Usage of Light-Emitting Diode Lighting and Visible Light Communication Technology for Temperature Control

Svetlana Grigoryeva, Alexander Baklanov, Aslima Alimkhanova

D. Serikbayev East Kazakhstan Technical University,
Faculty of Information Technology and Intelligent Systems
A. K. Protazanov Str. 69, 070004 Ust-Kamenogorsk, Kazakhstan
e-mail: {SGrigorieva, ABaklanov}@ektu.kz

Alexander Dmitriev

Novosibirsk State Technical University, Physics and Technology Faculty
20 Prospekt K. Marksa, 630073 Novosibirsk, Russia
e-mail: alexander_dmitriev@ngs.ru

György Györök

Óbuda University, Alba Regia Technical Faculty
Budai út 45, H-8000 Székesfehérvár, Hungary
gyorok.gyorgy@amk.uni-obuda.hu

Abstract: A new control system via light-emitting diode (LED), lighting devices, is reviewed in this article. In order to transmit data, Visible Light Communication technology was used. Function block diagrams and schematic diagrams, of transmitting and receiving optical channels were used for the temperature control. Implementation of the developed system has been accomplished by the regulation of the onsite temperature control. Serial interface of data transmission has been applied in order to provide stable and reliable link via physical channels with the usage of LED lighting devices. According to the protocol of this interface, the transmitted data transforms into consecutive impulses in the accordance with the UART standard. The present work shows that asynchronous data transmission has been carried out via microcontrollers. In order to conduct the experiment presenting controlling possibilities via white LED lighting, a breadboard prototype with two modules – transmitter and receiver – has been created. The conducted measurements have shown that optical channel of data transmission has high reliability, while the illumination level does not change. There is no flickering and a comfortable lighting regime is provided.

Keywords: control system; data transmission; lighting devices; LED; Visible Light Communication; temperature

1 Introduction

Currently, modern control systems obtain information, in order to monitor and manage the physical processes from sensors and actuation devices, which are spatially remote from one another. Classical data transmission between modules of managing system is carried through wires. The necessity of the physical connection of sensors and actuation mechanisms, limits their flexibility, scalability and reliability of the working system. Switching to wireless communication systems will allow for reduction of costs on assembly and failure risks during operation, the risk which will lead to the generation of a cost effective and reliable system, as well as, provision of an opportunity of the organization of intellectual infrastructure. For example, popular management platforms such as, “Internet of Things”, “SMART House”, “Device to Device” and “Machine to Machine” use wireless technologies to build their infrastructure.

However, with the increase in the number of mobile users and devices connected to the Internet, problems arise in ensuring communication, data exchange and the fast acquisition of information. Most international analytical agencies give high marks to the growth in the number of connected devices in the world. In 2016 during the “Internet of Things World” conference, which was held in Santa Clara (USA), company SigFox quoted the predicted data from different companies. The values differed by the order of magnitude. For example, “Gartner” company’s analytics claimed that the amount of connecting devices would reach 21 billion by 2020, while Intel company’s specialists predicted the value of 200 billion devices for the same year [1]. According to the report of Cisco Visual Networking Index, the seven times increase of volume of mobile data transmission is being expected in the period since 2016 to 2021, and the amount of mobile devices per capita would reach 11.6 billion to 2021 [2]. In May 2019, Strategy Analytics company’s experts published data on Internet connections of 22 billion devices [3]. Presented estimations, as well as, real data shows the necessity of the implementation of network based architectures, with high through-put, which will be able to satisfy rising exigency in the resource of wireless network.

Many types of technologies and standards have been developed, depending on the purpose and requirements for wireless systems. Technologies that transmit data in the radio frequency range have become widespread. The growth in service consumption in wireless networks has led to a shortage of radio frequency resources. To meet the growing demand, either an increase in bandwidth or an increase in spectral efficiency should be used. However, the increase in spectral efficiency is slow and cannot meet the rapidly growing demand. In parallel with the development of technology in the radio frequency field, there is a great potential for use in the optical range.

Additional usage of optical frequencies allows for the solution of problems related to the spectrum crunch, in wireless connections, on the basis of radio frequency. In Optical Wireless Connections (OWC) it is possible to use three main frequency bands – ultraviolet, infra-red and visible-light. Within the ambit of the last two bands, the link by means of visible-light (Visible Light Communications, VLC), wireless optics (Free Space Optics, FSO) and the link by means of optical camera (Optical Camera Communications, OCC) are possible [4] [5]. OWC technology has a variety of unique advantages such as broad spectrum, high data transmission speed, low delay, high security, immunity to radio frequent electromagnetic disturbance, available licensure, low cost and low power consumption [5-8]. It may be noted that there are publications which present comparison of OWC technologies on various aspects more carefully [9-13].

With the advent of white LEDs used for lighting, Visible Light Communication technology is rapidly developing. This technology allows a light source, in addition to lighting, to transmit information using the same light signal. Visible range of wavelength is 370-780 nm provides transmission capacity of ~400-800 THz, which is ten thousand times more than in the radio frequency bandwidth [5] [6] [11]. The first demonstration of wireless data transmission through a flashing LED was demonstrated by Harald Haas, a professor at the University of Edinburgh, during the TED World Conference in Edinburgh, Scotland in 2011 [14]. Despite the fact that the technology has been developed and researched over the last 10 years, the transmitting systems with the data transmitting speed of a few Gb/s have already been demonstrated [15-17]. Scientists claim data transfer rates from 20 Mbps to 40 Gbps, depending on the design of the LED, which is confirmed by the research results [18].

In addition to the listed advantages of OWC technologies, it is necessary to note the attractive aspects of VLC. For the deployment (organization) of the VLC system, there is a ready-made infrastructure of transmitting devices - LED lamps. LED lighting technology supports the concept of green energy and uses energy efficient lighting sources.

In the long term, the VLC systems can be used in the wide range of applications. There are problems which limit the implementation of the VLC technology. In article [7], two main tasks are identified that need to be solved - elimination of LED flickering during data transmission and support for dimming in the lighting system. These problems are caused by the simultaneous use of luminaires directly for lighting, and also additionally for data transmission. The problems associated with the optical properties of light sources are also highlighted: high losses in the path; intersymbol interference caused by multipath; interference caused by artificial light; nonlinearity of the electro-optical response of the LED. Article [19] discusses issues related to uplink transmission and integration into the existing information and communication structure. The articles [13] [20] provide an overview of the main problems arising in the design of communication using VLC technology. Light source problems (flicker, blackout, line of sight and interference) and wireless communication problems (uplink and mobility) are discussed.

Today VLC technology using LED lighting is in the research and development stage. For implementation in the field of communication and wide accessibility of users, it is necessary to solve many problems. The authors have published the results of the developed systems for the transmission of audio signals [21] and symbolic data [22]. The present work suggests a new idea to use the technology of signals transmission by means of LED lighting in automated control systems by physical parameters. Previously, the authors have developed and presented [23] an automated LED lighting control system on the basis of which the implementation of VLC technology is proposed.

2 The Construction of the Temperature Control System with the Usage of the VLC Technology

Automated control system actuator, heater, indoors, depending on the required temperature, has been suggested and developed, on the basis of the VLC technology. Structural scheme is shown in Fig. 1. The objectives of the control system are to measure the temperature in the room, control the heating element, and maintain the set temperature in the room. Temperature sensor is located in LED device which is also information transmitter – in the given case – current temperature. Actuator of the temperature regulator and information receiver are located remotely from the signal transmitter. Constructively, the transmitter and the temperature sensor are one device, while the receiver and the actuator are another device.

Fig. 2 shows the scheme of transmitting device. The temperature sensor is used for the temperature control, which allows to transmit measured data to the microcontroller. The values of the current temperature are shown on the indicator. The signal received from the microcontroller output is added in order to transmit data by means of the VLC technology to the power supply of LED. This signal is formed in accordance with the UART technology (Universal Asynchronous Receiver-Transmitter). Since the reliable work of LED requires the usage of constant current [24], a current stabilizer, has been added to the power supply. Thus, in order for LED to transmit data stably, the switch has been used, which made it possible to connect proceeding signals, by the UART and supply the constant current for the LED. The LED in the optical channel has alternating light in accordance with the signal proceeding from the microcontroller.

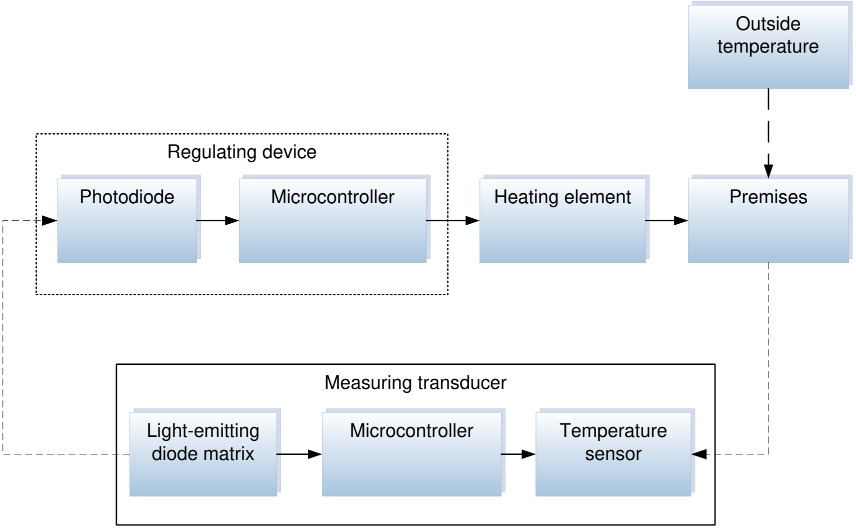


Figure 1

Structural scheme automated control system indoors temperature based on VLC technology

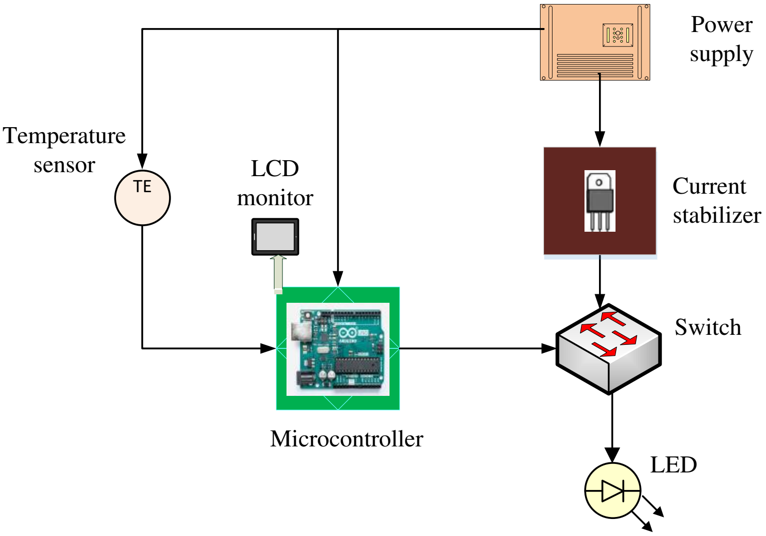


Figure 2

Transmitter function block diagram

Fig. 3 shows the scheme of the receiving device. The receiver's task is to transform light pulses received from emitter into electric signals. These signals by passing through amplifier are transferred to the shaper where they are turned into impulses of the TTL form (Transistor-Transistor Logic) for the further treatment by microcontroller. Microcontroller deciphers the information and processes the algorithm for the control of actuator depending on the assigned task. The received data on the current temperature is shown on the indicator.

Schematic diagrams have been developed in accordance with the function block diagrams.

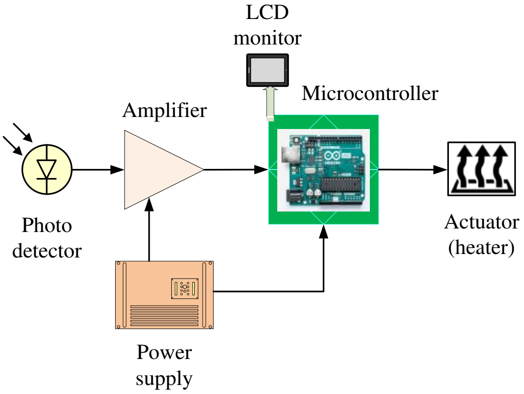


Figure 3
Receiver function block diagram

The schematic diagram of the transmitting device, is presented in Fig. 4. The transmitter's concept resides in the fact that the temperature sensor, with the transmitter, is inboarded with the LED illuminator. The transmitter sends the data of the temperature, by means of the LED of illuminator.

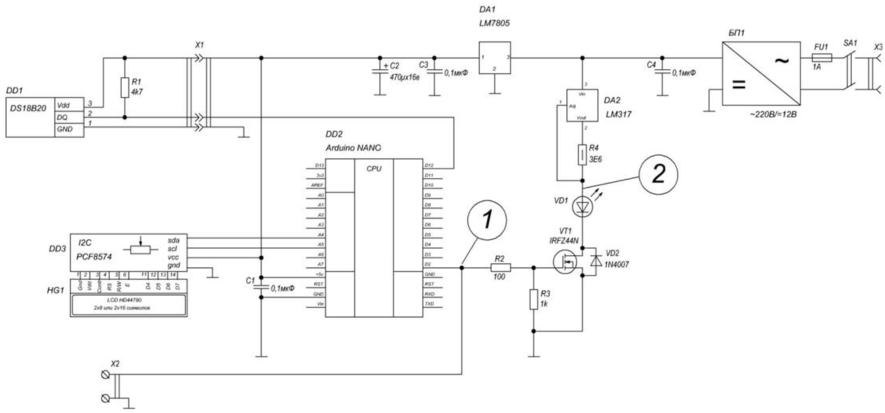


Figure 4
Transmitter circuit diagram

The values, upon request, are transmitted from the DD1 temperature sensor to the DD2 microcontroller, where they are recorded in temporal storage registers, processed and afterwards, displayed on the HG1 indicator. The information necessary for the monitoring and control is also displayed on the HG1 indicator. The value of temperature from the temporal storage registers after transformation proceeds into the FIFO buffer of the transmitter-receiver UART of the microcontroller, where, by the means of consequent asynchronous data transmission, bits proceed to the VT1 transistor gate through the TX microcontroller output and R2 resistor. A transistor switches, from allowing, to forbidding, the current flow through LED VD1, by working in key mode according to the potential on the gate.

The DS18B20 integral temperature sensor has been used as the DD1 temperature sensor. The required drivers and connecting programs for this sensor are publicly available. The R1 resistor with the nominal value of 4.7kΩ is recommended by the manufacturer of DS18B20 microchip.

ATmega328P microchip, which is a part of the Arduino NANO and Arduino Uno boards, has been chosen as the DD2 microcontroller.

The HG1 indicator is liquid-crystal indicator LCD1602 with the I^{2}C module.

IRFZ44N has been chosen as the VT1 transistor, but it is possible to use any n-channel MOSFET designed for the usage of light emitting diode VD1. The VD2 diode is necessary in order to shunt the transistor from current impulse, which occurs due to power-cut. This diode is located inside the IRFZ44N transistor case.

The R2 resistor in the gate circuit limits the current at transistor opening. The R3 resistor is connected to the gate in order to have reliable and fast transistor closure. It eliminates the possibility of the occurrence of the unconfigured state of the transistor and as a consequence will eliminate the presence of random errors in transmitter which might have occurred for that reason.

The VD1 - white LED powerful current of 700 mA which has possibility of installation on the radiator cooling system.

The DA2 microchip - integrated stabilizer LM317 is included in the current stabilization circuit. The R4 resistor is a shunt for the stabilizer. In our case, the R4's resistance is 3.6 Ω.

The DA1 microchip - an integral stabilizer LM7805 of fixed voltage of 5 V provides power supply for microprocessor's board, indicator and temperature sensor.

The power source PS1provides power supply from alternating current system of ~230 V, 50 Hz. It is impulse and produces the constant regulated voltage of 12 V and the current of 1.5 A.

C1, C3, C4 are filters capacitors, and C2 is a bulk capacitor.

Schematic diagram of the receiving device is presented in the Fig. 5.

The received light signal is transformed by the VD1 photodiode into photo current and through the DA1 amplifier-convertor of the current into voltage and through the C2 separating capacitor it proceeds to the DA2 amplifier's input. Afterwards it proceeds from the DA2 amplifier's output through the C3-R6 differentiating circuit to the DD1 logical gate's input, which performs the duty of buffering element and simultaneously inverts the signal for the further processing. Inverting the signal is necessary in order to bring the waveform of the required polarity to the UART transmitter-receiver, since photodiode connected with the reverse polarity generates inverse current impulses. The signal proceeds from the inverter's output to the UART transmitter-receiver input of DD2 microcontroller, where it is decrypted. The received data is recorded into the registers of temporary variables storage, where it afterwards proceeds for further processing and displaying onto the liquid-crystal sensor. The SB1-SB3 buttons are used for the changes of temperature regulation. Further data processing depends on which data is being transmitted and what it is used for. In our system, we transmit the value of the temperature to regulator – heating element EK1.

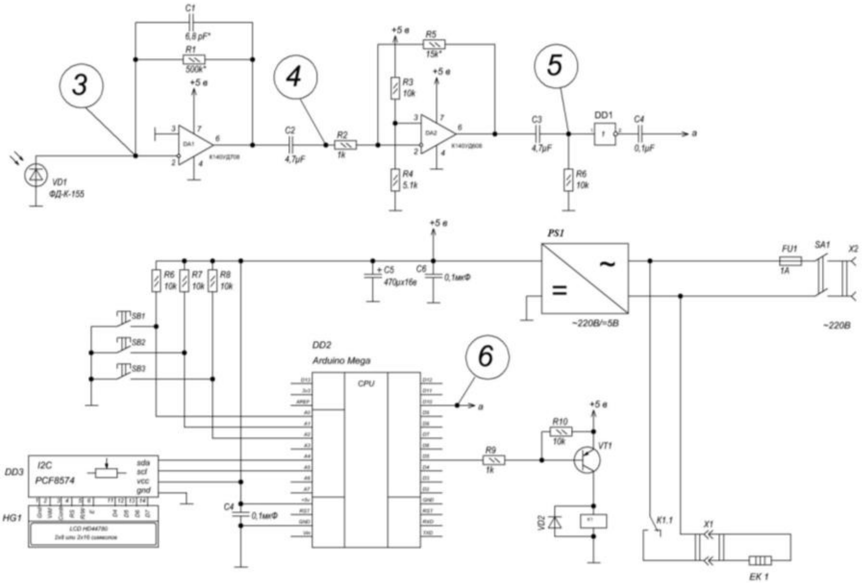


Figure 5

Receiver circuit diagram

The FD-K-155 photodiode has been chosen as the VD1 photodiode. The DA1 microchip has been developed on the basis of the KR140UD708 operational amplifier, while the DA2 has been developed on basis of the KR140UD608 operational amplifier.

The C1 condenser connected in parallel to the R1 resistor of negative circuit reentrancy is necessary in order to reduce the gain of high-frequency noise. Its nominal has to be matched specifically carefully. The R1, R2 and R5 resisters affecting the gain coefficient of operational amplifiers, as well as, the C2 and C3 condensers, which define the steepness of pulse front and droop, have been matched so that at the rate of 9600 baud, the form of the signal in the control points 4, 5 and 6 is approximated the most closely to the TTL impulses.

It is necessary to control a backhaul and a signal format whilst adjusting the receiving part of the device. Points for the recording of oscillograms, which are necessary for the research and calibration of the whole system in data transmission mode have been marked 1-6 in the schematic diagram.

3 The Experimental Results Settings of the Receiver

We have carried out research in order to optimize the system's values. Oscillograph has been controlling the presence and the form of the signal during the data transmission. The level and the form of the output signal from transmitter's microcontroller and the signal transmitted onto the receiver's microcontroller are crucial. The most crucial points of control are marked on the schematic diagrams of the transmitter and the receiver as markers 1-6. The signal proceeding from the transmitter's microcontroller is a pattern of the form and amount of impulses (marker 1). Oscillogram at this point is controlled by the second beam of oscillograph and is present at all oscillograms of control points for synchronization. All other oscillograms of control points are marked by the first beam of the oscillograph.

The signal to LED (marker 2) proceeds from transistor, which is the switch of the transmitting part. The control of this point provides the information on the switch working capacity. In case of the failure it is no longer able to be opened or closed or it happens incompletely. The form of the signal is shown in Fig. 6.

It is crucial to control the signal's level after all electrical components until entering the microcontroller during the work of the receiving device. The control of the signal after photodiode (marker 3) shows the working capacity of the photo current source – in our case the FD-K-155 photodiode of the receiving part of the apparatus. The oscillogram (Fig. 7) shows that the amplitude of the signal on the photodiode is very low (0.8 V), given the oscillogram was taken when the LED of the transmitter and photodiode were almost immediately adjacent. Along with the increasing of the distance between LED and photodiode the amplitude of the signal was decreasing significantly. Therefore, for the further work with the signal of such amplitude the amplifier is required.

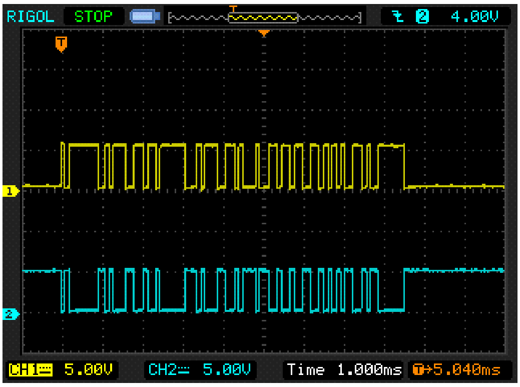


Figure 6
Waveform of the signal from the transmitter

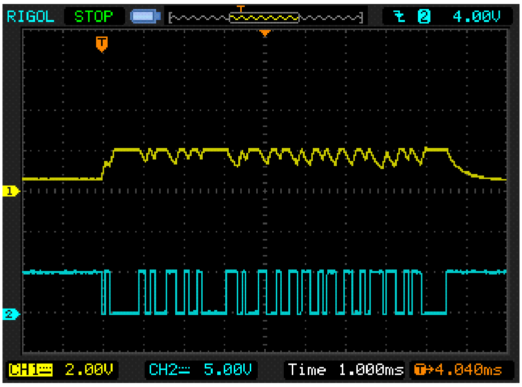


Figure 7
Waveform of the signal on the photodiode of the receiver

The measurement of the signal for the work adjustments of the first cascade of photo current amplification is included. The form of the signal at this point (marker 4) can be seen in Fig. 8. According to the form of the signal at this point, you can adjust the gain coefficient of the first stage on the K561LN2 microchip, by selecting a feedback resistor. The diagram has a condenser shunted to the ground and is meant for the elimination or the minimizing of the high-frequency noise, at the output of the first cascade of amplification.

According to the oscillogram's signal after the second cascade of amplification (marker 5) the coefficient of amplification at the second K561LN2 microchip is adjusted by the resistor selection. The form of the signal is shown in Fig. 9.

