analysis of related works

The specific features of the hyperconverged networks are considered in [3, 4]. Methods for improving QoS performance are presented in modern literature, for example, in [5-7]. However, these references do not take into account the specificity of hyperconverged networks. In [8, 9], the methods of resource distribution in hyperconverged networks are considered. References [10, 11] consider the features of the user tasks of computer system. Authors in [12] consider the problem of minimizing the average delay. Authors in [13, 14] deal with system reconfiguration. References [15, 16] propose approaches to addition of a new hardware. However, these sources did not consider the problem of optimal characteristics choosing for additional hardware in a hyperconverged network environment. Moreover, the existing models do not consider the possibility of taking into account cost constraints during the hardware maintenance.

C. Goals and structure

The problem of finding the most acceptable horizontal scaling method in hyperconverged networks is currently actual. It is targeted on improving QoS indicators. The purpose of the paper is to develop such a horizontal scaling method, which allows decreasing transaction processing time while total cost of data transmission does not exceed the specified values. Application of such method should allow determining the optimal characteristics of additional hardware.

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

The paper is structured as follows. Section 2 describes the mathematical model of message transmission process in a hyperconverged network environment. Section 3 describes a method for message transmission time minimization. Section 4 is devoted to discussion of the proposed method effectiveness.

II. MATHEMATICAL MODEL OF MESSAGE TRANSMISSION PROCESS IN A HYPERCONVERGED NETWORK ENVIRONMENT

A. Features of a hyperconverged network operation

To achieve an efficient allocation of resources, it is proposed to use, as a limiting condition, the cost of transmission of the certain amount of information per single unit of available bandwidth. Then the total amount of transmitted information will determine the total income from the use of communications. The total cost of the network can be determined at the final design stage, calculated, for example, as the sum of appropriate bandwidths. This approach will allow determining the payback period of the network, taking into account the costs of its implementation and income from its operation.

B. Development of a mathematical model

Based on Little's formula [11], according to the Kleinrock approximation [12], the queue of packets at the input to each elementary hardware module of hyperconverged network, we represent a queuing system (QS) of M/M/1 type. Consequently, the queue input of i -th elementary hardware module of a hyperconverged network ($i \in \overline{1,I}$, I is the quantity of elementary hardware modules of the hyperconverged network under synthesis) receives a Poisson stream of hypervisor's requests with the intensity of λ_i requests per second and μ_i average time of servicing requests per second. These random variables are distributed exponentially. Then the utilization factor of an elementary hardware module is defined as $\rho_i = \lambda_i / \mu_i$.

Therefore, the relative throughput (the average proportion of incoming requests to the system) can be defined [2] as

$$\bar{g}_i = \frac{1-\rho_i^{m_i+1}}{1-\rho_i^{m_i+2}}, \tag{1}$$

where m_i is the queue size for the i -th elementary module of hyperconverged network (number of places).

The average number of requests in the queue can be defined [2] as:

$$\bar{r}_i = \frac{\rho_i^2 \left[1-\rho_i^{m_i}\,(m_i+1-m_i\cdot\rho_i)\right]}{(1-\rho_i^{m_i+2})\,(1-\rho_i)}. \tag{2}$$

The average delay time $\bar{T}_i$ for the i -th elementary hardware module of hyperconverged network, that is equal to the staying of the request in QS, can be expressed by the general formula [13]:

$$\bar{T}_i = \bar{r}_i / \lambda_i + \bar{g}_i / \mu_i . \tag{3}$$

Taking into account (1) and (2), expression (3) after transformation takes the following form:

$$\bar{T}_i = \frac{1}{\mu_i} \frac{1-\rho_i^{m_i+1}\left[(m_i+2)-\rho_i\,(m_i+1)\right]}{\left(1-\rho_i^{m_i+2}\right)(1-\rho_i)}. \tag{4}$$

We denote

$$\sum_{k=0}^{m_i+1} \rho_i^k = \frac{1-\rho_i^{m_i+2}}{1-\rho_i} = \Sigma_{m_i}, \tag{5}$$

where Σ_{m_i} is a sum of geometric progression.