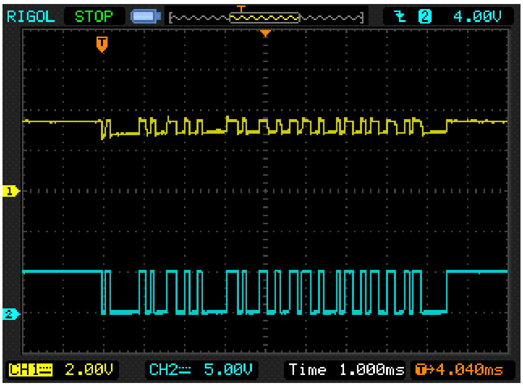


Figure 8
Waveform of the signal after the first cascade of receiver’s amplification

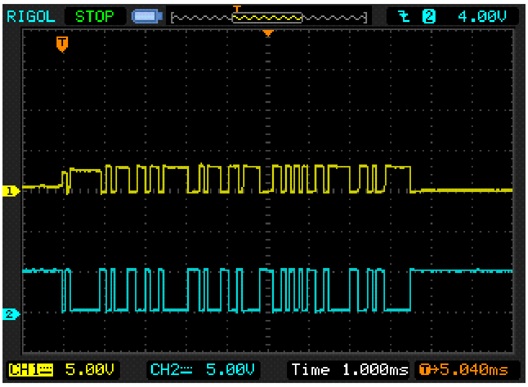


Figure 9
Waveform of the signal after the second cascade of the receiver’s amplification

After the buffer amplifier-inverter (marker 6), the signal that almost completely coincides with the etalon, can be seen. The form of the signal at this point is shown in Fig. 10.

Adjustment of all parts of the receiver’s scheme has been conducted with the synchronization of the signal through the second channel of the oscillograph (etalon one) in order to see each impulse separately during transmission. Adjusting cascades of the amplifier at each control point, we have achieved similar form of the signal and impulse length which provides stable work of the entire system. Identity of the transmitted and received signals ensures reliability and accuracy of the system’s performance.

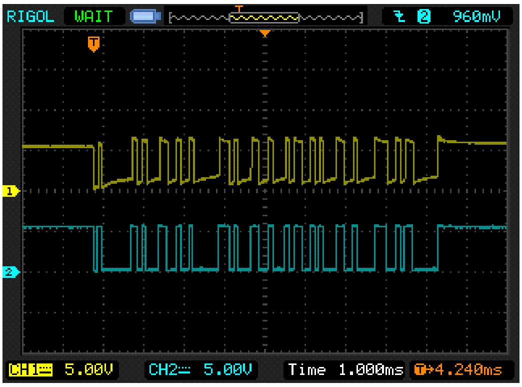


Figure 10
Waveform of the signal after the buffer amplifier-inverter of the receiver

4 Test Environment of Automated Control System of Temperature

Fig. 11 shows structural diagram of the test installation, of automated control system of a heater, in accordance with VLC technology, which has been developed on the basis of function blocks and schematic diagrams.

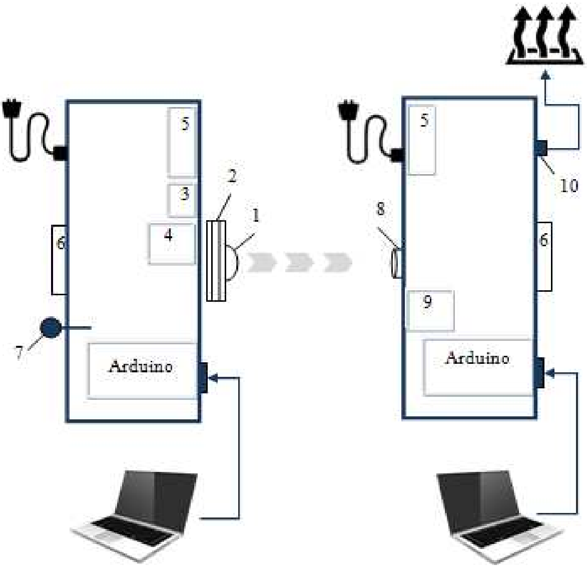
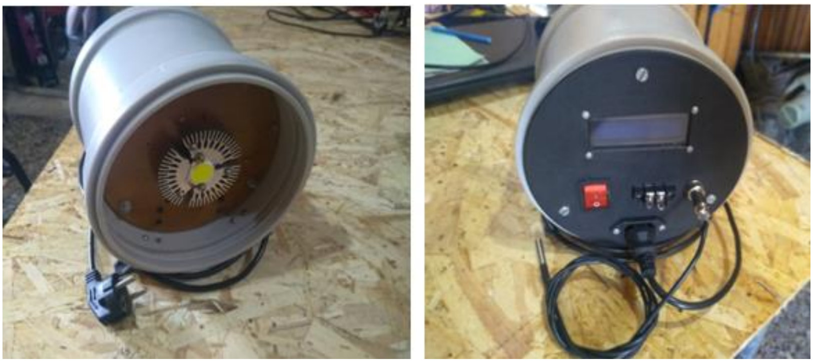


Figure 11
Structural diagram of test installation

Power of LED matrix (1) is 3 W and constant current of 700 mA is attached to the aluminum radiator (2). The radiator is installed with the gap on the plate of getinaks. The radiator has been designed and manufactured, so that there wouldn't be any overheating of the crystal's light-emitting array, due to continuous performance [25]. The gap is necessary for providing heat elimination, from the radiator. On the opposite side of the front panel, where the light emitting diode is placed, the current stabilizer (3), the commutator (4), the power source (5) and Arduino Nano board are suited. The sketch loading is provided through mini USB-port of the microcontroller.

The DS18B20 integral digital sensor of temperature manufactured by the Dallas Semiconductor company in the TO-92 (7) case, as well as, the LCD1602 with the I^{2}C (6) module for the monitoring of the control system values are attached to the front side of the back panel. Fig. 12 shows physical configuration of the transmitting device.

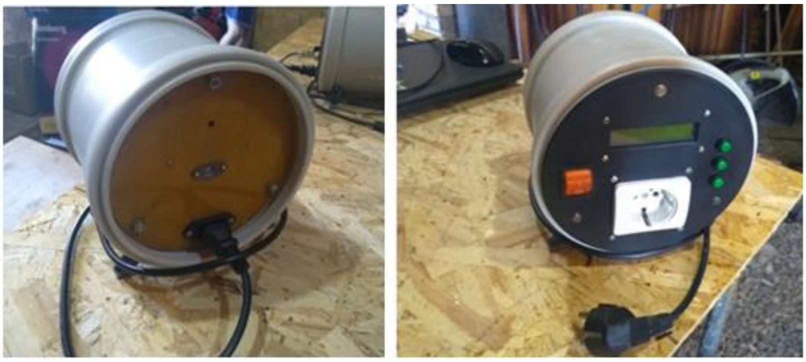


a) view of the working panel, b) view of the back panel

Figure 12
Physical configuration of the transmitting device

The FD-K-155 photodiode (8), which is attached to the external panel, is used for the receiving of LED impulses. The amplifier developed on the basis of the K561LN2 microchip (9), power source (5) and Arduino Nano microcontroller are suited onto the interior side. In order to control the signal of the exterior back panel the LCD monitor (7), the output for heating source control (10), the buttons for the assignment of required values of temperature have been placed. Fig. 13 shows physical configuration of the receiving device.

Modulating methods used for radio-frequency communication [26-28] can be used in the VLC systems, as well as, specific modulating methods such as laser femtosecond systems [29] can be implemented.



a) view of the working panel, b) view of the back panel

Figure 13
Physical configuration of the receiving device

In order to organize the system of the temperature monitoring and heater control through optical wireless connection with LED lighting element, the UART (Universal Asynchronous Receiver-Transmitter) technology has been used. In our case in accordance with the present technology the microcontroller has been programmed.

During the data transmission the converting of the transmitted data into coherent manner is carried out in the way so that it is possible to transfer it through the digital line to another similar device.

The data transmission in the UART is carried out by one bit at a time in equal time length. This time length is defined by target velocity of the UART. The choice of the UART speed depends on the physical capabilities (parameters) of the electronic components of the receive/transmit path. Data transfer speed was not the main criterion. The task was to obtain reliable data on the ambient temperature, as well as to maintain the required level of illumination and the absence of LED flickering. For our particular connection, the baud rate is set to 9600 baud. This velocity has been chosen experimentally as the most optimal, for the systems electrical elements, for transmitting and receiving devices.

The control system uses two devices for data transmission - a receiver and a transmitter. Since each device is equipped with a microcontroller, two separate projects were developed.

Projects of programs for Arduino Nano microcontrollers were developed in the visual programming environment FLProg [30] using the graphical programming language FBD (Function Block Diagram).

When programming the receiver according to the functional and logical principle, five blocks were allocated: receiving data from the UART port, separating the integer and fractional parts of the temperature value, presentation of data on the LCD display, setting the temperature in manual mode, program for the controller.

To operate the transmitter, three block diagrams were written: a program for polling a temperature sensor, a program for operating the UART port, a program for presenting data on the LCD display.

Here, we describe the algorithm for the operation of the temperature control system, using the developed experimental stand.

1) Turn on the receiver and transmitter. The receiving device must be located opposite the transmitting device. In our case, the maximum value of the distance between the receiver and the transmitter is 0.80-0.85 m.
2) The user sets the preset temperature using the buttons on the receiver. The set temperature is displayed on the monitor.
3) The temperature sensor located on the transmitting device determines the value of the ambient temperature and transmits it to the microcontroller.
4) The microcontroller processes the data, displays the temperature value on the monitor and transmits it through the LED.
5) A photodiode located on the receiving device reads the signal and transmits it through an amplifier to the microcontroller.
6) The microcontroller processes the signal and displays the received temperature data on the monitor. If the obtained temperature value is less than the set value, then the heating element turns on and the monitor displays “heat on”. If the value of the measured temperature is higher than the set one, then the heating element is not switched on / off and the monitor displays “heat off”. The range of regulation of the set temperature value is set by software.

The suggested control system is useful for the regulation of other physical processes, which at the present, rely on the wire connection systems or wireless systems with the radio wave usage. The transition to wireless control of technological processes is especially crucial. For example, the work [31], considers the usage of LED lighting in agricultural field for optimal provision or substitution of natural light for the plant production. Present artificial lighting infrastructure can become the platform for the organization of intellectual control systems of value, without a large financial investment.

Conclusions

As a result of the conducted experiments, stable data transmission, by means of white LED lighting and the coding by means of UART technology, has been achieved. The obtained results have allowed to provide for the regulation of temperature indoors, by means of optical wireless communication, via a LED lighting system. The base of the developed system, are inexpensive ATmega328P microcontrollers. Such an approach makes wireless systems, which are based on the usage of LED devices, a good prospect. This being said, the main features of the suggested system, is the simplicity of construction and low production cost of the radio-electronic elements. The suggested control system provides fast transmission of the data needed for the current temperature and regulation.

The advantages of the manufacture of control systems, with VLC technology is the organization of the values regulation system, providing the work of a “Smart House” and the intellectual energetic systems. The implementation of such developments is stimulated by the exponential growth in the use of LED lighting fixtures.

References

[1] World experience and development prospects of the industrial Internet of things in Russia. Link: https://www.crn.ru/news

[2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021, White Paper, 2017, p. 35

[3] Internet of Things, IoT, M2M world market. Link: http://tadviser.com/index.php

[4] Visible light for broadband communications. Report ITU-R SM.2422-1. Geneva, 2019, Link: https://www.itu.int/pub/R-REP-SM.2422

[5] H. Haas, J. Elmirghani, I. White: Optical wireless communication, Royal Society, Vol. 378, Issue 2169, 2020, doi.org/10.1098/rsta.2020.0051

[6] S. R. Teli1, S. Zvanovec, Z. Ghassemlooy: Optical Internet of Things within 5G: Applications and Challenges, IEEE International Conference on Internet of Things and Intelligence System, 2018, DOI: 10.1109/IOTAIS.2018.8600894

[7] I. Alimi, A. Shahpari, A. Sousa, R. Ferreira, P. Monteiro and A. Teixeira: Challenges and Opportunities of Optical Wireless Communication Technologies, 2017, DOI: 10.5772/intechopen.69113

[8] Z. Ghassemlooy, S. Amon, M. Uysal, Z. Xu, J. Cheng: Emerging optical wireless communications advances and challenges, IEEE Journal on Selected Areas in Communications, Vol. 33 (9), 2015, pp. 1738-1749

[9] M. Chowdhury, M. Hossan, A. Islam, Y. Jang: A Comparative Survey of Optical Wireless Technologies: Architectures and Applications, IEEE Access, Vol. 6, 2018, pp. 9819-9840, DOI: 10.1109/ACCESS.2018.2792419

[10] A. Sevincer, A. Bhattarai, M. Bilgi, M. Yuksel, N. Pala: LIGHTNETs: smart LIGHTing and mobile optical wireless networks – a survey, IEEE Communication Surveys & Tutorials, Vol. 15, No. 4, 2013, pp. 1620-1641, DOI: 10.1109/SURV.2013.032713.00150

[11] D. Karunatilaka, F. Zafar, V. Kalavally, R. Parthiban: LED based indoor visible light communication: State of the art, IEEE Communication Surveys & Tutorials, Vol. 17, No. 3, 2015, pp. 1649-1678

[12] M. Chowdhury, Md. Shahjalal, M. Hasan, Y. Jang: The Role of Optical Wireless Communication Technologies in 5G/6G and IoT Solutions: Prospects, Directions, and Challenges, Applied Sciences, 2019, DOI: 10.3390/app9204367

[13] O. Alsulami, A. T. Hussein, M. T. Alresheedi and J. M. H. Elmirghani: Optical Wireless Communication Systems, A Survey, 2018, DOI: 10.13140/RG.2.2.11751.09129

[14] H. Haas: Wireless data from every light bulb, TED Global, Edinburgh, July 2011

[15] A. M. Khalid, G. Cossu, R. Corsini, P. Choudhury, E. Ciaramella: 1-Gb/s transmission over a phosphorescent white LED by using rate adaptive discrete multitone modulation, IEEE Photon. J., Vol. 4, Issue 5, 2012, pp. 1465-1473

[16] A. Azhar, T. Tran, D. O'Brien: A Gigabit/s indoor wireless transmission using MIMO-OFDM visible-light communications, IEEE Photon. Technol. Lett., Vol. 25, No. 2, 2013, pp. 171-174

[17] R. X. G. Ferreira, E. Xie, J. J. D. McKendry, et al.: High bandwidth GaN-based micro-LEDs for visible light communication, IEEE Photonics Technology Letters, Vol. 28 (19), 2016, DOI:10.1109/LPT.2016.2581318

[18] D. O'Brien, S. Rajbhandari, H. Chun: Transmitter and receiver technologies for optical wireless, Royal Society, Vol. 378, Issue 2169, 2020, doi.org/10.1098/rsta.2019.0182

[19] S. U. Rehman; S. Ullah, P. Chong, S. Yongchareon: Visible Light Communication: A System Perspective-Overview and Challenges, Sensors, Vol. 19 (5), N 1153, 2019, DOI:10.3390/s19051153

[20] L. Matheus, A. Vieira; F. M. Luiz, M. Vieira, Omprakash Gnawali: Visible Light Communication: Concepts, Applications and Challenges, IEEE Communications Surveys & Tutorials, Vol. 21, Issue 4, 2019, pp. 3204-3237, DOI:10.1109/COMST.2019.2913348

[21] A. Baklanov, S. Grigoryeva, A. Alimkhanova, E. Grigoryev, V. Sayun: Audio Transmission System Using White LEDs, International Siberian Conference on Control and Communications (SIBCON), Tomsk, Russia, 2019, DOI:10.1109/SIBCON.2019.8729564

[22] E. A. Grigoryev, A. E. Baklanov, S. V. Grigoryeva, A. Zh. Alimkhanova, V. M. Sayun: A New Approach to Physical Encoding in VLC Data Transmission Technology, 21st International Conference on

Micro/Nanotechnologies and Electron Devices (EDM), Erlagol, Russia, 2020

[23] S. Grigoryeva, A. Baklanov, V. Sayun, D. Titov, Ye. Grigoryev: Analysis energy efficiency of automated control system of LED lighting, International Siberian Conference on Control and Communications (SIBCON), 2017, DOI:10.1109/SIBCON.2017.7998488

[24] A. Baklanov, S. Grigoryeva, Gy. Györök: Control of LED Lighting Equipment with Robustness Elements, Acta Polytechnica Hungarica, Vol. 13, No. 5, 2016, pp. 105-119, DOI:10.12700/APH.13.5.2016.5.6

[25] Y. A. Grigoryev, V. M. Sayun, S. V. Grigoryeva, D. N. Titov: Study of Illumination Properties of High-Power LEDs in Various Temperature Conditions, 18th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM), Erlagol, Russia, 2017, pp. 309-313, DOI:10.1109/EDM.2017.7981762

[26] F. A. Dahri, S. Ali, M. M. Jawaid: A Review of Modulation Schemes for Visible Light Communication, International Journal of Computer Science and Network Security, Vol. 18, No. 2, 2018, pp. 117-122

[27] A. Aliaberi, P. C. Sofotasios, S. Muhaidat: Modulation Schemes for Visible Light Communications, Advanced Communication Technologies and Networking (CommNet), 2019, DOI:10.1109/COMMNET.2019.8742376

[28] C. Manimegalai, S. Gauni, N. Raghavan, T. Rao: Investigations on suitable modulation techniques for visible light communications, International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 1818-1822, DOI:10.1109/WiSPNET.2017.8300075

[29] E. V. Baklanov, N. N. Golovin, A. K. Dmitriev, S. Grigor'eva: A femtosecond frequency standard with an external high-finesse interferometer, Optics and Spectroscopy, Vol. 121, No. 6, 2016, pp. 930-933, DOI:10.1134/S0030400X16120055

[30] FLProg. Link: https://flprog.ru/en/

[31] C. Kárász, J. Kopják: Comparative Study on Plant Type Specific LED Light Source Design Parameters, Acta Polytechnica Hungarica, Vol. 17, No. 3, 2020, pp. 71-90, DOI:10.12700/APH.17.3.2020.3.4

Design of Multidimensional Classifiers using Fuzzy Brain Emotional Learning Model and Particle Swarm Optimization Algorithm

Yuan Sun, Chih-Min Lin*

Department of Electrical Engineering, Yuan Ze University, Tao-Yuan 320, Taiwan,
s1058505@mail.yzu.edu.tw; cml@saturn.yzu.edu.tw* (corresponding author)

Abstract: This study presents a multidimensional classifier design using a fuzzy brain emotional learning model, combined with a particle swarm optimization (PSO) algorithm that allows a network to automatically determine the optimum values for the weights of the reward signal. The multidimensional fuzzy brain emotional learning classifier(MFBELC) is first described with corresponding fuzzy inference rules; then the PSO algorithm is applied for the optimum parameter choice. This PSO-MFBELC is evaluated for the Wine dataset and Iris dataset, which are publicly available from the UCI machine learning database. A comparison of simulations using the proposed PSO-MFBELC shows that this classifier is superior to other algorithms in the recognition accuracy aspect.

Keywords: brain emotional learning model; fuzzy inference system; particle swarm optimization algorithm; multidimensional classifier

1 Introduction

The inspiration for emotional learning in the brain comes from the anatomical discovery of LeDoux's emotional learning mechanism in the mammalian brain in 1991 [1]. In 2001, an algorithm based on the computational model of emotional processing - brain emotional learning (BEL) was initially developed by Moren, with the advantages of low computational complexity, fast convergence, and good stability [2].

In several studies, BEL model has been widely used for control [3] [6] [7] [8], prediction [9]-[10] [11] [12], identification [13], [14] and binary classification [15]-[16] [17] [18]. In recent years, this model has also been extended to overcome the multi-classification problem [19], [20]. However, in the previous application studies (including the authors' past papers [5], [17], [18], [20]), it is difficult to determine the appropriate parameters of the BEL model, and the parameters that need to be set for different samples are different. Whether the setting of these parameters is appropriate or not has a great impact on the results. Most scholars usually use the trial-and-error method to set the parameters, but it is time-consuming and unstable. Therefore, in order to make the parameter setting more efficient and stable, several optimization algorithms have been proposed; some examples are given as follows. A gray wolf optimizer (GWO) algorithm has been proposed for tuning the parameters of Takagi-Sugeno proportional-integral fuzzy controllers (PI-FCs) [21]. Iterative feedback tuning (IFT) and iterative learning control (ILC) have been used to minimize the objective function [22]. A weighted interest pattern (WIP) mining method has been proposed to improve the performance of data mining [23].

This paper uses the particle swarm optimization (PSO) algorithm to search the appropriate parameters for achieving desired classification performance. The PSO algorithm is a simplified model based on swarm intelligence, which is inspired by the regularity of bird swarm activities [24]. Previously, PSO was applied to some artificial intelligent algorithms [25]-[26] [27] [28] [29], and until recently, some researchers have utilized PSO to find the most suitable parameters in the structure of the BEL model [30]-[33]. In spite of these applications, an appropriate choice of optimal parameters or fitness function for a PSO-BEL algorithm is necessary for different applications.

For intelligent systems, there were a lot of modeling techniques and they have been applied in various fields. The combination of fuzzy logic, neural network, genetic algorithm, and statistical analysis is analyzed in [34]. A new feature-based expert system modeling method is proposed in [35]. The modeling of a multi-relational classifier has been proposed based on canonical correlation analysis [36]. In [36], the method of model transformation based on tensor product model is applied to magnetic levitation systems

This paper aims to propose a more efficient multidimensional classifier. The brain emotion model, fuzzy inference system, and PSO algorithm are combined to form a new intelligent model. Then, a multidimensional fuzzy brain emotion learning classifier with reward signal optimization is developed. The main contributions of this paper are as follows. (1) A multidimensional classifier based on fuzzy inference system and BEL model (MFBELC) is proposed, (2) The PSO algorithm is successfully applied to search the optimal values of the two weight factors of reward signal in MFBELC, and then the classification performance is obviously improved. (3) The effectiveness of the proposed classifier has been verified by two multidimensional classification examples, and it can achieve better accuracy than most other classification models.

The rest of this study is organized as follows: Section 2 introduces the overall structure of the PSO multidimensional fuzzy brain emotional learning classifier, including the updating algorithm and the implementation process. Section 3 introduces the simulation results in detail and compares the performance of the proposed classifier with other models. Conclusions are detailed in Section 4.

2 The PSO - MFBELC Model

The proposed PSO-MFBELC model consists of two parts: the multidimensional fuzzy brain emotion learning classifier and PSO algorithm. The PSO algorithm searches the optimal parameters through iterations and then assigns these parameters to the MFBELC model. The details of the algorithm are described in Section 2.3.

2.1 Multidimensional Fuzzy Brain Emotional Learning Classifier

2.1.1 Fuzzy Inference Rules of MFBELC

In a traditional brain emotional learning model, sensory input is calculated in the sensory cortex and sent directly to the orbitofrontal cortex and amygdala, without any learning process. Different from the traditional BEL model, for the proposed MFBELC, the fuzzy inference rules are proposed and defined as:

$$\text{If } I_1 \text{ is } S_{1j} \text{ and } I_2 \text{ is } S_{2j}, \ \cdots, \ \text{ and } I_{n_ij} \text{ is } S_{n_ij}, \text{ then } A_l = V_{ijl} \tag{1}$$

$$\text{If } I_1 \text{ is } S_{1j} \text{ and } I_2 \text{ is } S_{2j}, \ \cdots, \ \text{ and } I_{n_ij} \text{ is } S_{n_ij}, \text{ then } O_l = W_{ijl} \tag{2}$$

$$\text{for } i = 1,2,\cdots,n_i, \quad j = 1,2,\cdots,n_j, \quad l = 1,2,\cdots,n_l$$

where n_i is the input dimension, n_l is the output dimension and n_j is the number of neurons. S_{ij} is the fuzzy set for the i-th input and j-th neuron. A_l is the l-th output of the amygdala, and O_l is the l-th output of the orbitofrontal cortex. V_{ijl} is the amygdala weight for the l-th output corresponding to the i-th input and j-th neuron in the consequent part. Likewise, W_{ijl} is the orbitofrontal cortex weight for the l-th output corresponding to the i-th input and j-th neuron in the consequent part.

2.1.2 Structure of a Multidimensional Fuzzy Brain Emotional Learning Classifier

Figure 1 shows a multidimensional brain emotional learning classifier with six layers: the sensory input, sensory cortex, orbitofrontal cortex, thalamus, amygdala, and the output space. The following details the data transmission of MFBELC and the basic functions of each layer.

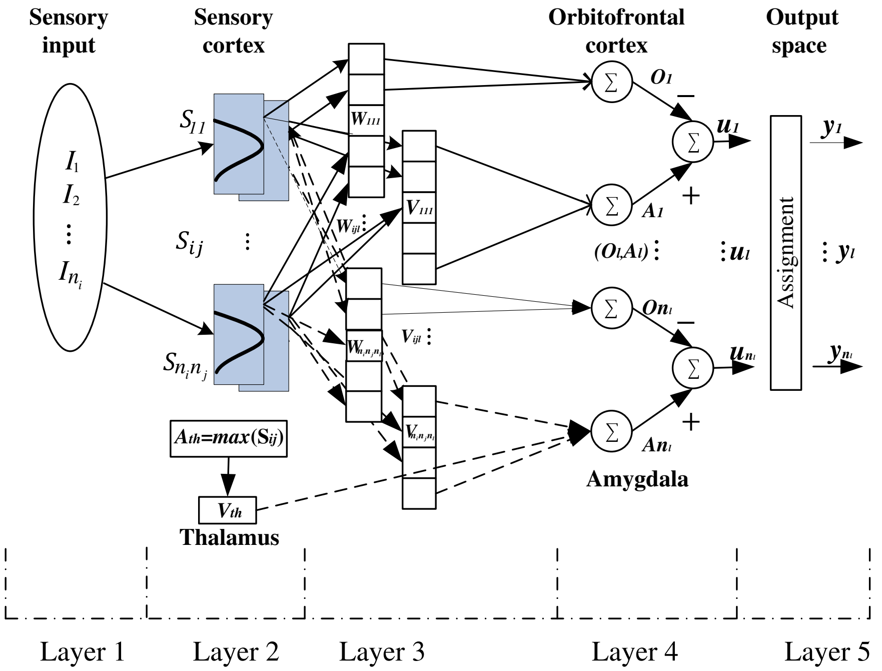


Figure 1

Structure of the multidimensional brain emotional learning classifier

a) Layer 1

Layer 1 is the sensory input space, where the input vector $I = [I_1, \cdots, I_i, \cdots, I_{n_i}]^T \in R^{n_i}$. In general, according to the given classification problem, the input dimension can also be regarded as the feature dimension.

b) Layer 2

Layer 2 is the sensory cortex space. Sensory input is transmitted to the orbitofrontal cortex and amygdala. In order to improve generalization ability and operation speed, The Gaussian function is used as a membership function as follows:

$$S_{ij} = exp\left[\frac{-(I_i - m_{ij})^2}{\sigma_{ij}^2}\right] \tag{3}$$

where m_{ij} and σ_{ij} correspond to the mean and variance of S_{ij}, respectively.

The thalamus receives the maximum signal from the sensory input layer, and it is known as the thalamic signal

$$A_{th} = max(S_{ij}) \tag{4}$$

c) Layer 3

Layer 3 is the weight space. Therein, a fuzzy output is represented by a block, which is the result of fuzzy inference rules.

For the amygdala system, this space is called the sensory weight space *V*, expressed in a vector form:

$$V = [V_{111}, \cdots, V_{ijl}, \cdots V_{n_i n_j n_l}]^T \in R^{n_i n_j n_l} \tag{5}$$

For the orbitofrontal cortex system, this space is called emotion weight space *W*, expressed in a vector form:

$$W = [W_{111}, \cdots, W_{ijl}, \cdots W_{n_i n_j n_l}]^T \in R^{n_i n_j n_l} \tag{6}$$

d) Layer 4

Layer 4 is the algebraic sum of input S_{ij} for the sensory cortex with activation weights.

For the orbitofrontal cortex, the corresponding nodes in the orbitofrontal cortex can receive the signals from the amygdala. The l-th output in the orbitofrontal cortex is

$$O_l = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} S_{ij} W_{ijl} \tag{7}$$

For the amygdala system, stimulation is received through the corresponding node in three parts: sensory input, reward signal, and thalamic signal. The l-th output in the amygdala is

$$A_l = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} S_{ij} V_{ijl} + A_{th} V_{th} \tag{8}$$

e) Layer 5

Layer 5 is the output space. It is the output of the brain emotional learning model, designed as

$$u_l = A_l - O_l \tag{9}$$

where u_l receives the l-th output from amygdala, and then subtracts the inhibitory outputs from the orbitofrontal cortex.

However, either in the training process or in the testing phase, the classification judgment result of each sample needs to be given, which means the output of the classifier should reflect the category label. For a binary classification problem, the sigmoid function is commonly adopted and a cut-off threshold is used to separate the two categories. For a multidimensional classifier, this method may produce some obstacles. Thus, the final output of this multidimensional brain emotional learning classifier is defined as

$$y_l = \begin{cases} 1, & u_l = max(u_1, u_2, \cdots, u_{n_l}) \\ 0, & else \end{cases} \tag{10}$$

From (10), obviously, the total output of the multidimensional classifier is presented as a multidimensional array and the index of the value 1 indicates the category label.

2.1.3 Learning Algorithm for MFBELC

Each emotional learning process in the amygdala and orbitofrontal cortex is a process of dynamic weight adjustment. According to an associative learning method [38], the l-th weight updating formulas of the amygdala and orbitofrontal cortex are respectively applied as

$$\Delta V_{ijl} = \lambda_v \ (S_{ij}\, max(0, REW_l - A_l)) \tag{11}$$

$$\Delta W_{ijl} = \lambda_w(S_{ij}(A_l - O_l - A_{th}V_{th} - REW_l)) \tag{12}$$

where λ_v and λ_w are the learning rates respectively for the amygdala and orbitofrontal cortex, which are the key elements that bear the influence on the learning speed. REW_l is the reward signal for the l-th output. Define the l-th output error as

$$e_l = t_l - y_l \tag{13}$$

where t_l and y_l are the l-th expected target and assignment output, respective.

Then, in this study, the reward signal can be a function of the error signal and the output of the model; it is selected as:

$$REW_l = k_1 e_l + k_2 u_l \tag{14}$$

where k_1 and k_2 are both weight factors, which are adjusted respectively for the expectation of error reduction and output. In general, the value of k_1 should be larger than that of k_2, because the error of model in learning process is always smaller than the output, and these two weight factors will be automatic searched by the PSO algorithm in this design.

Define the cost function

$$E_l = \frac{1}{2} e_l^2 \tag{15}$$

Because the gradient descent method can reduce the error as quickly as possible, the adjustment of the mean and variance of Gaussian function is generated by the gradient descent algorithm for minimizing the cost function, as

$$\begin{aligned}\Delta m_{ij} &= -\lambda_m \frac{\partial E_l}{\partial m_{ij}} = -\lambda_m \frac{\partial E_l}{\partial e_l}\frac{\partial e_l}{\partial y_l}\frac{\partial y_l}{\partial u_l}\frac{\partial u_l}{\partial s_{ij}}\frac{\partial s_{ij}}{\partial m_{ij}} \\ &= \lambda_m e_l \cdot y_l \cdot (1 - y_l) \cdot (V_{ijl} - W_{ijl}) \cdot S_{ij} \cdot \frac{2(I_i - m_{ij})}{\sigma_{ij}^2}\end{aligned} \tag{16}$$

$$\begin{aligned}\Delta \sigma_{ij} &= -\lambda_\sigma \frac{\partial E_l}{\partial \sigma_{ij}} = -\lambda_\sigma \frac{\partial E_l}{\partial e_l}\frac{\partial e_l}{\partial y_l}\frac{\partial y_l}{\partial u_l}\frac{\partial u_l}{\partial s_{ij}}\frac{\partial s_{ij}}{\partial \sigma_{ij}} \\ &= \lambda_\sigma e_l \cdot y_l \cdot (1 - y_l) \cdot (V_{ijl} - W_{ijl}) \cdot S_{ij} \cdot \frac{2(I_i - m_{ij})}{\sigma_{ij}^3}\end{aligned} \tag{17}$$

where λ_m and λ_σ are the learning rates.

2.2 Particle Swarm Optimization Algorithm

2.2.1 PSO Description

Particle Swarm Optimization (PSO) was inspired by the research of birds' foraging behavior in nature: a bevy of birds look for food in a random location in an area, and each bird can know what's the distance between their current position is and the location of the food. Therefore, as long as the bird closest to the food is found, searching around it is an effective way. Consequently, the heart of the matter of finding the optimal parameters is usually solved by using the PSO algorithm. In PSO algorithm, birds in search space are replaced by particles. Each particle has its direction and speed when searching. Then, the particle adjusts its position and direction according to the current best particle position and direction and searches the solution space [39].

In the process of each iteration, the particle swarm constantly adjusts and updates its position and speed by the following formula [40]:

$$V_{od}^{k+1} = \omega V_{od}^{k} + c_1 r_1 (p_{od}^{k} - L_{od}^{k}) + c_2 r_2 (G_{od}^{k} - L_{od}^{k}) \tag{18}$$

$$L_{od}^{k+1} = L_{od}^{k} + V_{od}^{k+1} \tag{19}$$

where $d = 1,2,\cdots,D$; $o = 1,2,\cdots,N$. V_{od}^{k} is the velocity of the o-th particle in the d-th dimensional space, L_{od}^{k} is the position of the o-th particle in the d-th dimensional space; L_{od}^{k+1} and V_{od}^{k+1} represent the updated values of L_{od}^{k} and V_{od}^{k}, respectively; k represents the number of iterations; c_1 and c_2 are acceleration factors, both of which are non-negative constants. They play a role in adjusting each particle to obtain the optimal individual step size and the optimal group step size, respectively. r_1 and r_2 are random numbers in the interval [0,1]; ω is the inertia weight, which means that the particle inherits the proportion of the previous speed. In PSO algorithm, the velocity and position of particles are often restricted to the region $[V_{min}, V_{max}]$ and $[L_{min}, L_{max}]$, so as to avoid the blind search of particles in the space of feasible solutions, which results in the loss of the superiority of the algorithm itself. The region of the velocity is often adjusted according to the range of position.

2.2.2 Fitness Function

The fitness function is applied to choose the best particle, and for the classification problem, it can be evaluated by the accuracy of the classifier. The following fitness function is applied for the whole particle search

$$fitness = 1 - ACC_{train} \tag{20}$$

where ACC_{train} is the training accuracy rate in the MFBELC. That is to say, the fitness function of PSO is equal to the error of the training results in MFBLEC, PSO finds the optimal parameters k_1 and k_2 according to minimizing the fitness value. If the error is small, it means that k_1 and k_2 can achieve good training results.

2.3 PSO-MFBELC

The flowchart of PSO-MFBELC is shown in Figure 2. The specific steps to achieve this algorithm are depicted as follows.

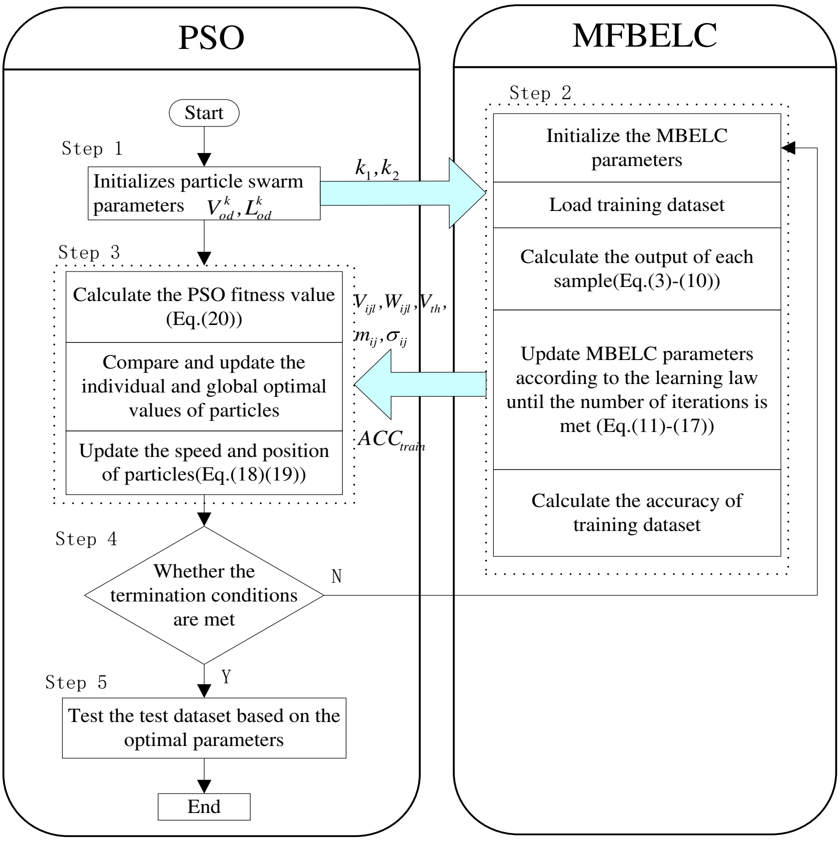

PSO
MFBELC
Start
Step 1
Initializes particle swarm parameters V_{od}^k, L_{od}^k
k_1, k_2
Step 2
Initialize the MBELC parameters
Load training dataset
Calculate the output of each sample(Eq.(3)-(10))
Update MBELC parameters according to the learning law until the number of iterations is met (Eq.(11)-(17))
Calculate the accuracy of training dataset
Step 3
Calculate the PSO fitness value (Eq.(20))
Compare and update the individual and global optimal values of particles
Update the speed and position of particles(Eq.(18)(19))
$V_{ijl}, W_{ijl}, V_{th}, m_{ij}, \sigma_{ij}$
ACC_{train}
Step 4
Whether the termination conditions are met
N
Y
Step 5
Test the test dataset based on the optimal parameters
End

Figure 2
Flowchart for a PSO-MFBELC

- Step 1. Generate the particle population for PSO, including particle swarm size, velocity vector, position, and other parameters. In this step, the initial values of k_1 and k_2, which are the parameters to be optimized, are transmitted to the MFBELC training stage.

- Step 2. Initialize the parameters of MFBELC and the training phase will be operated by the specified learning algorithms. Set the initial conditions and inputs. In general, the initial values of the parameters of MFBELC, such as $V_{ij}, W_{ij}, V_{th}, m_{ij}, \sigma_{ij}$ are chosen as random values. The features of training samples are put as the input of MFBELC, and the output are obtained by equations (3)-(10). Then, the parameters of MFBELC are modified according to the learning algorithm, by equations (11), (12), and (16), (17). When the training phase is finished, the accuracy of the training dataset is returned to PSO.

- Step 3. The fitness function is calculated by equation (20) and is assigned to each particle. The individual extreme values, global extreme values, and the speed and position of the particles are compared and updated in this step, by equations (18)-(19).

- Step 4. Determine whether the condition meets the number of iterations and the desired precision requirements. That is, the error value ≦ 0.2% or the set number of iterations 500 is achieved. If the stop condition is reached, jump out of the iteration; if not, move back to step 2.

- Step 5. Obtain the optimized parameters. The performance of the test is calculated and the effectiveness of the model is evaluated.

3 Simulation Results

In this section, to verify the performance of the proposed model, the proposed model is evaluated for two standard multiclass datasets: 1) Wine; 2) Iris, which are shared in the UCI machine learning database [40].

3.1 Description of Dataset

The description of the two datasets and training-testing proportion are shown in Table 1. There are 178 samples in the wine dataset, of which 100 samples are used for training and the other 78 samples are used for testing. Iris data set has 150 samples, of which 75 samples are used for training and the other 75 samples are used for testing.

Table 1

Description of datasets

Dataset	**No. of Sample**	**No. of training sample**	**No. of testing sample**	**No. of Cluster**	**No. of Attribute**
Wine	178	100	78	3	13
Iris	150	75	75	3	4

Before constructing the model, if the data has a high diversity among different attributes, it should be preprocessed. Such as the Wine dataset, the values of input features are normalized between the range [0, 1] first.

3.2 Experimental Results

From Figure 2, the proposed method begins with the initialization of parameters for PSO. The parameters applied for the two datasets are tabulated in Table 2, which shows the parameters and their numbers and ranges for the PSO to search k_1 and k_2. In general, the most important consideration of parameter setting is whether it can converge during the iteration. In most situations, the size of swarms, or the maximal iteration number may affect the speed of training. Besides, the parameters for the MFBELC, including m_{ij}, σ_{ij}, V_{ijl}, W_{ijl} are randomly initialized.

Table 3 shows the parameter setting for the MFBELC and the search results of PSO-MFBELC. It shows that due to the use of PSO for optimization, appropriate values of k_1 and k_2 are determined and the training epochs are much reduced. It has been demonstrated that the process of optimization is also beneficial to the learning of the MFBELC model and the burden of MFBELC could be reduced.

Furthermore, the optimal values of k_1 and k_2 for the two datasets are different. This also explains the difficulty of the selection of the two parameters and it is necessary to introduce PSO algorithm for parameter optimization.

Table 2

Parameters used for PSO

Parameter	Value	
	Wine dataset	**Iris dataset**
Swarm Size	20	20
Max of Generations	10	10
ω	[0.2-0.8]	[0.2-0.8]
c_1	2	2
c_2	2	2
$[V_{min}, V_{max}]$ for searching k_1	[-10,10]	[-10,10]
$[V_{min}, V_{max}]$ for searching k_2	[-1,1]	[-1,1]
$[L_{min}, L_{max}]$ for searching k_1	[0,200]	[0,1000]
$[L_{min}, L_{max}]$ for searching k_2	[-5,5]	[-5,5]

Table 3
Comparison of the parameters for different MFBELC models

Dataset	**Model**	k_1	k_2
Wine	MFBELC	150	1
	PSO-MFBELC	172.27*	0.64*
Iris	MFBELC	20	1
	PSO-MFBELC	615.68*	1.09*

"'' refers to the parameter value after PSO optimization*

The classification results of PSO-MFBELC are represented by the confusion matrix as shown in Figure 3. The green squares in the confusion matrix represent the number of correctly identified samples. For example, the number in the first row and the first column represents the sample target is class 1, and the prediction result is also class 1. The red square represents the number of samples with incorrect predictions. For example, the number in the second column of the third row indicates that the sample target is class 2, and the prediction result is that there is in class 3. In the multi-classification confusion matrix, the precision is divided by the value on the main diagonal by the sum of the row in which the value is located, and the sensitivity is equal to the value on the main diagonal divided by the sum of the column in which the value is located. Finally, the accuracy is the sum of the diagonal values.

The classification result for the PSO-MFBELC is shown in Figure 4 and Figure 5, respectively, for the two datasets, and the percentage below the number in the box represents the number divided by the total number of samples. There are two parts: one for training samples and the other for testing samples; both displayed in their respective confusion matrix. From Figure 4, on Wine dataset, in the training phase, there is one sample misclassified. Meanwhile, there are two samples misclassified in the testing phase. Therefore, the accuracy values for the training and testing dataset are 99.0% and 97.4%, respectively. Likewise, from Figure 5, on the Iris dataset, the accuracy values for training and testing are 93.3% and 98.7%, respectively.

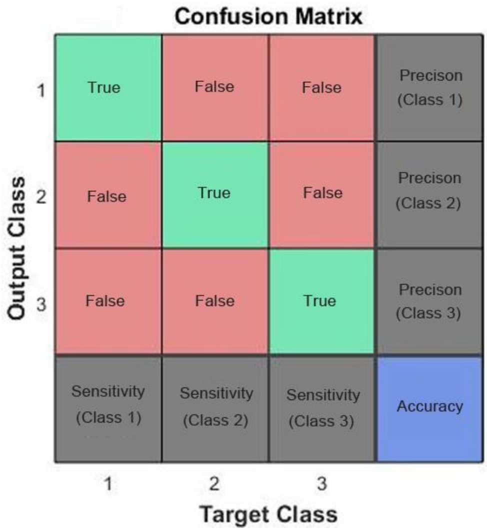
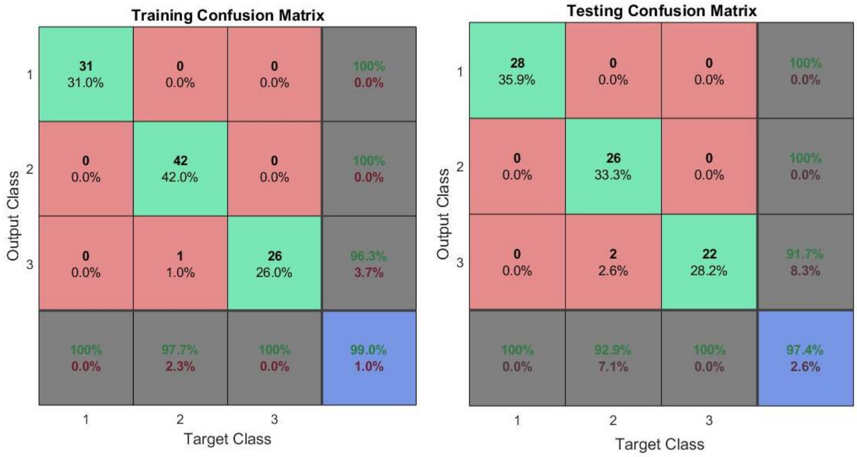
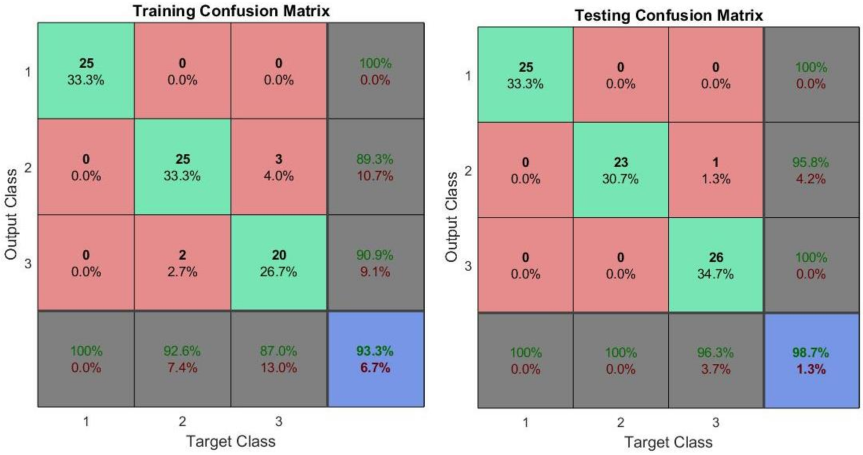


Figure 3
Confusion matrix description



(a) (b)

Figure 4
Classification result for the PSO-MFBELC on Wine dataset, (a) training phase, (b) testing phase



(a) (b)

Figure 5

Classification result for the PSO-MFBELC on Iris dataset, (a) training phase, (b) testing phase

3.3 Performance Evaluation

In order to illustrate the effectiveness of the new model, the PSO-MFBELC is compared with MFBELC without PSO. For the purpose to ensure a fair comparison, the simulations are repeated for 10 runs for both algorithms. The comparisons of testing accuracies are displayed in Figure 6 and Figure 7. Apparently, the accuracy of PSO-MFBELC is higher than that of MFBELC. Besides, the values of accuracies obtained by PSO-MFBELC are relatively more stable than that of MFBELC. Table 4 lists the training accuracy and testing accuracy of two datasets with 10 times running for the PSO-MFBELC.