Expression (4) can be represented in a modified form:

$$\bar{T}_i = \frac{1}{\mu_i} \cdot \frac{\sum_{\alpha=0}^{m_i} (1+\alpha)\,\rho_i^\alpha}{\sum_{\alpha=0}^{m_i+1} \rho_i^\alpha} = \frac{1}{\mu_i} \cdot \frac{\left(\sum_{\alpha=0}^{m_i+1} \rho_i^\alpha\right)'}{\sum_{\alpha=0}^{m_i+1} \rho_i^\alpha} = \frac{1}{\mu_i} \frac{\Sigma'_{m_i}}{\Sigma_{m_i}}, \tag{6}$$

where Σ' denotes the derivative of $\frac{\partial \Sigma}{\partial \rho}$.

Then, taking into account (6), the average delay time in the entire network can be found as

$$\bar{T} = \frac{1}{\gamma} \sum_{i=1}^{I} \left(\rho_i \frac{\left(\sum_{\alpha=0}^{m_i+1} \rho_i^\alpha\right)'}{\sum_{\alpha=0}^{m_i+1} \rho_i^\alpha} \right) = \frac{1}{\gamma} \sum_{i=1}^{I} \rho_i \frac{\Sigma'_{m_i}}{\Sigma_{m_i}}, \tag{7}$$

where γ is the total number of requests entering the network per second (total network traffic).

Dependence (6), when $m_i \to \infty$, is transformed into the following well-known formula [11]:

$$\overline{T} = \frac{1}{\gamma}\sum_{i=1}^{I}\frac{\rho_i}{1-\rho_i}, \quad (8)$$

which corresponds to the network model in the form of QS with an unlimited queue [11], i.e. corresponds to the QS network model without failures.

Function (7) is a convex function, but does not contain extrema, which does not allow finding the minimum of the average delay time by calculating the partial derivatives.

Thus, this problem is a conditional optimization problem. An analytical solution to the problem is possible with the appropriate choice of the cost function as the limiting condition. Numerical calculations [2] show that usually there is not much difference between the cases of using value functions of one kind or another. This means that ones should choose the cost function, which most fully corresponds to the conditions of a specific problem. Consider the cost function [10] for the i -th elementary module of hyperconverged network when streaming requests:

$$D_i = \nu \cdot \frac{F_i}{V_i}, \quad (9)$$

where ν is a normalizing factor, L is the average request size in bits, and for streaming requests:

$$F_i = L\lambda_i,\ V_i = L\mu_i\ . \quad (10)$$

Then the function of the total cost of transferring the amount of information per unit of hyperconverged network throughput takes the form of

$$D = \nu\sum_{i=1}^{k}\rho_i D = \nu\sum_{i=1}^{k}\rho_i \quad (11)$$

and is expressed in units of the cost of information unit transmission, that is, the information flow density.

Thus, the optimization task can be formulated in this way: calculate the values of the information flow density that minimizes the average delay

$$\overline{T} = \frac{1}{\gamma}\sum_{i=1}^{I}\rho_i\frac{\Sigma'_{m_i}}{\Sigma_{m_i}} \to \min \quad (12)$$

with a limitation on transmission cost of the total amount of information per unit of communication link capacity

$$D = \nu\sum_{i=1}^{I}\rho_i \le D_{req}\ . \quad (13)$$

III. Method for message transmission time minimization in hyperconverged network environment

To solve problem (12) - (13), the method of indefinite Lagrange multipliers was applied.

Let us compose the optimization functional:

$$\Phi = \frac{1}{\gamma}\sum_{i=1}^{I}\rho_i\frac{\Sigma'_{m_i}}{\Sigma_{m_i}} + P\cdot\nu\sum_{i=1}^{I}\rho_i\ , \quad (14)$$

where P is an indefinite Lagrange multiplier.

Calculating partial derivatives

$$\frac{\partial\Phi}{\partial\rho_i} = 0\ ,$$

we obtain a system of I equations of the following form:

$$\left(\rho_i\frac{\Sigma'_{m_i}}{\Sigma_{m_i}}\right)' + \gamma P\nu = 0\ ,\ \ i = \overline{1,I}\ . \quad (15)$$

Analysis of expression (15) shows that each equation of this system depends on ρ_i variable and m_i, γ, P, ν parameters.