Table 4

PSO-MFBELC running results (%)

Dataset	**Wine**		**Iris**	
Time	**Train Accuracy**	**Test Accuracy**	**Train**	**Test Accuracy**
1	99.00	97.44	93.33	98.67
2	98.00	98.72	93.33	98.67
3	98.00	97.44	94.67	98.67
4	99.00	97.44	93.33	97.33
5	98.00	97.44	93.33	98.67
6	98.00	96.15	93.33	98.67
7	99.00	97.44	93.33	98.67
8	98.00	97.44	94.67	98.67

9	98.00	97.44	93.33	98.67
10	98.00	96.15	93.33	98.67

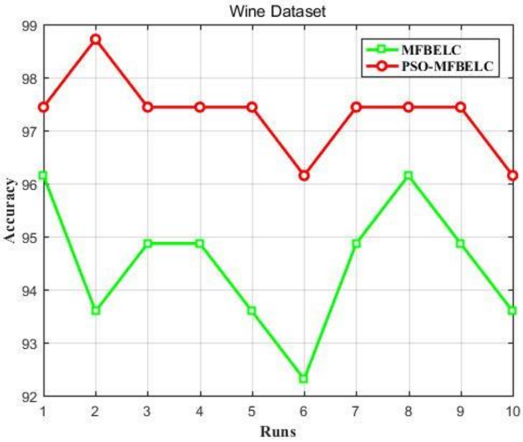


Figure 6

Comparison of testing accuracy values for different models on Wine dataset

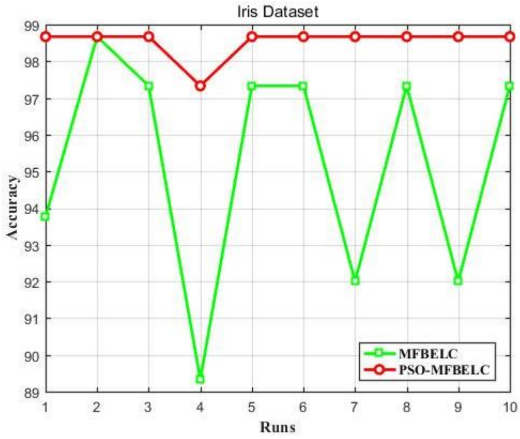


Figure 7

Comparison of testing accuracy values for different models on Iris dataset

Table 5 summarizes the results of accuracy for the two classifiers when operating for 10 times on different datasets. On the Wine dataset, the maximum training accuracy of PSO-MFBELC and MFBELC are100% and 99%, respectively; while the average values are 100% and 98.3%, respectively. The maximum testing accuracy of PSO-MFBELC and MFBELC are 98.72% and 96.15%, respectively; while the average values are 97.31% and 94.49%, respectively. On the Iris dataset, for PSO-MFBELC and MFBELC, although the maximum testing accuracy of the two algorithms is both 98.67%; however, the average values are 98.54% and 94.86%, respectively. Therefore, to summarize, the PSO-MFBELC model has better performance than the MFBELC model, which clearly indicates that taking advantage of the PSO algorithm to train the MFBELC model could enhance the performance of the brain emotional learning classifier.

Table 5

Performance comparisons (%)

Dataset	**Model**	**Train Accuracy (ACC)**			**Test Accuracy (ACC)**		
		Highest	**Lowest**	**Avg**	**Highest**	**Lowest**	**Avg**
Wine	MFBELC	100	100	100	96.15	92.23	94.49
	PSO-MFBELC	99	98	98.3	98.72	96.15	97.31
Iris	MFBELC	93.33	90.67	92.66	98.67	89.21	94.86
	PSO-MFBELC	94.67	93.33	93.6	98.67	97.33	98.54

In order to verify the proposed method, the comparison with other methods, shown in published literatures [42]-[51] which used the same data set, is summarized in Table 6. The classification accuracy of the proposed PSO-MFBELC is satisfactory compared to the other classifiers.

Table 6

Reported results in literatures

Dataset	**Author (Year)**	**Method**	**Test Accuracy (ACC)**
Wine	Hidayat et al.[42](2016)	ACO	89.90
	Chen et al. [43](2016)	DCQGA-SVM	90.4109
	Chakravarty et al. [44](2015)	CSFLNFN	93.5
	Guerrero-Enamorado et al. [45] (2016)	LOGIT-BOOST	97.07
	Xu et al. [46](2015)	FKNN+FLMDA	98.10
	Wongthanavasu et al. [47](2016)	CAC	98.18
	Xu et al. [48](2016)	Bayes net	98.59
Iris	Hidayat et al. [42](2016)	PREACO	90
	Zhang et al. [49](2015)	W-KNN	95.83
	Guerrero-Enamorado et al. [45] (2016)	MCGEP	96.53
	Xu et al. [48](2016)	DC-Core samples	96.67
	Yu et al. [50](2015)	SI-INNO + LGC	98
	Wongthanavasu et al. [47](2016)	Linear SVM	98.01
	Chakravarty et al. [44](2015)	CSMLP	98.1
	Chen et al. [43](2016)	DCQGA-SVM	98.3
	Syaliman et al. [51](2018)	LMKNN+DWKNN	98.33

Conclusion

Based on the fuzzy brain emotional learning model and PSO algorithm, this paper constructs a PSO-MFBELC for multiple classifications. PSO is used to automatically search for appropriate values of the weights of the reward signal of MFBELC, which affects the training speed and predictive accuracy of the classification model. Then, the proposed PSO-MFBELC is applied to two datasets. Numerical simulations show that the algorithm has high generalization ability and accuracy, not only the model structure is simple but also easy to implement. It is clear that this method can also be applied to other multidimensional classification problems.

Acknowledgement

This paper was supported in part by the Ministry of Science and Technology of the Republic of China under Grant MOST 106-2221-E-155-MY3.

References

[1] J. E. LeDoux, "Emotion and the limbic system concept," Concepts Neurosci, Vol. 2, pp. 169-199, 1991

[2] C. B. J. Moren, C. Balkenius, "Emotional learning:a computational model of the amygdala," Cybernetics & Systems, Vol. 32, No. 6, pp. 611-636, 2001

[3] M. A. Sharbafi, C. Lucas, R. Daneshvar, "Motion control of omni-directional three-wheel robots by brain-emotional-learning-based intelligent controller," IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), Vol. 40, No. 6, pp. 630-638, 2010

[4] P. K. Muthusamy, M. Garratt, H. Pota, et al. "Bidirectional fuzzy brain emotional learning control for aerial robots," 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 146-153

[5] Q. Wu, C. M. Lin, W. Fang, et al. "Self-organizing brain emotional learning controller network for intelligent control system of mobile robots," IEEE Access, pp. 59096-59108, 2018

[6] M. Jafari, R. Fehr, L. R. G. Carrillo, H. Xu, "Brain emotional learning-based intelligent tracking control for unmanned aircraft systems with uncertain system dynamics and disturbance," 2017 International conference on unmanned aircraft systems (ICUAS), IEEE, 2017, pp. 1470-1475

[7] M. Jafari, A. M. Shahri, S. B. Shuraki, "Speed control of a digital servo system using brain emotional learning based intelligent controller," Power Electronics, Drive Systems and Technologies Conference (PEDSTC), 2013 4th, IEEE, 2013, pp. 311-314

[8] C. F. Hsu, C. T. Su, T. T. Lee, "Chaos synchronization using brain-emotional-learning-based fuzzy control," Joint International Conference on

Soft Computing & Intelligent Systems, IEEE, 2016, pp. 811-816

[9] M. Parsapoor, U. Bilstrup, "Chaotic time series prediction using brain emotional learning based recurrent fuzzy system (BELRFS)," International Journal of Reasoning-Based Intelligent Systems, Vol. 5, No. 2, pp. 113-126, 2013

[10] E. Lotfi, M. R. Akbarzadeht, "Adaptive brain emotional decayed learning for online prediction of geomagnetic activity indices," Neurocomputing, Vol. 126, No. 3, pp. 188-196, 2014

[11] H. S. A. Milad, U. Farooq, M. E. El-Hawary, et al. "Fuzzy logic based parameter adjustment model for adaptive decayed brain emotional learning network with application to online time series prediction," 2017 IEEE Electrical Power and Energy Conference (EPEC), IEEE, 2017, pp. 1-6

[12] J. Ayubi, A. Omidi, S. M. Barakati, P. Ayubi, "Short term load forecasting based on brain emotional predictor," 2015 20th Conference on Electrical Power Distribution Networks Conference (EPDC), IEEE, 2015, pp. 37-41

[13] E. Lotfi, S. Setayeshi, S. Taimory, "A neural basis computational model of emotional brain for online visual object recognition," Applied Artificial Intelligent, Vol. 28, No. 8, pp. 814-834, 2014

[14] S. Motamed, S. Setayeshi, A. Rabiee, "Speech emotion recognition based on a modified brain emotional learning model," Biologically Inspired Cognitive Architectures, Vol. 19, pp. 32-38, 2017

[15] E. Lotfi, "Mathematical modeling of emotional brain for classification problems," Proceedings of IAM, Vol. 2, No. 1, pp. 60-71, 2013

[16] M. Parsapoor, U. Bilstrup, "Brain emotional learning based fuzzy inference system (Modified using radial basis function)," Eighth International Conference on Digital Information Management, IEEE, 2014, pp. 206-211

[17] Q. Q. Zhou, F. Chao, C. M. Lin, "A functional-link-based fuzzy brain emotional learning network for breast tumor classification and chaotic system synchronization," International Journal of Fuzzy Systems,Vol. 20, No. 2, pp. 349-365, 2018

[18] Y. Sun, C. M. Lin, "A fuzzy brain emotional learning classifier design and application in medical diagnosis," Acta Polytechnica Hungarica, Vol. 16, No. 4, pp. 27-43, 2019

[19] M. Asad, U. Farooq, J. Gu, et al. "Neo-fuzzy supported brain emotional learning based pattern recognizer for classification problems," IEEE Access, 2017, pp. 6951-6968

[20] J. Zhao, C. M. Lin, "Multidimensional classifier design using wavelet fuzzy brain emotional learning neural networks," Journal of Intelligent & Fuzzy Systems, Vol. 36, No. 2, pp. 1099-1107, 2019

[21] R. E. Precup, R. C. David, E. M. Petriu, et al. "Grey wolf optimizer-based approach to the tuning of PI-fuzzy controllers with a reduced process parametric sensitivity," IFAC Papers OnLine, Vol. 49, No. 5, pp. 55-60, 2016

[22] S. Preitl, R. E. Precup, Z. Preitl, et al. "Iterative feedback and learning control. Servo systems applications," IFAC Proceedings, Vol. 40, No. 8, pp. 16-27, 2007

[23] U. Yun. "Efficient mining of weighted interesting patterns with a strong weight and/or support affinity," Information Sciences, Vol. 177, No. 17, pp. 3477-3499, 2007

[24] J. Kennedy, R. C. Eberhart, "Particle swarm optimization," Proceedings of ICNN'95-International Conference on Neural Networks. IEEE, 1995, pp. 1942-1948

[25] D. R. Nayak, R. Dash, B. Majhi, "Discrete ripplet-II transform and modified PSO based improved evolutionary extreme learning machine for pathological brain detection," Neurocomputing, Vol. 282, pp. 232-247, 2018

[26] S. K. Satapathy, S. Dehuri, A. K. Jagadev, "EEG signal classification using PSO trained RBF neural network for epilepsy identification," Informatics in Medicine Unlocked, Vol. 6, pp. 1-11, 2017

[27] S. Saraswathi, S. Sundaram, N. Sundararajan, et al. "ICGA-PSO-ELM approach for accurate multiclass cancer classification resulting in reduced gene sets in which genes encoding secreted proteins are highly represented," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 8, No. 2, pp. 452-463, 2010

[28] C. M. Lin, T. L. Le, "WCMAC-based control system design for nonlinear systemsusing PSO," Journal of Intelligent & Fuzzy Systems, Vol. 33, No. 2, pp. 807-818, 2017

[29] R. J. Wai, J. D. Lee, K. L. Chuang, "Real-time PID control strategy for maglev transportation system via particle swarm optimization," IEEE Transactions on Industrial Electronics, Vol. 58, No. 2, pp. 629-646, 2010

[30] H. T. Dorrah, A. M. El-Garhy, M. E. El-Shimy, "PSO-BELBIC scheme for two-coupled distillation column process," Journal of Advanced Research, Vol. 2, No. 1, pp. 73-83, 2011

[31] T. L. Le, C. M. Lin, T. T. Huynh, "Self-evolving type-2 fuzzy brain emotional learning control design for chaotic systems using PSO," Applied Soft Computing Journal, Vol. 73, pp. 418-433, 2018

[32] Y. Mei, X. Yang, Z. Liu, et al. "Real-time facial expression recognition based on the improved brain emotional learning model," 2018 37th Chinese Control Conference (CCC), IEEE, 2018, pp. 3924-3928

[33] S. H. Fakhrmoosavy, S. Setayeshi, A. Sharifi, "An intelligent method for generating artificial earthquake records based on hybrid PSO–parallel brain emotional learning inspired model," Engineering with Computers, Vol. 34, No. 3, pp. 449-463, 2018

[34] C. A. Laurentys, C. H. M. Bomfim, B. R. Menezes, et al. "Design of a pipeline leakage detection using expert system: A novel approach," Applied Soft Computing, Vol. 11, No. 1, pp. 1057-1066, 2011

[35] C. Pozna, R. E. Precup. "Applications of signatures to expert systems modelling," Acta Polytechnica Hungarica, Vol. 11, No. 2, pp. 21-39, 2014

[36] Zall R, Mohammad Reza K, "On the construction of multi-relational classifier based on canonical correlation analysis," International Journal on Artificial Intelligence Tools, Vol. 17, No. 2, pp. 23-43, 2019

[37] E. L. Hedrea, R. E. Precup, C. A. Bojan-Dragos. "Results on tensor product-based model transformation of magnetic levitation systems," Acta Polytechnica Hungarica, Vol. 16, No. 9, pp. 93-111, 2019

[38] C. Lucas, D. Shahmirzadi, N. Sheikholeslami, "Introducing BELBIC: brain emotional learning based intelligent controller," Intelligent Automation & Soft Computing, Vol. 10, No. 1, pp. 11-21, 2004

[39] E. A. Grimaldi, F. Grimaccia, M. Mussetta, et al. "PSO as an effective learning algorithm for neural network applications," 2004 3rd International Conference on Computational Electromagnetics and Its Applications, IEEE, 2004, pp. 557-560

[40] Y. Shi, R. Eberhart, "A modified particle swarm optimizer," 1998 IEEE international conference on evolutionary computation proceedings, IEEE world congress on computational intelligence, IEEE, 1998, pp. 69-73

[41] UCI repository of machine learning databases. Available at: http://archive.ics.uci.edu/ml/datasets/

[42] D. T. Hidayat, C. Fatichah, V. Raden, "Pattern reduction enhanced ant colony optimization clustering algorithm," International Seminar on Application for Technology of Information & Communication, IEEE, 2016, pp. 317-322

[43] P. Chen, L. Yuan, Y. He, et al. "An improved SVM classifier based on double chains quantum genetic algorithm and its application in analogue circuit diagnosis," Neurocomputing, Vol. 211, No. 26, pp. 202-211, 2016

[44] S. Chakravarty, P. Mohapatra, "Multi-class classification using cuckoo search based hybrid network," Power, Communication & Information Technology Conference, IEEE, 2015, pp. 1-8

[45] A. Guerrero-Enamorado, C. Morell, A. Y. Noaman, et al. "An algorithm evaluation for discovering classification rules with gene expression programming," International Journal of Computational Intelligence

Systems, Vol. 9, No. 2, pp. 263-280, 2016

[46] J. Xu, Z. Gu, K. Xie, "Fuzzy local mean discriminant analysis for dimensionality reduction," Neural Processing Letters, Vol. 44, No. 3, pp. 1-18, 2015

[47] S. Wongthanavasu, J. Ponkaew, "A cellular automata-based learning method for classification," Expert Systems with Applications, Vol. 49, pp. 99-111, 2016

[48] X. Xu, J. Zheng, J. Yang, et al. "Data classification using evidence reasoning rule," Knowledge Based Systems, Vol. 21, No. 8, pp. 1-8, 2016

[49] L. Zhang, C. Zhang, Q. Xu, et al. "Weigted-KNN and its application on UCI," IEEE International Conference on Information & Automation, IEEE, 2015, pp. 1748-1750

[50] C. Yu, F. Li, G. Li, et al. "Multi-classes imbalanced dataset classification based on sample information," 2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC), IEEE, 2015, pp. 1768-1773

[51] K. U. Syaliman, E. B. Nababan, O. S. Sitompul, "Improving the accuracy of k-nearest neighbor using local mean based and distance weight," Journal of Physics: Conference Series, 2018, pp. 1-6

Methodology for the Water Injection System Design Based on Numerical Models

Miroslav Spodniak[1], Ladislav Főző[1], Rudolf Andoga[2], Karol Semrád[1], Károly Beneda[3]

[1]Department of aviation engineering, Faculty of Aeronautics of Technical University of Košice, Rampová 7, 04001 Košice, Slovakia
miroslav.spodniak@tuke.sk, ladislav.fozo@tuke.sk, karol.semrad@tuke.sk

[2]Department of avionics, Faculty of Aeronautics of Technical University of Košice, Rampová 7, 04001 Košice, Slovakia
rudolf.andoga@tuke.sk,

[3]Department of Aeronautics, Naval Architecture and Railway Vehicles, Faculty of Transport Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary
kbeneda@vrht.bme.hu

Abstract: Jet engines are nowadays one of the most popular ways of propulsion for aircraft. This type of propulsion is widely used also in other sectors of industry. The main challenging task for designers is to design reliable and also powerful propulsion units. There are many manners for increasing the thrust of the particular jet engine and one of them is water injection into the compressor. In order to design such a system, it is necessary to have information about the flow parameters in the compressor. The proposed article deals with the CFD modeling of the radial compressor in order to estimate flow parameters for further research in the field of increasing thrust. One of the aims of the article is to develop a numerical model of the compressor and carry out the numerical analysis using CFD software ANSYS CFX. The analysis can be in further research performed multiple times for multiple regimes and results can be compared with the experimental measurements of thermodynamical values if the proper CFD model is developed. The main target of the paper is to establish the methodology for the amount of water estimation for particular engines. The methodology is introduced in the third chapter of the paper. Following the CFD model presented in the article and the methodology for water amount estimation, further research is presented in the fourth chapter of the article.

Keywords: temperature model; pressure model; CFD simulation; water injection system; centrifugal compressor

1 Introduction

Centrifugal compressors are nowadays extensively used in an area of applications including small jet engines, turboprop engines, power generators, and auxiliary power units, air conditioning, etc. [1] The construction of this device is an important factor in terms of the performance and efficiency of the system. Centrifugal compressors have to meet some specific requirements to ensure reliable engine operation, such as reliable work during the particular phases of the flight, durability, resistance to high loads. [2, 1] Jet engines have to also provide enough thrust during their operation for the specific phases of the flight. [1] There are some special cases when a range of phenomena is occurring, including noise generation, tip clearance, surge, and unsteady flow and, also the compressor performance. In order to investigate the mentioned processes numerical modeling using the CFD method is a convenient tool. CFD modeling is not the only appropriate way to estimate characteristics of the compressors that are already made but also for the modeling flow during the design phase of the centrifugal compressors. An important aspect of the numerical calculation process is evaluating the results according to the experimental data [2, 3].

Centrifugal compressor is a dynamic axisymmetric machine, which achieves a pressure rise by adding kinetic energy or velocity to a continuous flow of fluid through the impellers. Apart from the pressure rise in the impeller, the kinetic energy is converted to increase the pressure using the diffuser that is decreasing the flow speed and increasing the pressure [4].

The thrust of the jet engine is one of the fundamental parameters and is defined as the force to be the change in momentum of an object with a change in time. Momentum is the object´s mass times the velocity. [2] There are some ways for increasing the thrust and one of them is water injection into the compressor inlet of the engine. This principle has been well known for many years, so it is possible to implement the methodology for a particular engine. In order to design a water injection system, the thermodynamic parameters of the system should be well known, therefore, in this study, the flow parameters of the centrifugal compressor will be estimated. According to the flow parameters obtained from the CFD simulation, the methodology for the water injection system will be established [4]. In [4] different parameters of the centrifugal compressor were estimated using CFD analysis in ANSYS, also the impact of the flow on the compressor operation and phenomena were studied in [4]. Using a similar methodology the 3D modeling of the flow in the compressor is carried out. [6, 7].

There is a large number of centrifugal compressor types, investigated centrifugal compressor of the iSTC-21v jet engine consists of an inlet, impeller, diffuser, and the casing. [8] The compressor hub with impeller and stator vanes is shown in Figure 1. The model is partially modified for the CFD analysis. The impeller is semi-enclosed with 20 blades. Research is based on the data of iSTC-21v engine but the methodology is also applicable on TKT-1 turbojet (which is almost the same as iSTC-21v). The TKT-1 is a single spool, centrifugal compressor, axial turbine jet engine, which is equipped with an annular combustion chamber [9, 10].

The object of the CFD study is the compressor section with the inlet. The inlet of TKT-1 got a vertical cylindrical air inlet duct due to the testbed location [9] and that is the main difference in comparison with iSTC-21v engine. Such an inlet allows air inlet from the undisturbed regions and minimizes foreign object damage possibilities. The Centrifugal compressor with the diffuser is presented in Fig. 1. The 3D model of the compressor has to be modified for the CFD analysis [9, 1].

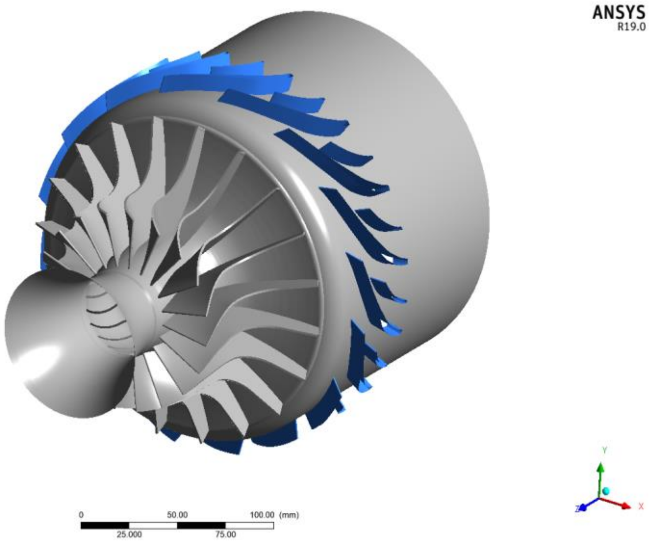


Figure 1
The geometry of the compressor TKT-1

2 Numerical Study

Currently, 3D virtual modeling is popular among designers due to its advantages, for instance, in the proposed study CFD modeling is an effective tool in order to estimate flow parameters. For the study, the ANSYS R19.0 software is used to carry out a steady-state analysis. Centrifugal compressor analysis involves the rotating flow domain. In ANSYS, the flow features associated with this problem can be also analyzed using the multiple reference frame capability. The analysis of the problem is including stationary and dynamic parts, compressor stator, and impeller, so the interaction of the parts is described in the following lines.

The analyses between the stationary and dynamic parts as the centrifugal compressor also often involves the examination of the transient effects. Transient effects are caused due to flow interaction between the rotor and stator parts. In the study, the sliding mesh capability of ANSYS is used to compute the transient flow in a centrifugal compressor. The interaction between the rotor and stator is modeled by allowing the mesh which is associated with the rotor to rotate relative to the stationary mesh associated with the stator part of the compressor [10, 11, 12].

In the presented paper the Singe Phase study is performed, which means the water is not included yet. In this type of study ANSYS fluent solves for an arbitrary scalar Φ_k the equation:

$$\frac{\partial \rho \phi_k}{\partial t}+\frac{\partial}{\partial x_i}\left(\rho u_i \phi_k-\Gamma_k \frac{\partial \phi_k}{\partial x_i}\right)=S_{\phi_k} \quad k=1,\ldots,N \tag{1}$$

Where Γ_k and $S_{\Phi_k}$are the diffusion coefficient and source term you supplied for each of the N scalar equations. Note that Γ_k is defined as a tensor in the case of anisotropic diffusivity. The diffusion term is, therefore ∇ . (Γ_k .Φ_k). For isotropic diffusivity, Γ_k could be written as $\Gamma_k I$ where I is the identity matrix. For the steady-state case, ANSYS Fluent will solve one of the three equations, depending on the method used to compute the convective flux [10, 11].

ANSYS compute the following equation in the case that the flux is not concluded:

$$-\frac{\partial}{\partial x_i}\left(\Gamma_k \frac{\partial \phi_k}{\partial x_i}\right)=S_{\phi_k} \quad k=1,\ldots,N \tag{2}$$

where Γ_k and S_{Φ_k} are the diffusion coefficient and source term you supplied for each of the 5 scalar equations. If the convective flux is to be computed with a mass flow rate, ANSYS Fluent will solve the equation:

$$\frac{\partial}{\partial x_i}\left(\rho u_i \phi_k-\Gamma_k \frac{\partial \phi_k}{\partial x_i}\right)=S_{\phi_k} \quad k=1,\ldots,N \tag{3}$$

The software ANSYS is working as most CFD software solutions based on the Navier-Stokes equations for viscous flow. These consist of the continuity equation, the momentum equation, and the energy equation (2). The equations are reflecting the changes in flow in each solution element of computational mesh [10, 11, 12]. Thermodynamics parameters can be estimated also analytically one of the methods is represented by the following formulas. The pressure in the inlet can be assumed as follows:

$$p_{0t} = p_0 \cdot \left(1 + \frac{\kappa-1}{2} . M_0^2\right)^{\frac{\kappa}{\kappa-1}} \tag{4}$$

Where p_0 is the atmospheric pressure, κ is the adiabatic exponent and M is the Mach number [10, 11, 12]. The temperature in the inlet can be estimated according to the formula:

$$T_{0t} = T_0 \cdot \left(1 + \frac{\kappa-1}{2} . M_0^2\right) \tag{5}$$

The T_0 is the atmospheric temperature. The pressure at the inlet of the impeller can be estimated by multiplying formula (4) by the pressure retention factor [10]. After modifying the formulas we can writhe the formula for the temperature at the outflow of the compressor as:

$$T_2 = T_{2t} - \frac{c_2^2}{2.c_p} \tag{6}$$

Where T_{2t} is the temperature at the inlet of the bladeless diffusor of the compressor, c_4 is the speed at this point and cp is the heat capacity [11]. Then the overall pressure at the outflow of the compressor can be estimated according to the formula:

$$p_2 = p_{2t} \cdot \left(\frac{T_2}{T_1}\right)^{\frac{n_1}{n_1-1}} \tag{7}$$

p_2 states for the overall pressure at the outlet of the compressor and p_{2t} [12]. The formulas are showing one of the ways for calculation thermodynamical parameters, which are essential for the methodology described in Section 3. However, the results would be not precise due to calculation only in a particular point that is why CFD methods for calculation are used. The model is described in the following chapters also this model will be used in Chapter 3.

2.1 CFD Modeling of Centrifugal Compressor

The object of an investigation is axisymmetric assembly, which means that the numerical model can be divided according to the symmetry into 20 symmetric parts. The model represents single-stage radial compressor consists of two blade row - rotor blades and stator vanes. The geometry (Figure 1) is adapted for the cyclic symmetry analysis as the case of this study, thus one sector is created for both components stator and rotor.

The computational domain consists of the inlet, rotating impeller, and stationary diffuser with frozen rotor interfaces between adjacent domains. The rotational speed of the impeller is set according to the experimental setup.

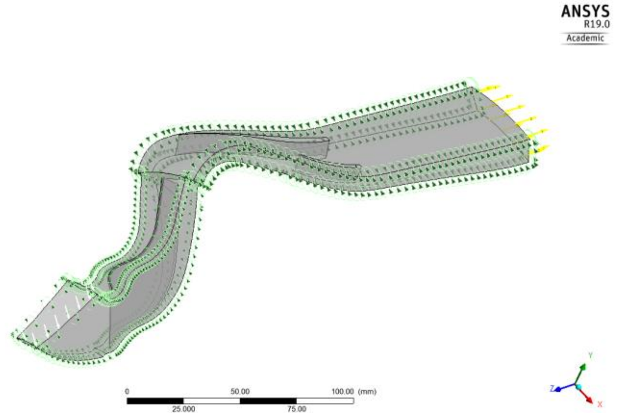


Figure 2
Computational domain of the compressor

2.2 Boundary Conditions

Boundary conditions are clearly seen in Figure 2, the temperature and pressure are applied as the boundary condition at the inlet interface. Between the adjacent blade passages, the periodic boundary condition is applied. The study is performed for teen different regimes according to the measured data (Figure 5). For instance one of them 3rd regime is presented, where boundary conditions are defined by the inlet pressure and temperature at 39 960 rpm.

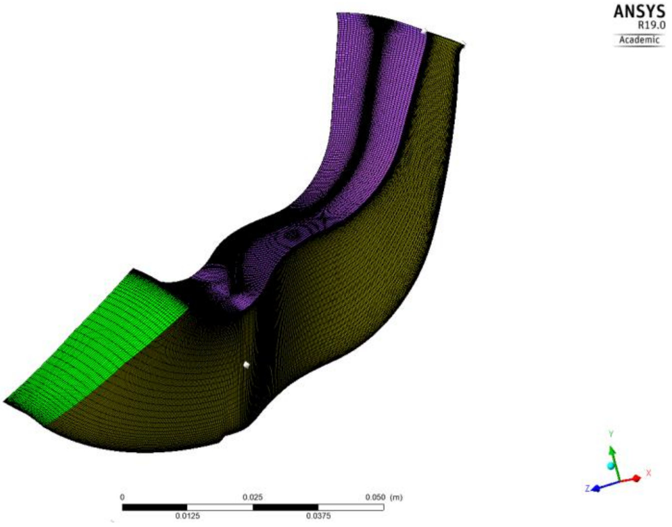


Figure 3
Radial compressor mesh

The mesh is created in the TurboGrid software. For mesh creation, it is essential to have some information about the system. Such information includes the location of the geometry files for all components (hub, shroud, and blades) the mesh topology type, and the distribution of mesh nodes. One of the crucial parts of the meshing process is the preparation of the geometry. Hub, shroud, and the blades have to be imported to TurboGrid software as coordinates with the axis of rotation Z. Subsequently, the geometry is generated in the software and the meshing process can be started. Once the geometry is defined, the next step is to create the topology that guides the mesh. Afterward the topology creation number of layers for both parts is defined. The mesh is created using the same ideology for both components stator and rotor and consists of the hexahedral elements. Mesh is represented by Figures 3 and 4. For the boundary condition application, ANSYS CFX software is selected, which is a highly accurate solver for robust solutions such as compressor assembly [10, 11, 12].

The mesh is generated for both components stator and roto separately and during the preprocessor process, the computational domain from the two parts is created.

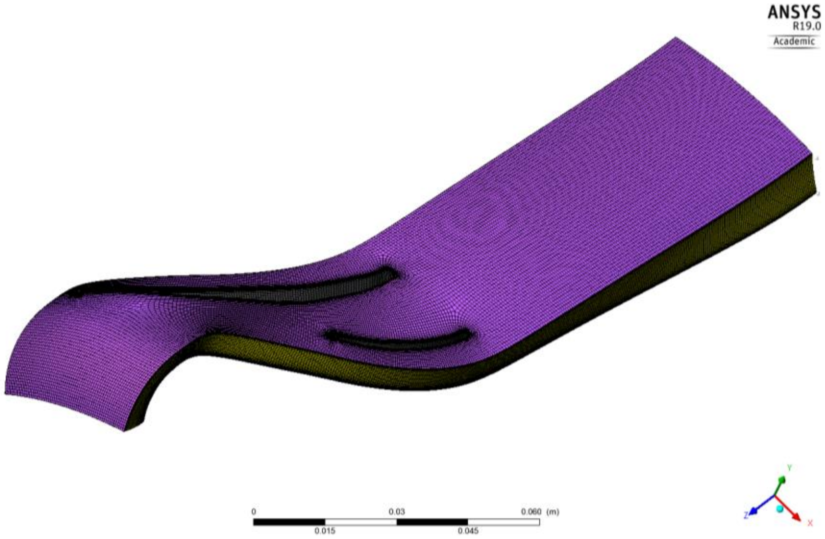


Figure 4
Computational mesh of the stator

The mesh of the stator consists of 927 405 elements and the rotor of the centrifugal compressor consists of 1 055 168 elements.

The analysis is performed for multiple regimes according to the measured data and for instance, in the proposed study the third regime is presented. The rpm (Figure 5), temperature, and pressure in the compressor section are measured. The boundary conditions are set according to the measured data in front of the centrifugal compressor. In Figure 5 the rpm for one run is measured using this data it is possible to simulate multiple regimes of centrifugal compressor operation by CFD methods [13].

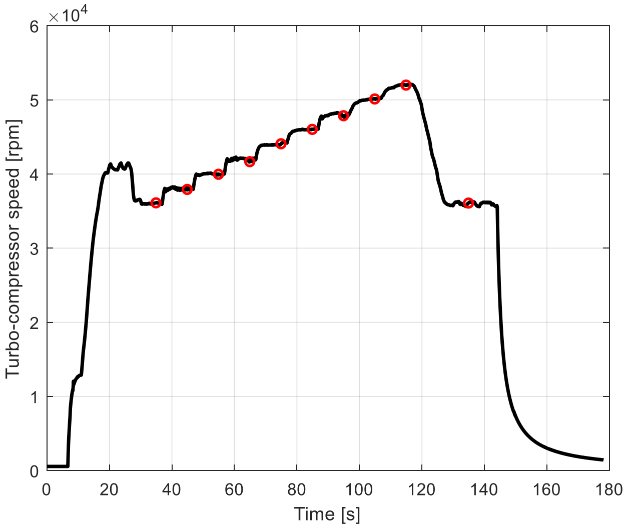


Figure 5
Measured data of iSTC-21v jet engine

In the presented study one regime is taken into an account in terms of CFD simulation. The third regime is simulated in ANSYS CFX software with 39 960 rpm and the measured pressure at this operating point is 276 414,6 Pa.

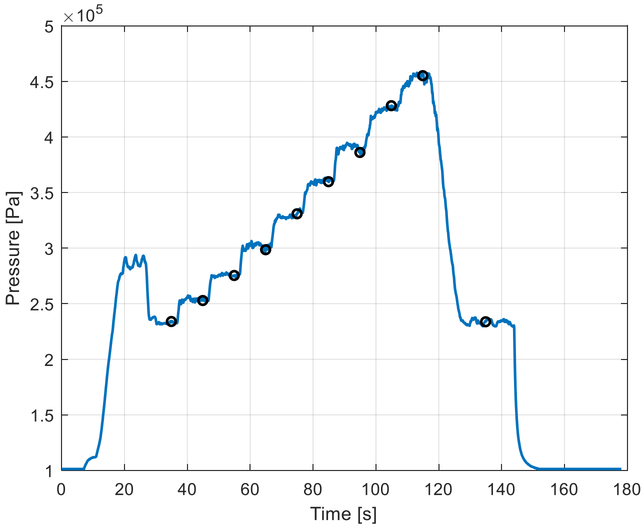


Figure 6
Measured pressure of iSTC-21v jet engine

2.3 Results of CFD Analysis

There are few most important parameters of the thermodynamic cycle of a jet engine that have a significant impact on the performance of the engine. During the analysis, the goal is to monitor mainly the temperature and pressure. These two parameters will be used for water injection methodology, which will be described in Chapter 3.

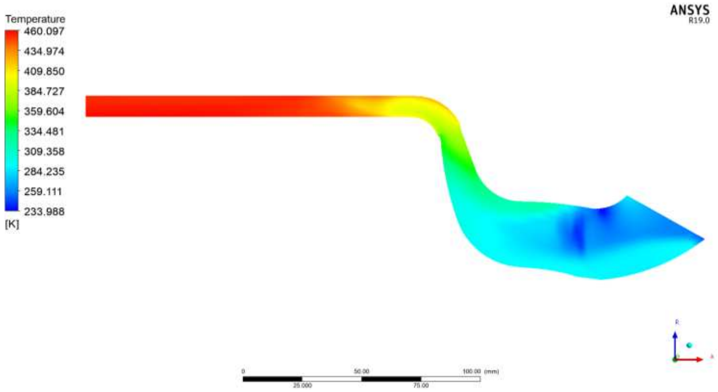


Figure 7
Temperature map of the compressor in meridional plane

The temperature is one of the crucial aspects when the thermodynamic cycle of the engine is investigated. The temperature field computed using CFX is shown in Figure 7. The results are comparable with the measurement of the temperature in iSTC-21v jet engine. The second estimated thermodynamical parameter during the CFD simulation is the pressure at the outflow of the compressor. Maximal pressure during the simulated regime of the engine is 277 000 Pa, which is in the comparison with the measurement of highly accurate results (Figure 8).

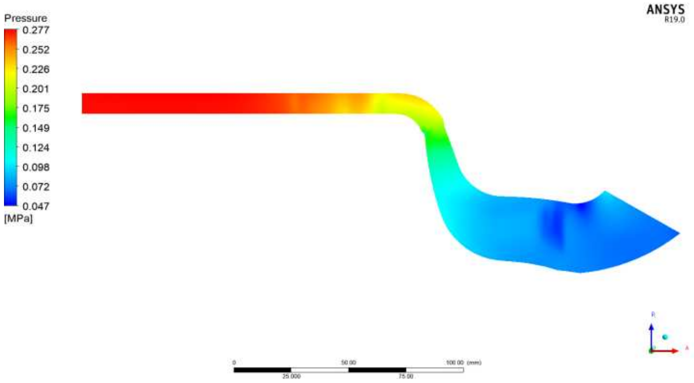


Figure 8
Pressure field of the compressor in meridional plane

In Figure 9 the Mach Number is presented on the blade to blade plot. The plot shows an increasing trend of Mach number from inlet to the impeller and in the diffusor area the Mach Number is decreased, according to the theory.

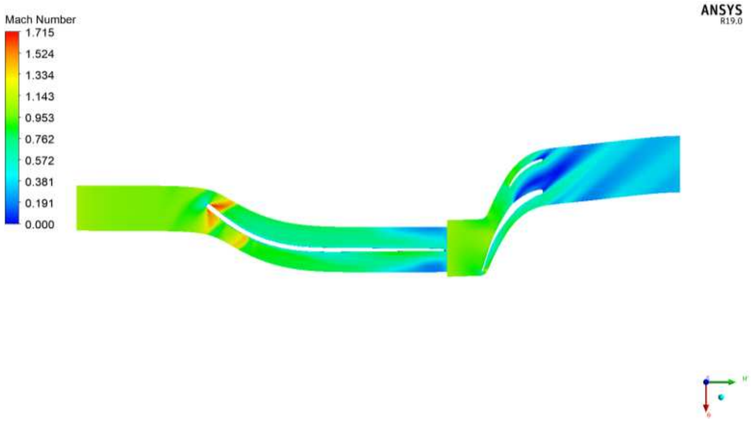


Figure 9
Mach Number of the compressor

Model is created for 10 regimes of the compressor operation, for instance, the comparison for the first four runs is in Table 1. The estimated temperature and pressure are compared using CFD methods. According to the results it can be clearly seen that the accuracy of the models is high and models of temperature and pressure will be used in the future while water injection system designing. Results for the rest regimes are in a similar range.

Table 1
Results Comparison of the Measured and Estimated Data

Run	Measured Temp. [°C]	Measured Pressure	Temp. [°C] CFD	Pressure CFD	Error Temp. [%]	Error P.[%]
1	423,8	234 000	421,4	237 100	0,566	1,325
2	427,3	252 800	425,5	251 110	0,421	0,669
3	437,5	276 414	447,3	277 000	2,240	0,212
4	443,7	298 600	450,1	299 100	1,442	0,167

In Figure 10 the behavior of the air in the centrifugal compressor of the iSTC-21v jet engine is shown. Streamlines are showing vortices that are generated behind the stator vanes. From the picture the character of the flow can be seen, there is no significant turbulence and we can consider also the numerical results as relevant.

The presented results with the rest of the computed states are necessary for the creation of the new methodology, thus the results will be used in further research. Based on the computed temperatures and pressures error coefficient will be estimated according to the comparison with the measured temperatures and pressures. The coefficient will be used for the model with the water injection system, the idea is described in Chapter 3.

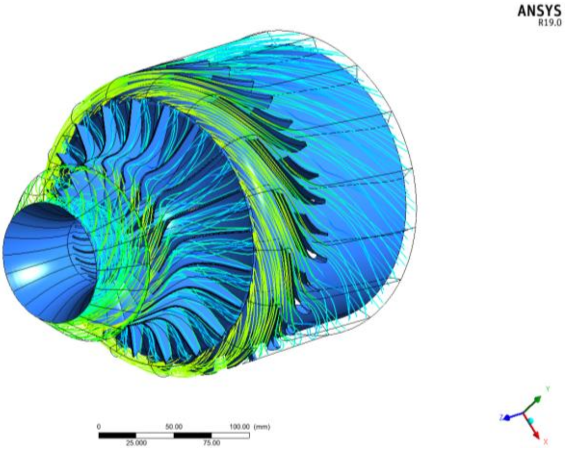


Figure 10
Streamlines of the air in the compressor

3 Methodology for the Water Injection System Design

The water injection into the compressor was already used for the thrust increase and multiple studies were performed, the amount of water was estimated mainly experimentally, however until nowadays there is no unified methodology for the water amount estimation. Also, the methodology for water injection which would be based on numerical models is not developed. Using obtained models from the previous chapters the method will be developed. The numerical model (Figure 2) from the previous chapters is used as one part of the scheme in Figure 11. There are some articles, in which the water injection impact on the compressor is investigated, for instance in the paper [14], the effect of the water vaporization on the compressor is investigated. In [15] are presented the results of a calculation study of motion and evaporation of water in the compressor flow path of a GT-009 gas-turbine. Other articles deal with the use of water for cooling engine parts [16], but also with the water injection into the other parts as for example the combustion chamber. The impact of the water injection is tested and the results are compared in [17] for the microturbine, the results are compared in characteristic maps and running curves of the compressor, the contribution of the water on the power augmentation was proved. Also, in paper [18] is presented the water effect of the environmental conditions on the thermodynamic cycle processes of a gas turbine, by using analytical relations. There are many studies of the water impact of the turbine performance, the evaporation of the water, experiments, etc. [19] but there is not a methodology of the water injection, which is based on the estimation of the amount of water using CFD methods.

The main goal of the proposed article is the development of a new methodology for assessing the amount of the injected water into the compressor for a thrust increase. Injected water is usually estimated experimentally [20, 21] but in the presented article the new approach using modern tools such as CFD modeling and experimental modeling is revealed. The methodology is based on the numerical simulations that are compared with experimental results. The main idea behind the methodology is presented in Figure 11.

As was already mentioned the whole process is depended on the numerical simulation, and that is the reason why the basic element of the scheme (Figure 11) is a 3D model. The methodology can be divided into three sections, the first one is the CFD simulation without the water injection, the second part is the simulation with the water injection and the third one is the experiment. For both simulations (with and without water injection) the thermodynamical parameters should be estimated for multiple regimes. For particular working conditions of the compressor, the temperature and the pressure should be computed.

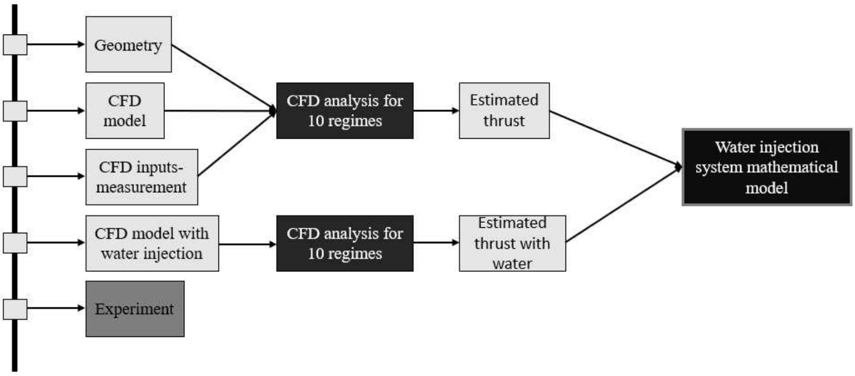

Geometry
CFD model
CFD inputs-measurement
CFD model with water injection
Experiment
CFD analysis for 10 regimes
CFD analysis for 10 regimes
Estimated thrust
Estimated thrust with water
Water injection system mathematical model

Figure 11
Methodology for the water injection system development

The computed temperature T_{2t} and pressure p_{2t} have to be compared with measurement in the real engine, so the error of CFD simulation can be estimated. It means we can assess the error coefficient by performing a sensitivity study between the CFD simulation (without water injection) and experimental data for multiple regimes.

Once the error coefficient of the CFD study is estimated it is possible to create an accurate CFD simulation of the compressor with a water injection system. It is possible to create a CFD simulation of the system for water injection and inject different amounts of water into the compressor (also perform few regimes the same as without the injection) and multiply the results with error coefficient [22]. Using this methodology the impact of water on the T_{2t} and p_{2t} can be obtained. Subsequently, according to the computed T_{2t} , p_{2t} (with water injection) [23] and measured data T_{2t} , p_{2t} it is possible to estimate the thrust of the engine [24, 25]. The described method in Figure 11 is applicable to each engine, by creating particular parts of the method for the particular engine. The method is innovative due to the fact that the amount of water is possible to estimate using numerical modeling [26, 27].