If we assume the same $m_i = m$ for all nodes (which, in principle, is permissible for a hyperconverged network), then these parameters will not depend on i index, i.e. as a result of its solution with respect to ρ_i, we obtain $\rho_i = F(m, \gamma, P, \nu)$. This allows us to conclude that for a hyperconverged network $\rho_i^{onm} = \rho = const$, i.e. the optimal values of the information flow densities are the same for all branches and do not depend on the number of the communication branch (isotropic network).

After differentiation and some transformations, we obtain a second-order differential equation for each branch. Omitting i index, we have

$$\frac{\Sigma'_m}{\Sigma_m} + \rho\frac{\Sigma''_m}{\Sigma_m} - \rho\frac{\left(\Sigma'_m\right)^2}{\Sigma_m^2} + \gamma P\nu = 0\ . \quad (16)$$

By changing the variable

$$\frac{\Sigma'_m}{\Sigma_m} = Z \quad \text{and} \quad \Sigma''_m = Z'\Sigma_m + Z\Sigma'_m \quad (17)$$

equation (16) is transformed into an inhomogeneous linear equation of the 1st order

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

$$Z' + \frac{1}{\rho} Z = -\frac{1}{\rho} \gamma P \nu . \quad (18)$$

The general solution of equation (18) is found by the method of variation of an arbitrary constant [11]. The corresponding homogeneous equation is

$$Z' + \frac{1}{\rho} Z = 0 , \quad (19)$$

with separable variables has a general solution in the following form:

$$Z = \frac{a_1}{\rho} . \quad (20)$$

Suppose $a_1 = a_1(\rho)$, i.e some continuously differentiable function of ρ, then

$$Z = \frac{a_1(\rho)}{\rho} . \quad (21)$$

Let us choose a function of $a_1(\rho)$ so that expression (21) satisfies equation (18). Substitute (21) into (18). Then, after transformations, we obtain

$$a_1(\rho) = \gamma P \nu . \quad (22)$$

Integrating (22) over ρ, we have

$$a_1(\rho) = -\gamma P \nu \rho + a_2 \quad (23)$$

and therefore

$$Z = -\gamma P \nu + \frac{a_2}{\rho} . \quad (24)$$

Returning to the old variable (18), we get:

$$\frac{\partial \Sigma_m'}{\Sigma_m} = -\gamma P \nu + \left(a_2 / \rho\right) . \quad (25)$$

Separating the variables in expression (25), we get the following equation:

$$\frac{d\Sigma_m}{\Sigma_m} = -\gamma P \nu \, d\rho + a_2 \frac{d\rho}{\rho} , \quad (26)$$

integrating which, we finally obtain

$$\Sigma_m = a_3 \rho^{c_2} \cdot e^{\gamma P \nu \rho} , \quad (27)$$

where a_3 is an integration constant of the 2nd quadrature.

Using equation (27), we find arbitrary constants of integration (a_2 and a_3) by solving the Cauchy problem for given initial conditions.

In further calculations, we restrict to dependence (27), from which ρ values for each branch of the considered network are determined:

$$\rho = \frac{a_2}{\gamma P \nu + \left(\Sigma_m' / \Sigma_m\right)} . \quad (28)$$

Let us determine a_2 values from the initial condition of $\rho_o = 1$ and the equation (7):

$$\begin{aligned} \left(\Sigma_m' / \Sigma_m\right)\Big|_{\rho=1} &= \frac{1+2+\cdots+(m+1)}{m+2} = \\ &= \frac{(m+1)(m+2)}{2(m+2)} = \frac{m+1}{2}\left(\Sigma_m' / \Sigma_m\right)\Big|_{\rho=1} = \\ &= \frac{1+2+\cdots+(m+1)}{m+2} = \frac{(m+1)(m+2)}{2(m+2)} = \frac{m+1}{2} . \end{aligned} \quad (29)$$

Expression (29) takes into account that the numerator

$$1 + 2 + \ldots + (m_i + 1) = \frac{(m+1)(m+2)}{2} \quad (30)$$

is a sum of arithmetic progression.