As for further research in the field, the methodology will be applied to the TKT-1 turbojet engine and according to the simulations, the water injection system will be designed. The number of the nozzles and their diameter will be estimated as well as the pressure of the water that will be injected into the compressor. The preliminary design of the system is shown in Figure 12, the methodology will be applied to the model in the following figure.

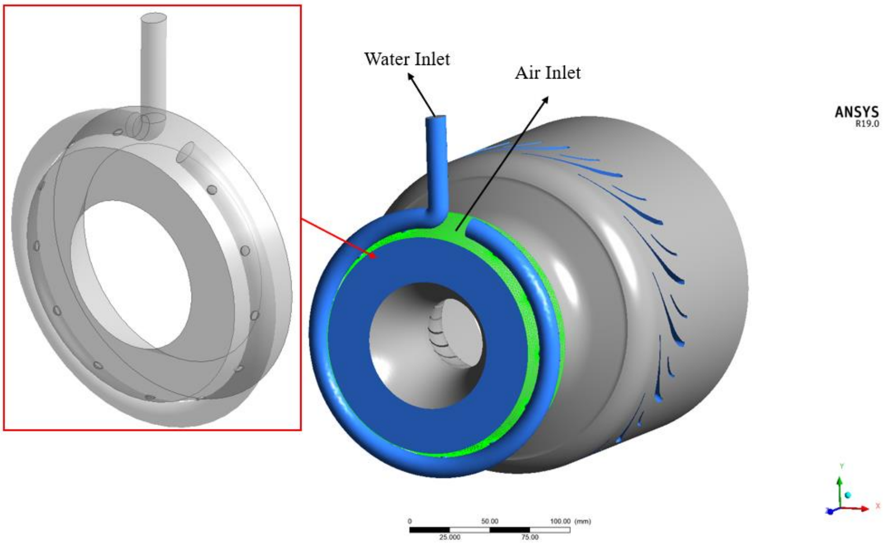


Figure 12
Preliminary design of the water injection system

Conclusion

Future research is strongly based on the proposed methodology in the article. According to the 3D CFD models, it is possible to create other analyses for the research in the field of water injection and its impact on the thermodynamic cycle of the jet engine. The CFD models created in the first part of the work are highly valuable, according to this model it will be possible to create mathematical dependence between the measured data and CFD models.

The results presented in the paper are highly valuable for further research in terms of the water injection system. Figures 7 and 8 are showing that the results are credible due to the small deviation between the measured data in the engine iSTC-21v and estimated temperature and pressure using CFD analysis. The analysis was performed for the third regime and the measured pressure during this operating point is 276 414,6 Pa, estimated pressure using CFD analysis is 277 000 Pa so the error between the analysis and measured data is 0.21%. Also, data in Table 1 represents the high accuracy of the created model of iSTC-1v turbojet. Based on the results presented in the first section of the paper the error coefficient will be estimated for the model with a water injection system (Figure 12). Using this dependence, it will be possible to estimate an accurate CFD model with a water injection system.

Further research will be based mainly on the proposed methodology and the created CFD models, the preliminary idea of the water injection system is in Figure 12. The ideology is based on the tested CFD model with an added domain in front of the compressor. The domain is a preliminary water injection system. According to the methodology in Figure 11 and the model in Figure 12 the amount of water will be estimated. The simulation will be carried out for multiple regimes (Figure 5) and the results will be used for the water injection system design. The number of injection nozzles, the water pipe, water pump, etc. will be estimated according to the results.

Acknowledgment

The project was supported by KEGA 044TUKE-4/2019 - A small unmanned airplane - the platform for education in the area of intelligent avionics. And by the Hungarian National EFOP-3.6.1-16-2016-00014 project titled “Investigation and development of the disruptive technologies for e-mobility and their integration into the engineering education” (IDEA-E). The work was also support of research and development potential in the area of transport means with ITMS project code: 313011T557

References

[1] E. Sundröm, B. Semlitsch, M. Mihaescu, Generation Mechanisms of Rotating Stall and Surge in Centrifugal Compressors, Flow, Turbulence Combustion, Vol. 100, pp. 705-719, 2018

[2] R. A. Togh, A. M. Tousi, M. Soltani, Design and CFD analysis of centrifugal compressor for a microgasturbine, Aircraft Engineering and Aerospace Technology, pp. 137-143, 2007

[3] M. T. Shobhavathy, P. Hanoca, CFD Analysis to Understand The Flow Behaviour of a Single Stage Transonic Axial Flow Compressor, Proceedings of ASMEGTINDIA, 2013

[4] M. Omidi, H. J. Liu, S. Mohtaram, H. T. Lu, H. Ch. Zhang, Improving Centrifugal Compressor Performance by Optimizing the Design of

Impellers Using Genetic Algorithm and Computational Fluid Dynamics Methods, Sustainability, pp. 1-18, 2019

[5] B. Semlitsch, M. Micaescu, Flow Phenomena Leading to Surge in a Centrifugal Compressor, Energy, Vol. 103, pp. 572-587, 2016

[6] X. Cheng and R. S. Amano, Studz of the Flow in Centrifugal Compressor, International Journal of Fluid Machinerz and Systems, Vol. 3, pp. 260-270, 2010

[7] Z. Sun, X. Zheng, Z. Linghu, Flow Characteristics of a Pipe Diffuser for Centrifugal Compressors, Journal of Applied Fluid Mechanics, Vol. 10, No. 1, pp. 143-155, 2017

[8] A. Hafaifa, R. Belhadef, M. Guemana, Modelling of surge phenomena in a centrifugal compressor: experimental analysis for control, Systems Science & Control Engineering An Open Access Journal, Vol. 2, pp. 631-641, 2014

[9] L. Főző, R. Andoga, K. Beneda, J. Kolesar, Effect on Operating Point Selection on Non-linear Experimental Identification of iSTC-21v and TKT-1 Small Turbojet Engines, Periodica Plytechnica Transportation Engineering, pp. 141-147, 2017

[10] ANSYS CFX-Solver Theory Guide, ANSYS, Inc. Release 16.0 January 2015

[11] ANSYS CFX Tutorials, Release 15.0, ANSYS, Inc., 788 pages, 2014

[12] ANSYS TurboGrid Tutorials, ANSYS, Inc. Release 16.0, 2015

[13] I. Roumeliotis, K. Mathioudakis, Evaluation of water injection effect on compressor and engine performanc eand operability, Applied Energy 87, pp. 1207-1216, 2010

[14] A. J. White, A. J. Meacock, An Evaluation of the Effects of Water Injection on Compressor Performance, Journal of Engineering for Gas Turbines and Power, 2004, pp. 748-754, 2004

[15] Yu. M. Anurov, A. Yu. Peganov, A. V. Skvortsov, A. L. Berkovich, and V. G. Polishchuk, Calculation Study of Water Injection on Compressor Characteristics of a GT-009 Gas-Turbine Installation, Thermal Engineering, Vol. 53, pp. 964-969, 2006

[16] U. Metha, J. Bowles, J. Melton, L. Huynh, P. Hagseth, Water injection pre-compressor cooling assist space access, The Aeronautical Journal, Vol. 119, pp. 145-171, 2015

[17] K. Suzuki, S. Nakano, K. Seki, Y. Tekeda, T. Kishibe, Effects of Water Injection on Generator Output Power Augumentation in a Microturbine, International Symposium on Transport Phenomena and Dynamics of Rotating Machinery, 2017

[18] R. Kadi, A. Bouam, S. Aissani, Analzye of gas turbine performances with the presence of the steam water in the combustion chamber, Revue des Energles Renouavelables, pp. 327-335, 2007

[19] S. Schuster, D. Brillert, U Martens, V. Hermes, F. K. Benra, Investigation of the evaporation process of liquefied hzdrocarbons in front of a compressor, International Szmposium on Transport Phenomena and Dynamics of Rotating Machinerz, ISROMAC, United states, 2019

[20] T. Ous, E. Mujic, N. Stosic, Experimental investigation on -water injected twin screw compressor for fuel cell humidification, International Journal of Hzdroge, 2011

[21] M. Obermuller, K. J. Schmidt, H. Schulte, D. Peitsch, Some Aspects on Wet Compression – Physical Effects and Modeling Strategies Used in Engine Performance Tools, Deutscher Luft- und Raumfahrtkongres, 2012

[22] R. Andoga, L. Főzo, L. Madarász, T. Karol', “A Digital Diagnostic System for a Small Turbojet Engine,” Acta Polytechnica Hungarica, Vol. 10, No. 4, 2013, ISSN 1785-8860

[23] E. Kiyak, A. Kahvecioglu, F. Caliskan, “Aircraft Sensor and Actuator Fault Detection, Isolation, and Accommodation,” in Journal of Aerospace Engineering, Vol. 24, No. 1, 2011, pp. 47-58

[24] X. Wei, G. Yingqing, “Aircraft Engine Sensor Fault Diagnostics Based on Estimation of Engine's Health Degradation,” Chinese Journal of Aeronautics, Vol. 22, No. 1, 2009, pp. 18-21, ISSN 1000-9361

[25] L. Főző, J. Judičák, R. Bréda, S. Szabo, R. Rozemberg, M. Džunda, Intelligent Situational Control of Small Turbojet Engines, Hindawi, International Journal of Aerospace Engineering, 2018

[26] M. Spodniak, M. Klimko, M. Hocko, P, Žitek, Low cycle fatigue numerical estimation of a high pressure turbine disc for the AL-31F jet engine, EPJ Web of Conferences, Vol. 143, 2017, ISSN 2101-6275, pp. 1-5

[27] Roman, R.-C., Radac, M.-B., Precup, R.-E., Petriu, E. M.., “Data-driven Model-Free Adaptive Control Tuned by Virtual Reference Feedback Tuning,” Acta Polytechnica Hungarica, Vol. 13 No. 1, 2016, ISSN 1785-8860, DOI: 10.12700/APH.13.1.2016.1.7

Cavitation Measurement in a Centrifugal Pump

Nikolett Fecser, István Lakatos

Széchenyi István University, Egyetem tér 1, 9026 Győr, Hungary
fecser.nikolett@sze.hu; lakatos@sze.hu

Abstract: One of the causes of centrifugal pump instability lies in the phenomenon of cavitation. Cavitation in the centrifugal pump can produce undesirable effects such as deterioration in hydraulic performance. In order to prevent the emergence of cavitation, it is required to know the onset point of the cavitation, in the pump. Our study presents the process and result of the cavitation measurement performed on a centrifugal pump in a closed loop test system. In this study, we applied CFD analysis, which serves as a means of measuring flow in the impeller of the centrifugal pump. The results gained from CFD analysis correspond, approximately, to the measured results.

Keywords: centrifugal pump; cavitation; NPSH; CFD

1 Introduction

Water is a determinative and vital element of human life. Therefore, its movement from the place of extraction to the consumer is an important task, and seeking solutions for it has been an issue for all of human development. The issue of water supply has reasonably been a complex task spanning thousands of years, which still has relevance in our current times. Water pumps determine nearly all aspects of our life directly or indirectly. Some of the areas in which water pumps play a significant role are, non-exhaustively, water supply, activities related to water management, health care, crisis and disaster management, firefighting, agriculture, industry, wine sector, energy production, building engineering, food industry, etc.

Water transportation systems are the critical elements of an infrastructure, the operation is also important from the aspect of security, therefore, the uninterrupted, trouble free, transportation, is of great importance. This demand is influenced by several factors, such as uninterrupted, steady, or irregular water consumption, as well as regular or irregular change in the flow characteristics of the transported water.

Cavitation occurring in pipelines, water transportation systems, and more typically, in water pumps, shut-off and regulating valves could be damaged due to this phenomenon. On one hand, it exerts noise and vibration effects on the environment and, on the other hand, generates physical damage to the machinery. Harmful vibrations during cavitation also damage the other related equipment and affects the overall operation, negatively.

Based on the data available in scientific literature and the foregoing research results, we can draw the conclusion that in the years between 1960 and 2000, this topic was more extensively dealt with, then subsequently less and less timely research explored this area, with a lesser consideration of the current practice [1-6].

In the course of our research work, we observed that there were few laboratories with measurements on this subject, as well as, accessible measurement results. With regard to measurement methods, we also encountered a number of difficulties, in the field of measuring device development and the elaboration of related measuring methods for instance.

2 The Phenomenon of Cavitation

On the grounds of the Knapp-Daily–Hammitts definition, cavitation occurs when bubbles grow in a static or flowing fluid and a collapse process follows this growth. If the collapse of the bubbles does occur, then due to the outgassing and formation of vapor bubbles characterizing the growth process, the phenomenon of bubbling or boiling occurs. On this basis, we distinguish gaseous and vaporous cavitation, which differ in their damaging effect on the material. The occurrence of gaseous cavitation is prior to the vaporous cavitation. The bubbles at a certain pressure start to grow awhile due to the dissolved gases from the fluid. As they reach a critical size, their static equilibrium resolves, and an explosive growth process starts. This is vaporous cavitation, which has a destructive effect [7].

Due to the damaging effects of cavitation, there has been an increasing demand for elaborating cavitation research. Cavitation can be a source of significant noise and vibration, and even the hydrodynamic characteristics can change. The phenomenon of cavitation can be examined based on these characteristics, as well.

The phenomenon of cavitation can occur for various reasons. We summarized the major underlying causes hereunder:

- High local liquid flow velocity
- Drop in pressure on the suction side
- Increase in geodetic suction head
- Temperature increase in pumped fluid

The two main types of measuring cavitation are the open and closed-loop test circuits. In open-loop systems the phenomenon of "cavitation detachment", namely when the pressure on the suction side decreases gradually causing a sudden drop in the pump delivery head (characteristic curve breaks down) and damaging effect occurs on the suction side of the machinery, cannot be measured precisely. It is difficult to change the pressure of open-loop equipment on the suction side. In our study, we deal with closed-loop cavitation testing.

2.1 Cavitation Measurement in Closed-Loop Test Circuit

Normally, cavitation testing is based on the ISO 9906 Standard and GOST 6134, however, it should be also taken into account which market the pump involved in the measurement is made for. The specifications and requirements of the relevant market are to be considered at all times [8] [9].

The cavitation measurement that we performed consists of the following steps:

- In the course of the measurement, we apply a few volume flow rates. We set at least ten different volume flow rates. These are called operating points.
- We read the values required for the characteristic curves:
    - Static / Differential pressure
    - Real-time shaft rotation speed
    - Electric power / torque input
    - Water level head (for cavitation measurements)
    - Water temperature
    - Environment temperature
- Apart from the above, other values requested by the customer will be measured. Generally, it is sufficient to do this only once per measurement after a 30-minute operation. These are as follows:
    - Bearing temperature
    - Bearing and/or baseplate vibration
    - Acoustic pressure values
    - Leechate

In the hydraulic laboratory, the complete characteristic range can be measured on the pump model built into the test circuit. The power output of the machine model itself cannot be too low, meaning a power of 50-100 kW. Its minimal size and method of measurement are regulated by international standards. These models strictly require the accurate scale modelling of the effective transition cross sections, and most importantly, that of the most relevant elements in power transformation (impeller, volute, guide vanes, back-plates, and draft tube). Our measurements were performed as follows: we performed the cavitation measurement on 10 operating points, which based on standards and practical experience: three points below the 97% pump head rise and seven of them from the start-up to the cavitation detachment. The throttle valve at on suction side was closed while on the discharge side was opened. The only change made, compared to the traditional cavitation measurement, was that a feed pump was built in on the suction side (it was necessary due to low water level). As a result, upstream pressure occurred on the suction side, but it did not affect the rest of the measurement. Figure 1 shows the photo of a closed-loop test circuit.

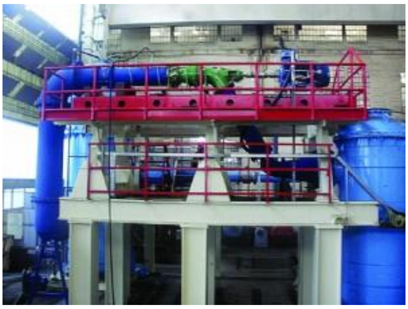


Figure 1
Closed-Loop Test Circuit

Figure 2 illustrates the schematic diagram of a closed-loop test circuit.

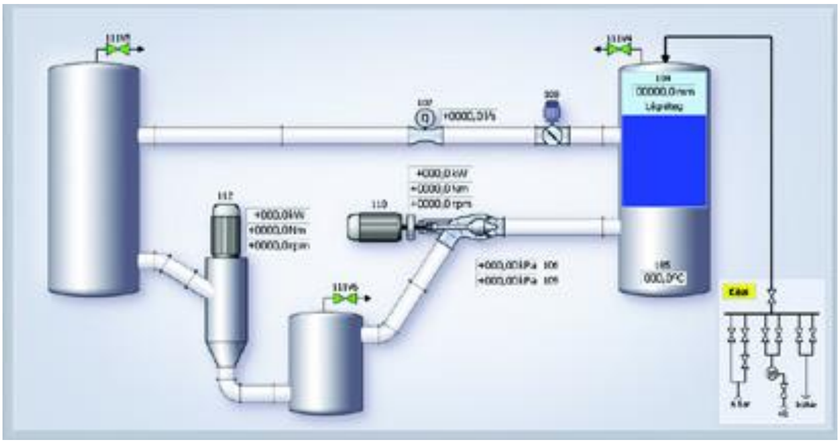


Figure 2
Schematic Diagram of a Closed-Loop Test Circuit

The measurement results were recalculated from the rated speed, not the real-time values were implemented. It was necessary to have a measurement result that is compatible with the tender curve characteristic submitted to the customer.

In the course of the measurement, we monitored how the results evolved as shown in Figure 3.

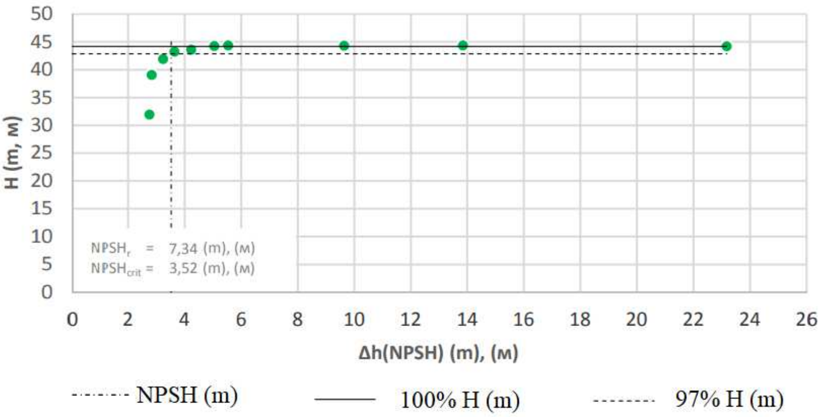


Figure 3
Variation of H-NPSH in the Course of the Measurement

Table 1 demonstrates the calculated data required for the H-NPSH curve of the pump.

Table 1
Calculated Data of the Pump

	Calculated Values	
No.	H [m]	Δh [m]
1	44.2	23.2
2	44.4	13.8
3	44.3	9.6
4	44.4	5.5
5	44.2	5.0
6	43.6	4.2
7	43.2	3.6
8	41.9	3.2
9	39.1	2.8
10	32.0	2.8

From the recalculated results we determined the H-ΔH (NPSH) curve, which is shown in Figure 4.

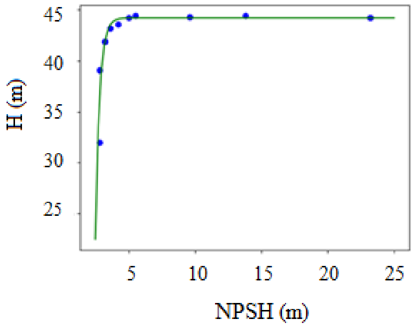


Figure 4
The H-ΔH (NPSH) Curve [Authors compilation]

One important characteristic of a centrifugal pump is the NPSH (Net Positive Suction Head). It specifies the minimum pressure at the pump inlet that is required by this particular pump type to operate cavitation-free, meaning the additional pressure, which is required to prevent the fluid evaporation and keep the fluid in the ideal state. From pump aspect the NPSH is affected by the impeller type and the speed of the rotation, whereas from environmental aspect by the fluid temperature, water coverage and atmospheric pressure. The main instruments used in the measurements were:

- Flow rate meter
- Shaft Torque Meter
- Manometer
- Tachometer

An inductive flow meter is installed in the system that ensures accurate and high repeatability flow measurement. Figure 5 shows the measuring scheme of the inductive flow meter.

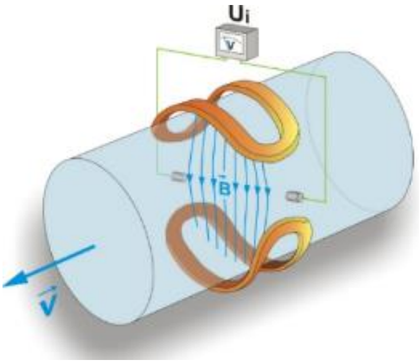


Figure 5
Inductive Flow Meter Measuring Scheme

The so-called, torque disk is used to measure the shaft power output in a closed loop test circuit.

Figure 6 shows the meter bridge principle, which allows the accurate measurement of slight changes in resistance.

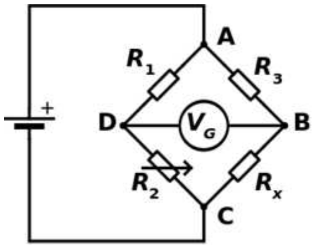


Figure 6
Meter Bridge Principle

In the course of the cavitation measurement, we measured three pressures as the main parameters:

- Suction side pressure
- Differential pressure
- Atmospheric pressure

Figure 7 illustrates the principle of pressure measurement.

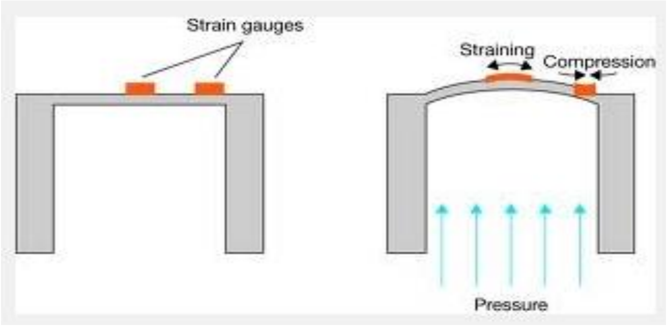


Figure 7
Principle of Pressure Measurement

The speed of rotation is measured by counting the impulses per unit time. The closed loop equipment has an incremental encoder mounted at the end of the motor shaft. It sends 4096 signals per revolution.

Figure 8 shows the measuring principle of the incremental encoder.

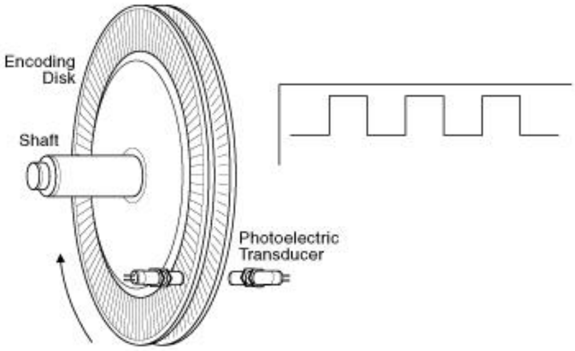


Figure 8
Measuring Principle of the Induction Flow Meter

Figure 9 demonstrates the change of the pressure line in the closed-loop test circuit. By defining one of its points, the line becomes more contoured. At point A, a tank is connected with an air cushion inside. The pressure of the air cushion is kept at a permanent rate by a controller. By gradually reducing the pressure of the air-cushion, the line is sinking in direct ratio. The operating point of the pump remains unchanged until the suction side pressure (specifically NPSH, net positive suction head) is close to the critical value. At this moment, the pump cavitates, its delivery head decreases. In general, we reduce the pressure in 8-10 steps to find the breakdown point. This is cavitation measurement.

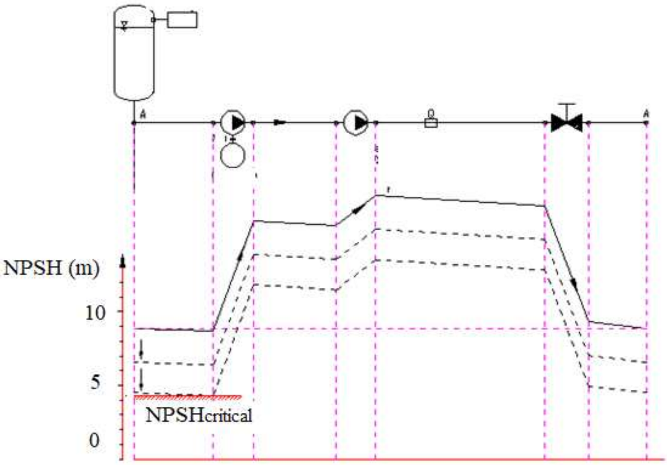


Figure 9
Change of Pressure Line in Closed-Loop Test Circuit

The pressure valve has a two-stage backpressure regulator. The first (rough) stage performs a huge pressure change in a quick way. The second (fine) stage operates after the first stage ends, and only fine adjustments are needed. Both the first and the second stage regulate between the pressures of 0.2 bars, absolute vacuum, and 2.0 bars, compressed air. The fist (rough) stage operates valves with ON/OFF function. The second (fine) stage adjusts the opening of small sized valves. The vacuum is provided by a pneumatic water-ring pump; the compressed air is supplied by the air network of the factory.

Measurement accuracy is a crucial factor. When performing technical acceptance procedures, one major error in the measurement can hinder the acceptance. When improving, we have to detect even slight changes in efficiency. In a closed-loop test circuit, the dominant source of error is the error of the water volume measurement, which has a rate of $\pm 0.2\%$. Its decrease therefore is a priority.

Within the resulting measurement error, the measuring accuracy of the efficiency is $\leq \pm 0.35\%$, and that of the critical cavitation rate is $\leq \pm 1\%$.

2.2 Cavitation Model

2.2.1 The Description of the Ansys Cavitation Model

For cavitation modelling, we chose the ANSYS-FLUENT CFD software, which is suitable for multiphase flow modelling [10].

The cavitation model is suitable for determining the cavitation characteristics of water turbines and pumps under vapor formation and condensation with low vapor content due to pressure change [11].

The equation of continuity for vapor phase is as follows:

$$\frac{\partial}{\partial t}(\alpha\rho_v) + \nabla\cdot(\alpha\rho_v v_v) = \pm\frac{\rho_v\rho_l}{\rho}\alpha(1-\alpha)\frac{3}{\left(\frac{\alpha \quad 3 \ 1}{1-\alpha 4\pi n}\right)^{\frac{1}{3}}}\sqrt{\frac{2}{3}\frac{\pm(p_v-p)}{\rho_t}} \qquad (1)$$

where α denotes vapor volume fraction, ρ_v vapor density, v_v vapor phase velocity, ρ_l liquid density, n bubble volume fraction, p_v saturated vapor pressure (function of temperature), p mixture pressure, and ρ mixture density.

Three multiphase cavitation models are available in ANSYS-FLUENT software system:

- Singhal model
- Zwart-Gerber-Belamri model
- Schnerr and Sauer model

The first simulation was based on the Singhal model. The results of the simulation and the measured values showed a large difference. We also performed a simulation based on the Schnerr and Sauer model, which produced much more accurate results. The Schnerr and Sauer model follows similar approach to the Singhal model in terms of determining the exact expression for the net mass transfer from liquid to vapor.

The Schnerr and Sauer model implements the following standard equations.

The general equation form of vapor volume fraction:

$$\frac{\partial}{\partial t}(\alpha\rho_v) + \nabla\cdot\left(\alpha\rho_v\vec{V}\right) = \frac{\rho_v\rho_l}{\rho}\frac{D\alpha}{Dt}$$

(2)

where α marks vapor volume fraction, p_v vapor density, ρ_l liquid density, and ρ mixture density.

$$R = \frac{\rho_v\rho_l}{\rho}\frac{d\alpha}{dt} \tag{3}$$

$$\alpha = \frac{n_b\frac{4}{3}\pi R_B^3}{1+n_b\frac{4}{3}\pi R_B^3} \tag{4}$$

where n_b is bubble quantity.

$$R = \frac{\rho_v\rho_l}{\rho}\alpha(1-\alpha)\frac{3}{R_B}\sqrt{\frac{2}{3}\frac{(p_v-p)}{\rho_l}} \tag{5}$$

where R represents the total interphase mass transfer rate per unit volume. R_B is the bubble radius.

$$R_B = \left(\frac{\alpha}{(1-\alpha)}\frac{3}{4\pi}\frac{1}{n}\right)^{\frac{1}{3}} \tag{6}$$

$$R_e = \frac{\rho_v\rho_l}{\rho}\alpha(1-\alpha)\frac{3}{R_B}\sqrt{\frac{2}{3}\frac{(p_v-p)}{\rho_l}} \tag{7}$$

where R_e is a mass transfer source term connected to the growth and collapse of the vapor bubbles.

$$R_c = \frac{\rho_v\rho_l}{\rho}\alpha(1-\alpha)\frac{3}{R_B}\sqrt{\frac{2}{3}\frac{(p-p_v)}{\rho_l}} \tag{8}$$

where R_c is a mass transfer source term connected to the growth and collapse of the vapor bubbles.

2.2.2 Simulation Performed with Ansys Fluent Software

Using ANSYS-FLUENT v19.1 software system, the simulation of the two-phase cavitation turbulent flow was run on the specified geometry by following the discretization method based on the finite volume method (FVM). The first simulation followed the Singhal model. There was a large difference between the simulation and the measured values, for instance, compared to the measured values, a twofold dynamic pressure occurred on the vane leading edge.

Figure 10 shows the spectra of the pressure values on the impeller.

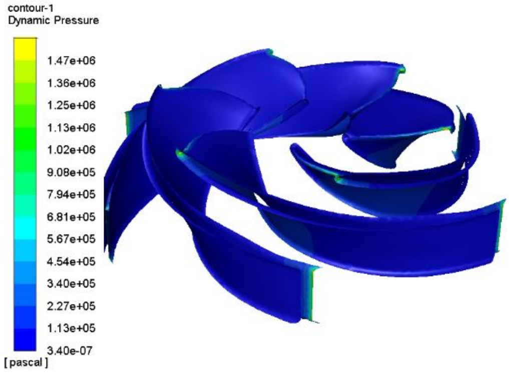


Figure 10

Spectra of Pressure Values on the Impeller [Author compilation]

Figure 11 depicts the velocity rate spectra on the impeller.

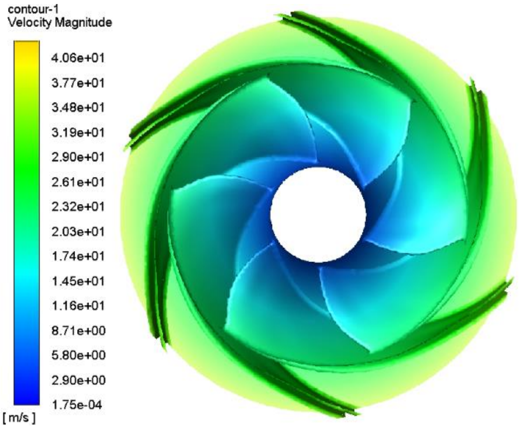


Figure11

Velocity Rate Spectra on the Impeller [Author compilation]

We also performed the simulation based on the Schnerr and Sauer model, which produced much more accurate results. We ran the two-phase turbulent flow with Realizable k-epsilon turbulence model considering standard wall function. In the two-phase flow the primary phase was the liquid and the secondary was the vapor phase. The friction coefficient was calculated by applying the Schiller-Naumann Correlation, whereas the cavitation flow was simulated with the Schnerr-Sauer model. We made the simulation on the assumption of a non-stationary flow. We evaluated the results by using CFD-Post software system.

Figure 12 presents the spectra of pressure rates on the impeller.

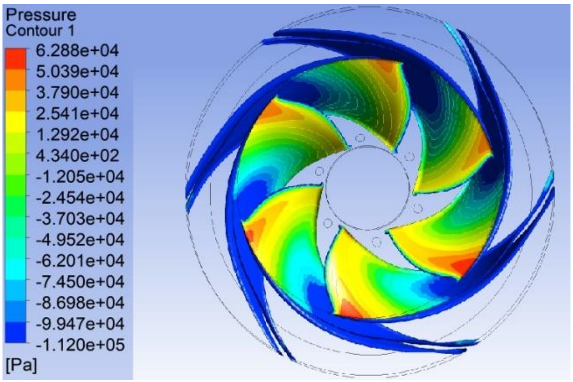


Figure 12
Pressure Rate Spectra on the Impeller [Author compilation]

Figure 13 shows the liquid velocity rate spectra on the impeller.

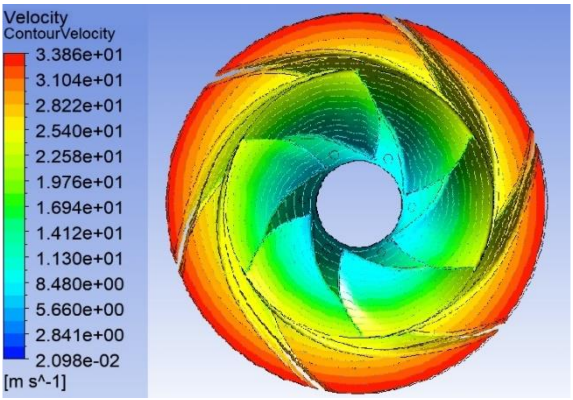


Figure 13
Spectra of Liquid Velocity Rate on the Impeller [Author compilation]

Figure 12 shows the pressure rate spectra on the impeller. The pressure rate is -62.01 kPa on the vane leading edge, 112 kPA on the vane trailing edge. The measured pressure rate on the suction side -62.5 kPa and on the discharge side 262.3 kPa, which indicates that the simulation closely corresponds to the measured values.

Figure 13 depicts the spectra of liquid velocity on the impeller. The liquid velocity on the vane leading edge is 2.52 m/s, whereas on the vane trailing edge 5.66 m/s. The values measured on suction and discharge side are around 5.12 m/s, which also demonstrates that the simulation closely approximates the measured values.

Conclusions

Pumps are designed and manufactured so that they operate remotely from the state of cavitation therefore, the factors affecting the initial stage of cavitation are important. Our study dealt with the phenomenon of cavitation occurring in centrifugal pumps. We presented the process and result of a closed-loop cavitation measurement performed with a centrifugal pump in a hydraulic laboratory.

In this study, we used CFD analysis. The results of the CFD analysis are approximately the same as the measurement results. Based on the measurements, we also performed a cavitation simulation that allows to determine the processes occurring during cavitation and the effects on the pump.

Our future goal is to further analyze and refine the simulation results and try to optimize the results obtained. In the course of our research, we contacted several companies and received the information that they applied simulation only in the event of a failure. As a result, we concluded that it would be practical to implement simulations, when measuring at the end of the production, since the results could be beneficial for the subsequent developments and also, filter out the dangerous structural elements of the pump.

We would also like to examine whether there is this relationship between our research and areas of transport science involving on pipeline systems, as a subsystem of transport. In addition to pumps, turbines also play an important role in everyday life. The types of turbines are gas turbines, steam turbines, water turbines and wind turbines, all having different flow mediums. Another important subject of research could be the examination of the cavitation in wind turbines and the related dynamic effects of the air [12-18].

References

[1] A. A. B. Al-Arabi, S. M. A. Selim, R. Saidur, S. N. Kazi, G. G. Duffy: Detection of Cavitation in Centrifugal Pumps , Australian Journal of Basic and Applied Sciences, 5(10): 1260-1267, 2011 ISSN 1991-8178

[2] M. ČDINA: Detection of cavitation phenomenon in a centrifugal pump using audible sound, Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva 6, 1000, Ljubljana, Slovenia, Received 26 February 2002, Accepted 25 July 2002, Available online 13 August 2003. DOI:10.1006/MSSP.2002.1514, Corpus ID: 122177370

[3] Tan Lei1, Zhu Bao Shan1, Cao Shu Liang1, Wang Yu Chuan1, Wang Bin Bin: Numerical simulation of unsteady cavitation flow in a centrifugal pump at off-design conditions, First Published December 2, 2013, https://doi.org/10.1177/0954406213514573

[4] S. R. Shah, S. V.Jain, R. N. Patel, V. J. Lakhera: CFD for Centrifugal Pumps:A Review of the State-of-the-Art, Available online 25 April 2013, https://doi.org/10.1016/j.proeng.2013.01.102

[5] T. Capurso, L. Bergamini, M. Torresi: Design and CFD performance analysis of a novel impeller for double suction centrifugal pumps, Received 9 May 2018, Revised 22 October 2018, Accepted 2 November 2018, Available online 9 November 2018, https://doi.org/10.1016/j.nucengdes.2018.11.002

[6] Pranav Vyavahare, Lokavarapu Bhaskara Rao, Nilesh Patil: CFD Analysis of Double Suction Centrifugal Pump with Double Volute, Received 26 August 2017; accepted after revision 05 December 2017, https://doi.org/10.3311/PPme.11425

[7] Robert T. Knapp, James W. Daily, Frederick G. Hammitt: Cavitation,

McGraw-Hill Book Company, New York, 1970

[8] Rotodynamic pumps. Hydraulic performance acceptance tests. Grades 1, 2 and 3 (ISO 9906:2012), MSZ EN ISO 9906:2013

[9] Rotodynamic pumps. Methods of testing, GOST 6134-87, 2007

[10] ANSYS-FLUENT v19.1 Tutorial guide, 2009

[11] József Nyers, G. Laszló: Analysis of heat pump's condenser performance by means of mathematical model, Acta Polytechnica Hungarica, 2014, ISNN 1785-8860

[12] László Pokorádi: Fluid transport system linear parameter sensitivity analysis, Szolnok Scientific Publications XVII, pp. 43-55, 13 p. (2013)

[13] László Pokorádi, Boglárka Molnár, Monte-Carlo Simulation Analysis of parametric uncertainity of hidraulic system, In: Pokorádi, László (ed.) Technical Science in the Northeast Hungarian Region 2013: conference presentations, Debrecen, Hungary: Debrecen Academic Committee Technical Committee (2013) 518 p. pp. 171-180, 10 p.

[14] László Pokorádi, The uncertainty analysis of the pipeline system, Polytechnical University of Bucharest. Scientific Bulletin. Series D: Mechanical Engineering 73: 3 pp. 201-214, 14 p. (2011)

[15] Moustafa El-Gindy, Hossam Ragheb, Rear wing spoiler effects on vehicle stability and aerodynamic performance, International Journal of Vehicle Systems Modelling and Testing, 2020

[16] Szauter, Ferenc; Péter, Tamás; Bokor, József Complex Analysis of the Dynamic Effects of Car Population Along the Trajectories, In: ASME (szerk.) ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference: Mechatronics for Electrical Vehicular Systems. New York (NY), American Society of Mechanical Engineers (ASME) (2016) Paper: DETC2015-47075, 6 p.

[17] Szauter, Ferenc; Péter, Tamás; Bokor, József Examination of complex traffic dynamic systems, ACTA TECHNICA JAURINENSIS 6: 2 pp. 11-17, 7 p. (2013)

[18] Szauter, Ferenc; Peter, Thomas; Bokor, József Investigation of a Complex Transport Dynamics System, In: Péter, Tamás (ed.) Innovation and Sustainable Surface Transport Conference: IFFK 2012 Budapest, BMF, Hungarian Academy of Engineering, (2012) pp. 108-111, 4 p.

Identifying the Possibilities for Superior Recovery by Pelletization of Industry Related Small and Powdery Iron Containing Waste

Sorina Gabriela Șerban, Imre Kiss

University Politehnica Timișoara, Department of Engineering & Management,
Faculty of Engineering Hunedoara, 5, Revolutiei, Hunedoara, Romania
e-mails: sorina.serban@fih.upt.ro, imre.kiss@fih.upt.ro

Abstract: Maximal recovery of iron contents waste is an important problem in siderurgy, since its needed transformation into by-products, thus into valuable economic goods, can lead to a rational exploitation of raw material, thus, ensuring industrial needs, as well as environmental protection. Powder waste can be processed in the form of pellets or briquettes and then is used in the steelmaking processes. Thus, the powdery ferrous wastes resulting from the materials industry, from the point of view of granulation, corresponds to processing by pelletization. In line with the superior recovery of waste, our research has focused on identifying the possibilities for the pelletization of industry related the small and powdery iron containing waste that exists in very large quantities, in Hunedoara County area (Hunedoara and Calan) and beyond. This article presents the results relating to the possibilities of pelletization, of industrial small and powdery iron containing waste, from the steel industry (steel dust, mixed agglomerating and blast furnace dust, landfilled in siderurgical ponds, in vicinity of the former industrial areas of Hunedoara and Calan, Hunedoara County). The origin of the material is the minerals (bauxite residue / red mud, landfilled in vicinity of Oradea, Bihor County) or from metal surface treatment operations (coating anticorrosive dust, collected in Oradea) in the industry. The waste utilized in the laboratory experiments was processed using a series of installed equipped in the laboratories at the Faculty of Engineering in Hunedoara, University Politehnica Timisoara.

Keywords: steel dust; mixed agglomeration and blast furnace dust; bauxite residue (red mud); anti-corrosive sludge; pelletizing process; experimental pellets; compression test

1 Introduction

In the production stream of steelmaking, steel operators generate iron containing waste, continuously, in appreciable quantities, proportional to their steel production [1-7]. The main small and powdery ferrous waste, with the reported iron content are: steel dust, agglomerate dust and sludge, blast furnace dust and sludge, converter dust and sludge, electrical steel mill electrofilter dust or a mixture of water with fine particles of material resulting from the mechanical preparation of ores or coal [7-13].

In addition to the ferrous waste resulting from the metallurgical industry, waste with ferrous content is also generated from other industries (mining and processing of minerals, energy sector, chemical industry, etc.), such as: pyritic ash from chemical industry, iron concentrate from thermal power plant ash, other waste such as lime dust, dolomite, bauxite residue (red mud), anti-corrosive sludge, galvanic sludge, etc. [1-6, 8-10].

Waste resulting from various technological processes, in particular those resulting from industrial metallurgical, mining or ore processing processes, can be processed by pelletization, briquetting and/or agglomeration, meaning that it can be used in the development of cast iron and steel production [7-26]. By processing this waste, as well as processing it into by-products, qualitatively suitable for use as raw materials or auxilliary materials in the steel industry, the areas currently occupied, can be rendered back to the surrounding landscape, thus, contributing to the greening of the general environment, as well as to the expansion of the raw material base [1-10, 13].

Pelletizing is an integral non-pressure agglomeration technique, in the mineral processing industry, that transforms various iron content materials into a premium by-products, a larger particles/agglomerates form “iron pellets”, suitable for use in an iron-making furnace, at a steel mill, such as a blast furnace or electric arc furnace [7, 9-10, 13-26]. This processing transforms the small and powdery ferrous materials into a concentrated raw material suitable for pelletizing, as opposed to being disgarded. Such minerals or raw materials, that are commonly pelletized include [1-10, 13]:

- Mined iron ore
- Other source of iron such as dust collected from blast furnaces, steel plants, convertors or agglomerating sectors
- Sludge collected from mine sites such as limestone, alumina, bauxite etc.
- Sludge collected from mineral processing sectors such as red mud, galvanising sludge, coating anticorrosive sludge etc.

In order to pelletize the small and powdery ferrous materials, a binder is needed, bentonite clay being a common choice [7, 9-10, 13-26]. Various additives may also be included with the feedstock to improve performance in the steel making (graphite, lime, dolomite etc.).

In general, three primary tehnological phases occur in any pelletizing process:

- The mixing process (Figure 1) – that sets the stage for creating a homogeneous mixture from the small and powdery raw materials that will allow for a uniform by-product (i.e. pellets) to be created.

- The pelletizing (also named “balling”) process, in which the agglomeration operations are made using a pelletizing equipment (or balling device). At this stage, the obtained balls are referred to as “green” pellets.
- The induration process (Figure 1) – which involves a thermal treatment that heats the “green” pellets to just before their melting point, causing them to become extremely hard. In fact, the “green” pellets will be fired in order to cure into their hardened form, as strong “fired” pellets.

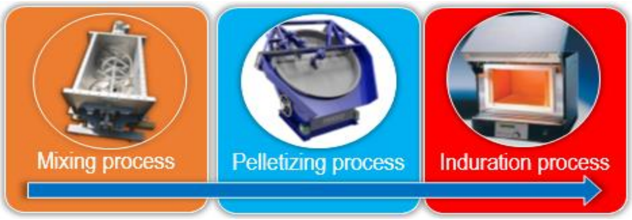


Figure 1
The primary tehnological phases in any pelletizing process

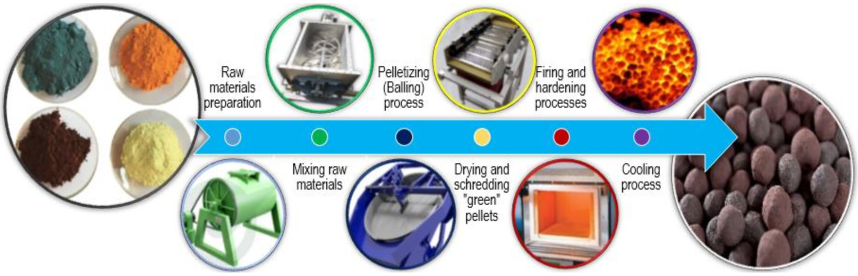

Raw materials preparation
Pelletizing (Balling) process
Firing and hardening processes
Mixing raw materials
Drying and schredding "green" pellets
Cooling process

Figure 2
A complete pelletizing process

Figure 3
The by–products flow: “green” pellets – “fired” pellets – “cooled/hardened” pellets

A complete pelletizing process is typically comprised of a series of unit operations in a specific process sequence (Figure 2), in a generalized approach that can be divided into the following stages [7, 9-10, 13-26]:

- Stage of procurement and receiving raw material, the location of a pelletizing unit affecting the method of receiving raw materials such as fine and pulverous iron oxides content dust or sludge, additives and binders

- Stage of pre-processing (pre-treatment process and preparation of raw materials), suitable for pelletizing (typically involves dry grounding into finer size and screening the pulverous materials)
- Stage of pre-wetting operation that includes adding an adequate amount of water homogeneously into the dry ground material to prepare pre-wetted material suitable for balling with optimum moisture.
- Stage of proportionating and mixing the components (main ferrous content wastes, additional ferrous content wastes) with other materials called binders and additives, creating a homogeneous mixture to prepare the green balls and to achieve the required quality in final by-product (i.e. iron oxide pellet).
- Stage of pelletizing on disc pelletizer (balling process) and producing the "green" pellets from the pre-wetted mixed material prepared in the previous operation (Figure 3).
- Stage of post-processing (mainly consisting in mechanical screening operations), the "green" balls produced on the disc pelletizer being not uniform in diameter, a significant portion of the discharge (about 70%) being smaller (undersized) than target size and must be returned to disc pelletizer, after screening. If oversized balls are resulted (bigger than target size) will be returned to stage of pre-processing for dry grounding into finer size and re-mixing.
- Stage of indurating process (firing and hardening operations) the "green" pellets, establishing the binding of fine and pulverous particles at an elevated temperature, and producing the "fired" pellets (Figure 3). This is typically preceded by a drying stage, which may be carried out in the induration unit, or in a separate device, mainly in laboratory muffins furnace. After induration, the pellets can be cooled or left to cool.
- Stage of testing, often an essential part of the development of a successful pelletizing operation.
- Stage of handling of by-product (typically involves final screening for end user size requirements and storage).