From equation (23) under the condition (24) we define the arbitrary constant of a_2:

$$a_2 = \gamma P \nu + \frac{m+1}{2} . \quad (31)$$

Finally we get:

$$\rho = \left(\gamma P \nu + \frac{m+1}{2}\right) \Big/ \left(\gamma P \nu + \left(\frac{\Sigma_m'}{\Sigma_m}\right)\right) . \quad (32)$$

To determine the indefinite Lagrange multiplier, we use condition (13) for the limiting cost value:

$$\nu \sum_{i=1}^{I} \frac{\gamma P\nu + (m+1)/2}{\gamma P\nu + \left(\Sigma_m^{'}/\Sigma_m\right)} = \\ = \nu \cdot I \cdot \frac{\gamma P\nu + (m+1)/2}{\gamma P\nu + \left(\Sigma_m^{'}/\Sigma_m\right)} = D_{req}. \quad (33)$$

After transformations, we obtain the value of P Lagrange multiplier:

$$P = \frac{(m+1)/2 - \left(\Sigma_m^{'}/\Sigma_m\right) \cdot \frac{D_{req}}{\nu \cdot I}}{\gamma\nu\left(\frac{D_{req}}{\nu \cdot I} - 1\right)}. \quad (34)$$

Substituting (32) into (34), we obtain the conditions for the extrema $\bar{T}_i$ of expression (7):

$$\left(\rho_{opt} - \frac{D_{req}}{\nu \cdot I}\right)\left(\frac{m+1}{2} - \Sigma_m^{'}/\Sigma_m\right) = 0. \quad (35)$$

Conditions (35) are satisfied if any of their factors is equal to zero, that is

$$\rho_{opt} - \frac{D_{req}}{\nu \cdot I} = 0; \quad (36)$$

$$\Sigma_m^{'}/\Sigma_m - (m+1)/2 = 0. \quad (37)$$

Condition (36) determines the optimal value of the flow in the branches

$$\rho_{opt} = \frac{D_{req}}{\nu \cdot I}, \quad (38)$$

providing the lowest average latency:

$$\bar{T}^{\min} = \frac{1}{\gamma} \cdot \frac{D_{req}}{\nu} \cdot \left(\frac{\Sigma_m^{'}}{\Sigma_m}\right)_{opt}, \; 0 < \rho_{opt} < 1. \quad (39)$$

Condition (37) corresponds to the maximum value of the network delay under $\rho = 1$:

$$\bar{T}^{\max} = \frac{I}{\gamma} \cdot \frac{m+1}{2}, \quad (40)$$

which is achieved regardless of the cost of the network.

IV. Discussion and Conclusion

The results obtained in the previous section show that for the minimization of the average delay during HCN's horizontal scaling it is necessary to choose an isotropic expansion. This approach ensures a constant density of information flow, which is served by all new hardware modules, i.e.

$$\rho_{opt} = \frac{D_{req}}{kn} < 1. \quad (41)$$

If the flows in the branches during the horizontal scaling of hyperconverged network synthesis are given in the form of $\|\lambda_i\|$ gravitational matrix, the capacities of the corresponding branches are directly proportional to the values of the flows of these branches, that is

$$V_i = \frac{\nu \cdot I}{D_{req}} \cdot F_i, \quad (42)$$

which is a necessary condition for eliminating network blockages (that is $V_i > F_i$) and the degree of this excess is determined by the ratio of the number of network branches to their cost.

The assumption that m does not depend on the number of a node or branch is valid, since according to (42), an increase in F_i flow leads to the need for a proportional increase in throughput. This, in turn, leads to faster freeing of buffers, so that the number of requests at the entrance to each link remains unchanged and the required number of buffers remains constant.

To select the characteristics of a standard HCN extension module that satisfy the QoS requirements for the minimum average latency and the probability of denial of service, a combined graph of the dependences of such values on module utilization coefficient and required number of buffers is proposed (Fig. 1).

Appropriate delay in a non-buffered system provides a lower bound for delay in a wide variety of multiple access systems with buffering and flow control. However, the latency in a non-buffered system provides a lower bound on latency for a wide variety of buffered, flow controlled multiple access systems. However, limiting the number of buffers in switching nodes inevitably leads to the fact that the node rejects some of the packets. To take this into account, Fig. 1 shows the combined curves of the dependence between $P(\rho)$ probability of failure and $\bar{T}^{\min}(\rho)$ function for the same m values of the number of buffers, constructed in accordance with the following expression:

$$P_{fail}(\rho, m) = \rho^{m+1} \Big/ \sum_{i=0}^{m+1} \rho^i. \quad (43)$$

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

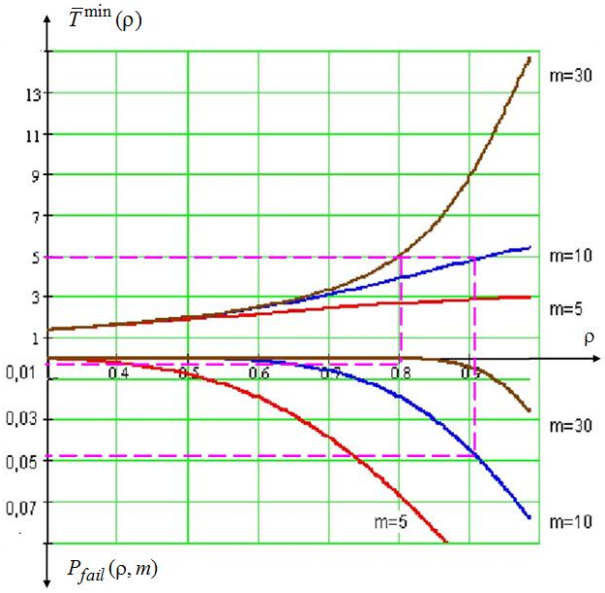


Figure 1. Selection of module characteristics for HCN horizontal scaling

Let us consider an example of choosing the characteristics of a module. The analysis of the curves allows us to conclude that the minimum average delay $\bar{T}^{\min}(\rho) = 5$ s, that corresponds to the values of $\rho'_{opt} = 0,8$ and $m = 30$, can be achieved with a higher density of information flow transmitted over the network ($\rho'_{opt} = 0,9$ and $m = 10$) with a threefold decrease in the number of buffers in the switching nodes. In this case, the probability of failure increases from $P_{fail}(\rho, m) = 0.015$ to $P_{fail}(\rho, m) = 0.045$. If this probability is acceptable, then this method will allow choosing cheaper modules.

So, the proposed horizontal scaling method for a hyperconverged network allows determining the characteristics of additional hardware modules that, in turn, will provide minimum average delay in message transmission while maintaining acceptable levels of costs for additional hardware and the probability of denial of service.

ACKNOWLEDGMENT

The presented study is partly supported by the grant of the Slovak Research and Development Agency SK-SRB-18-0002

REFERENCES

[1] Merlac, V., Smatkov, S., Kuchuk, N. and Nechausov, A. “Resourses Distribution Method of University e-learning on the Hypercovergent platform”, *Conf. Proc. of 2018 IEEE 9th Int. Conf. on Dependable Systems, Service and Technologies*, DESSERT’2018, Kyiv, 2018, pp. 136-140, doi: https://dx.doi.org/10.1109/DESSERT.2018.8409114.

[2] Kuchuk, N., Gavrylenko, S., Lukova-Chuiko, N. and Sobchuk, V. “Redistribution of information flows in a hyperconvergent system”, *Advanced Information Systems*, Vol. 3, No. 2, 2019, pp. 116-121, doi: http://dx.doi.org/10.20998/2522-9052.2019.2.20.

[3] Shmatkov S.I., Kuchuk, N.G. and Donets V.V. “Model of information structure of the hyperconvergent system of support of electronic computing resources of university e-learning”, *Control systems, navigation and communication*, PNTU, Poltava, No. 2 (48), 2018, pp. 97-100, doi: https://dx.doi.org/10.26906/SUNZ.2018.2.097.

[4] Kuchuk, N., Nechausov, A. nd Mamusuc, I. “Synthesis of the air pollution lever control system on the basis of hyperconvergent ifrustructures”, *Advanced Information Systems*, Vol. 1, No. 2, 2017, pp. 21-26, doi: https://dx.doi.org/10.20998/2522-9052.2017.2.04.

[5] Kianpisheh, S. and Glitho, R.H. “Cost-efficient server provisioning for deadline-constrained VNFs Chains: A parallel VNF processing approach”, *Proceeding of 2019 16th IEEE Annual Consumer Communications & Networking Conference*, 2019, doi: https://doi.org/10.1109/CCNC.2019.8651799.

[6] Kovalenko, A., Shamraev, A., Shamraeva, E., Dovbnya, A. and Ilyunin, O. “Green Microcontrollers in Control Systems for Magnetic Elements of Linear Electron Accelerators”, *Green IT Engineering: Concepts, Models, Complex Systems Architectures. Studies in Systems, Decision and Control series*, Springer International Publishing Switzerland, 2017, pp. 283-305, doi: https://dx.doi.org/10.1007/978-3-319-44162-7_15/

[7] Lee, S. R. “Dispersion-managed links formed of SMFs and DCFs with irregular dispersion coefficients and span lengths”, *Journal of Information Communication Convergence Engineering*, vol. 16, no. 2, 2018, pp. 67-71, doi: https://dx.doi.org/10.6109/jicce.2018.16.2.67.

[8] Franti, P. Efficiency of random swap clustering, *Journal of Big Data*, vol. 5, is. 13, 2018, pp. 1–29, doi: https://doi.org/10.1186/s40537-018-0122-y

[9] Ye, Q. and Zhuang, W., “Distributed and adaptive medium access control for internet-of-things-enabled mobile networks”, IEEE Internet of Things Journal, vol. 4, no. 2, pp. 446-460, 2017, doi: http://doi.org/10.1109/JIOT.2016.2566659

[10] Tkachov, V., Hunko, M. and Volotka, V. “Scenarios for Implementation of Nested Virtualization Technology in Task of Improving Cloud Firewall Fault Tolerance”, *2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology* (PIC S&T), Kyiv, Ukraine, 2019, pp. 759-763, doi: https://doi.org/10.1109/PICST47496.2019.9061473.

[11] Lima, A.A., de Barros, F.K., Yoshizumi, V.H., Spatti, D.H. and Dajer, M.E. “Optimized artificial neural network for biosignals classification using genetic algorithm”, *Journal of Control, Automation and Electrical Systems*, 30(3), 2019, pp. 371–379, doi: https://doi.org/10.1007/s40313-019-00454-1.

[12] Raskin, L., Sira O., and Parfenyuk, Y., “Selection of the optimum route in an extended transportation network under uncertainty”, *Advanced Information Systems*, Vol. 5, No. 1, 2021, pp. 62-68, doi: https://dx.doi.org/10.20998/2522-9052.2021.1.08.

[13] Donets, V., Kuchuk, N. and Shmatkov, S., Development of software of e-learning information system synthesis modeling process, *Advanced Information Systems*, Vol. 2, No. 2, 2018, pp. 117-121, doi: https://doi.org/10.20998/2522-9052.2018.2.20.

[14] Kianpisheh, S. and Glitho, R.H., “Cost-efficient server provisioning for deadline-constrained VNFs Chains: A parallel VNF processing approach”, Proceeding of 2019 16th IEEE Annual Consumer Communications & Networking Conference, 2019, doi: https://doi.org/10.1109/CCNC.2019.8651799.

[15] Mukhin, V., Kuchuk, N., Kosenko, N., Kuchuk, H. and Kosenko, V. “Decomposition Method for Synthesizing the Computer System Architecture”, *Advances in Intelligent Systems and Computing*, AISC, vol. 938, 2020, pp 289-300, doi: https://doi.org/10.1007/978-3-030-16621-2_27.

[16] Chen, Y., Ran, R. and Oh, S. “Performance analysis for pilot decontamination of massive MIMO using alternative projection”, *Proceedings of IEEE International Conference on Ubiquitous and Future Networks*, Vienna, Austria, 2016, pp. 297–300, doi: https://doi.org/10.1109/ICUFN.2016.7537036.