2 Quality of Pellets

As a by-product, a typical iron pellet is roughly spherical in shape, measuring from 6 mm to 20 mm in diameter and having a minimum compression resistance in range of 180-200 daN/pellet or more, although some variations in these typical parameters can be specified and targeted in their preparation process [7, 9, 10, 13].

The quality of the pellets is influenced by the nature of the solid starting materials, the type and quantity of additions, the humidity of the mixture in the formation of “green” pellets and their hardening treatment [7, 10, 13]. Taking into account the demands to which pellets are subjected during transport, from the pelletization equipment to the elaboration aggregate, as well as the influence of temperature and transformations that take place in the processes of casting irons and steels manufacturing, it is necessary that they correspond to qualitative characteristics namely (chemical composition and optimal dimensions), as well as a good resistance to compression etc. [7, 10, 13].

Compression resistance is a factor of qualitative assessment of the pellets and is expressed in daN/pellet. Compression resistance is a feature that depends on the mineralogic composition of the used materials, the finesse of this material (the granulation), the addition of binder and additives, the working conditions applied and the size of the pellets [7, 10, 13]. For pellets with a diameter between 10 and 15 mm the minimum compression resistance must be in range of 180-200 daN/pellet.

Plasticity is also an important property of “green” pellets during their formation. The compression resistance of raw pellets is somewhat influenced by the plasticity of “green” pellets [7-13]. A certain degree of plasticity is required to support the growth rate of “green” pellets. If plasticity increases, the compression resistance of wet pellets decreases. The amount of moisture is required to create a certain degree of plasticity that depends on properties of particles and their distribution.

3 Materials & Methodology

The wastes from mineral processing related sectors (bauxite residue/red mud, coating anticorrosive sludge), together with the wastes from siderurgy related sector (steel dust and the mix of agglomerating/furnace dust) were subjects of the pelletizing process, in presence of graphite, used as the reducing agent, respectively bentonite and lime, used as binders [7, 9-26].

Figure 4

Main solid components: (a) steel dust / electric arc furnace dust, (b) blast furnace dust / agglomerating dust (mix), (c) bauxite residue / red mud, (d) coating sludge / anticorrosive sludge

Figure 5
Binders, additives and water: (a) bentonite clay, (b) graphite powder, (c) hydrated–lime, (d) water

For the pellets production, the ferrous waste used (Figure 4) and the manufacturing formulations are shown in Table 1, which it also shows the proportions of bentonite and lime (experiment #5 and #6), used as binders (Fig. 5), as well as the proportion of graphite used as a reducer (experiment #5 and #6). For each formulation there were produced 4 series of pellets, so a total of 24 pelletization charges.

At preparation of the experimental recipes, we have in consideration the following preliminary remarks:

- The choice of waste was made according to its chemical and mineralogic composition. The purpose was to obtain the pellets by varying the amount of very fine waste, the binder content and the amount of water used. The binder was used to increase the resistance of the pellets, in laboratory experiments being used bentonite and lime [7, 10, 13]. The results are presented in Table 1.
- With the increase in the addition of bentonite and the proportion of fine fraction, the values for compression resistance also increase. Depending on the value of the compression resistance we want to obtain and the proportion of fine fraction in the pelletization shear we choose the addition of bentonite and water. Increasing the proportion of fine fraction in the pelletization shear, requires an increase in the addition of bentonite [7, 10, 13].
- The recipes compositions and their percentual participation for the pelletizing charges in Table 1 and Figures 6-8 are presented.

The 1st experimental experimental recipe is prepared with steel dust as main component (72%), and equal percentages of red mud and anti-corrosive mud (12%). This experimental recipe is prepared without agglomerating–furnace dust, graphite and lime, but is used 4% bentonite, as is presented in Table 1 and Figure 6. The 2nd experimental recipe is prepared with agglomerating–furnace dust as main component (72%), and equal percentages of red mud and anti-corrosive mud (12%). This experimental recipe is prepared without steel dust, graphite and lime, but is used same 4% bentonite (Table 1 and Figure 6).

Table 1
Participation of solid components in experiments / recipes

Components	Participation in experiments / recipes, [%]					
	Sample #1	Sample #2	Sample #3	Sample #4	Sample #5	Sample #6
Steel dust	72	0	72	0	70	0
Agglomerating/ furnace dust	0	72	0	72	0	70
Bauxite residue (red mud)	12	12	24	0	0	24
Anti–corrosive protection mud	12	12	0	24	24	0
Graphite	0	0	0	0	1	1
Bentonite	4	4	4	4	3	3
Lime	0	0	0	0	2	2

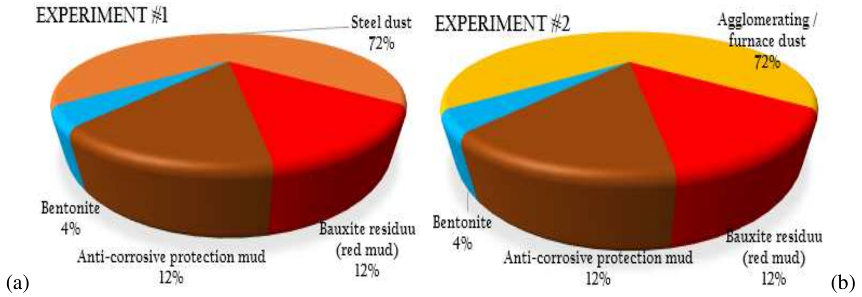


Figure 6
Participation (percentage) of solid components in Recipe#1 (a) and in Recipe #2 (b)

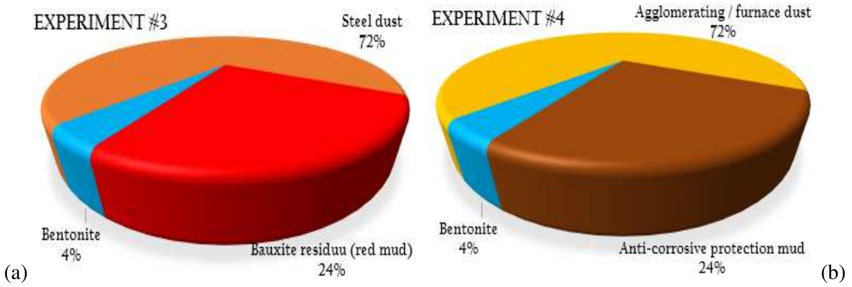


Figure 7
Participation (percentage) of solid components in Recipe #3 (a) and in Recipe #4 (b)

The 3rd experimental recipe is prepared with steel dust as main component (72%) and red mud (24%). This recipe is prepared without agglomerating–furnace dust, anti-corrosive mud, graphite and lime, but is used the same 4% bentonite (Table 1 and Figure 7). The 4th recipe is prepared with agglomerating–furnace dust as main component (72%) and anti-corrosive mud (24%). This experimental recipe is prepared without steel dust, red mud, graphite and lime, but is used the same 4% bentonite (Table 1 and Figure 7).

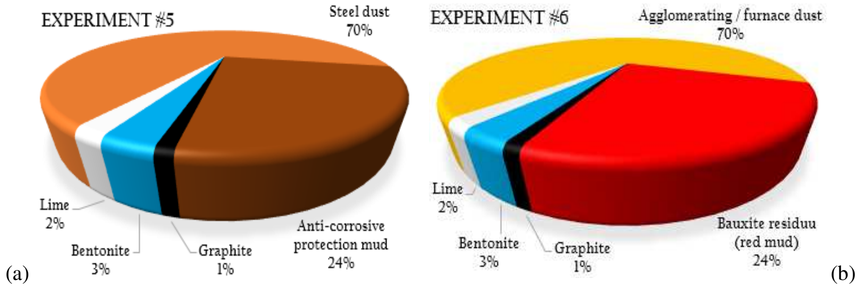


Figure 8

Participation (percentage) of solid components in Recipe #5 (a) and in Recipe #6 (b)

The 5th experimental recipe is prepared with steel dust as main component (70%) and anti-corrosive mud (24%). This recipe is prepared without agglomerating–furnace dust and red mud, but graphite (1%), lime (2%) and bentonite (3%) are used (Table 1 and Figure 8). The 6th experimental recipe is prepared with agglomerating–furnace dust as main component (70%) and red mud (24%). This recipe is prepared without steel dust and anti-corrosive mud, but graphite (1%), lime (2%) and bentonite (3%) are used (Table 1 and Figure 8).

In our research, laboratory and pilot tests were carried out on the possibilities of recovery of pulverous ferrous waste in the form of pellets [7, 10, 13]. The influence of granulation and the weight of materials of different granulometric classes in the the pelletization recipe, the proportion of binder and the addition of water on the compression resistance of "green" and "hardened" pellet was studied [7-26]. The waste envisaged in the laboratory experiments was processed, according to the technological flow shown in Figure 2.

After homogenisation (mixing the solid components), the material (40 kg material/recipe) is inserted into the experimental equipment which is a disk pelletizer that belongs to the Laboratory of Materials Processing, in Faculty of Engineering Hunedoara [7, 9-26]. The pelletizing equipment (or "balling" device) used is a disc pelletizer, for production of balls (pellets), consist of an inclined, rotating disc mounted on a stationary structure. Nucleation, compaction, size enlargement, and spheroidization of the pellets ("balling" process) take place in the course of balling and related agglomeration processes [7, 9-26].

The duration of the pelletization process is range of 10-20 min. Throughout the process, the way of wetting the composition and the formation of the pellet ("balling") is followed. It was intended to determine the optimal additions of bentonite and water in order to obtain the most compression–resistant pellets. Depending on the humidity of the material, the water flow is adjusted. Balling has proved to be a highly versatile technique for large–scale production of particulate spheroids at relatively low capital and operational costs. Once pellets reach the desired size, they exit the pelletizing device [7, 9-26].

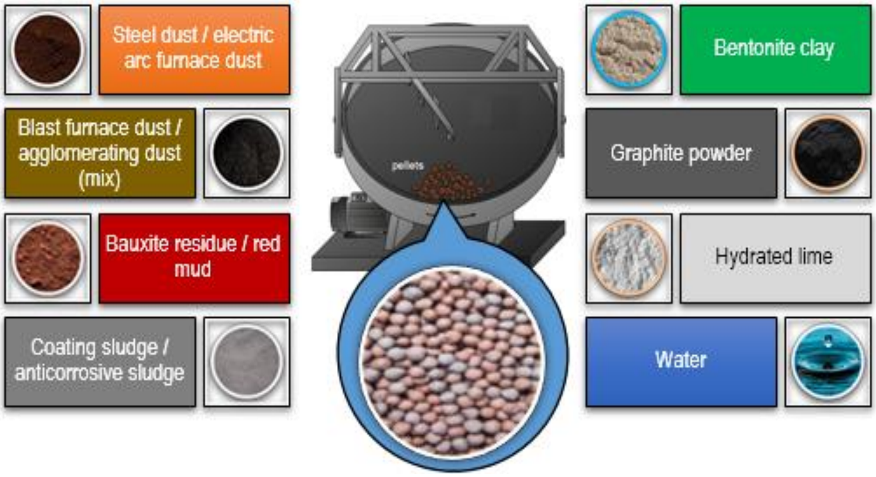

Steel dust / electric arc furnace dust
Blast furnace dust / agglomerating dust (mix)
Bauxite residue / red mud
Coating sludge / anticorrosive sludge
pellets
Bentonite clay
Graphite powder
Hydrated lime
Water

Figure 9
The pelletizing equipment and the "ingredients"

The waste used in the pelletisation process must have a fine granulation in order to form the proper pellet. For the obtained pellets, dusty wastes were used, their granulometric classes being presented in Table 2. Experimentally it was established that the size of the pellets was between 10 and 25 mm. The dimensions of 10-15 mm, a fraction representing at least 80% of the experimental recipes, were also considered to be optimal [7, 10, 13]. After completion of the pelletization operation, the obtained pellets are subjected to the granulometric sorting operation. The fraction less than 10 mm, is reintroduced into the circuit (in the pelletization stream) and the fraction greater than 10 mm will be subjected to hardening process [7, 9-26].

Table 2
Characteristics of pelletizing process and pellets

Experiment no.	Pelletizing time [min]	Humidity of green pellets [%]	Average diameter of green pellets [mm]	Granulometric classes proportion in the experimental recipes [%]					Compression resistance of pellets [daN/pellet]
				0–5 [mm]	5–10 [mm]	10–15 [mm]	15–20 [mm]	over 20 [mm]	
Sample #1	13	10.23	12.42	4	10.8	82.12	1.88	1.08	194
Sample #2	14	9.78	12.66	4.12	10.41	83.34	1.76	0.37	179.5
Sample #3	15	8.68	11.47	2.77	11.79	82.44	2.55	0.45	205
Sample #4	16	9.34	14.23	3.8	12.34	80.14	3.76	0.38	210.5
Sample #5	14	9.67	10.31	3.65	10.3	82.64	2.83	0.56	178
Sample #6	13	10.25	10.23	4.21	13.78	80.25	1.78	0.78	178.5

The hardening of the pellets are made by burning [7, 9-10, 13-26]. Burning hardening is done in electric (resistance) or flame ovens, following a proper treatment diagram (heating – maintenance – cooling), established on the basis of its own experiments (heating at 1150 ºC, for 2 hours, maintenance 30 minutes and cooling in the air).

After hardening, qualitative characteristics (chemical composition, dimensional analysis) and a mechanical characteristic (compression resistance of burned pellets) were determined in our laboratories. Also, the pellets are re-examined by the granulometric sorting operation, the fraction greater than 10 mm being dispatched to the beneficiary and the fraction less than 10 mm is recirculated (reintroduced into a new pelletisation circuit) [7, 10, 13-26].

5 The Experimental Results

We start to discuss by following the graphs from the Figures 10-12, which although conclusive, are not suggestive, because it presents a cumulative analysis of all factors that compete to achieve high quality pellets, i.e. the granulation of mixed components, humidity and the duration of pelletization:

- In making the recipes it was ensured that 80% of the solid components had a grain of 10-15 mm (Figure 10, according to Table 2)
- The duration of the pelletization operation, in the average of the 4 tests/recipe, was between 13-16 minutes (Figure 11, according to Table 2), which means a good time for such processes
- Diameters obtained for pellets, in a „green” state, are between 10-15 mm (Figure 11, according to Table 2)
- The necessary humidity for the plasticity of the „green” pellets was ensured at 8-11% proportions, relative to the volume of solid quantities (Figure 11, according to Table 2)
- At first view, the compression resistance that ensure sufficient quality of the pellets are obtained with formulations #1, #3 and #4, being over 200 daN/pellet (Figure 12, according to Table 2). The other recipes provide a lower limit of quality requirements (180-200 daN/pellet). It is considered in the assessments that in most cases the pellets are intended for use in steelmaking aggregates.

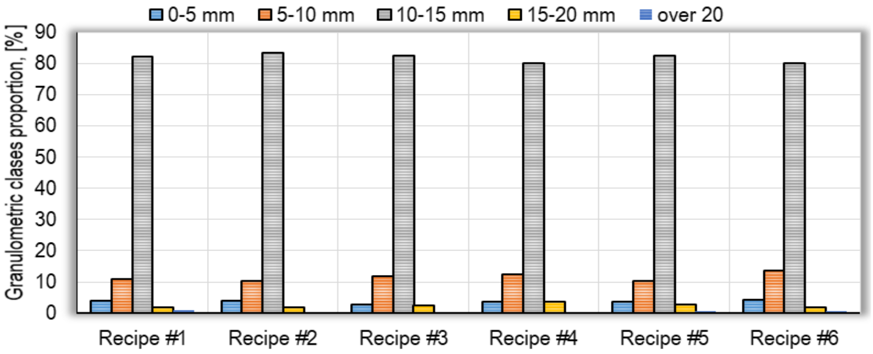


Figure 10

Granulometric classes proportion in the experimental recipes

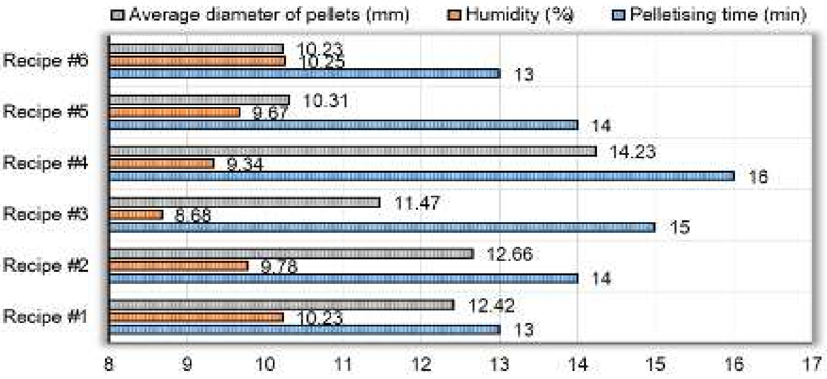


Figure 11

Average diameter of "green" pellets, humidity of "green" pellets and the pelletising time

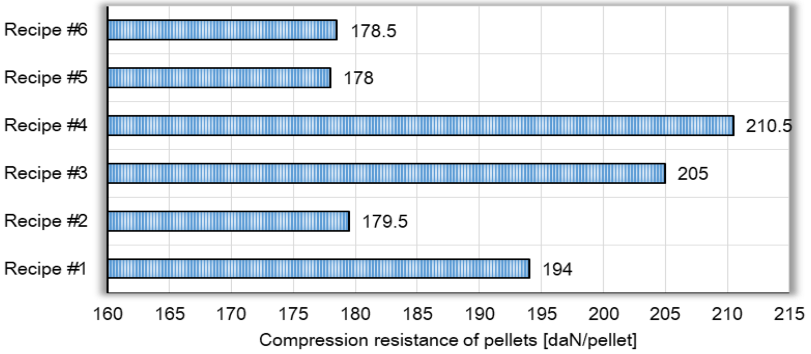


Figure 12

The compression resistance of the harnened pellets

6 Analysis and Discussion

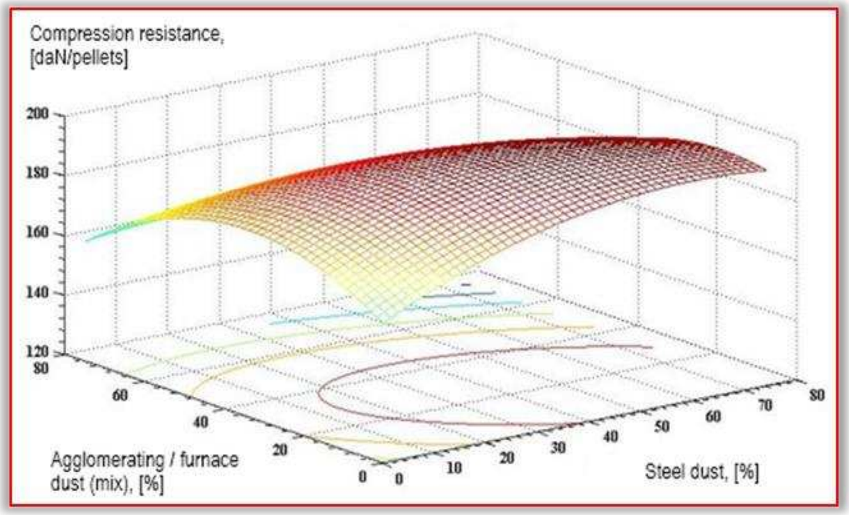
Our analysis of the study of the influence of solid components with iron content (steel dust, agglomeration–furnace dust, red mud/bauxide residue and anti-corrosive sludge) on compression resistance, are graphically represented in Figures 13-17, based on the results obtained in our experiments.

In diagram presented in Figure 13, [Rc = f(steel dust, agglomeration–furnace dust)], it is noted that there is an area with a maximum point, and the range with values greater than 180 daN/pellet is quite extensive, so that the variation limits for the two components are quite high (approximately 60 agglomeration–furnace dust, over 70% steel dust). It is noted that only at low values of the two components (approximately 60% agglomeration–furnace dust, or over 70% steel dust), the resistance to compression is less than 180 daN/pellet. However, at values of only 40% agglomeration–furnace dust and between 30-60% steel dust, compression resistance reaches values of almost 200 daN/pellet.

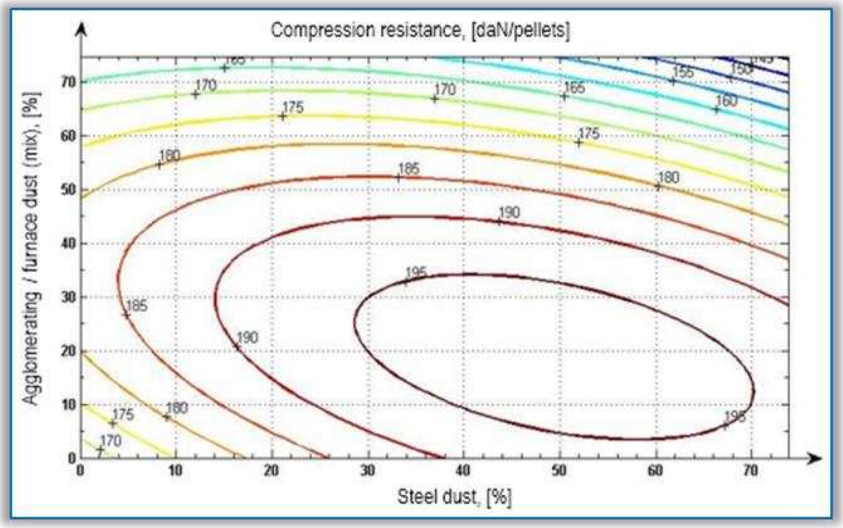
In the diagram presented in Figure 14, [Rc = f(steel dust, bauxide residue)], it is found that there is an increase in compression resistance with an increase in the proportion of red mud/bauxide residue and steel dust, the best values being obtained for the upper limits of the two components of the recipe. The technological explanation lies in the high finesse of both materials and the content of Al_2O_3 in bauxide residue. Depending on the availability of waste assortments it would be preferable to place their proportion at quantities that provide compression resistance above 200 daN/pellet – in the right corner, top of the diagram (10-14% bauxide residue and 30-70% steel dust) or more than 180-190 daN/pellet (10-14% bauxide residue and between 10-70% steel dust).

The diagram presented in Figure 15 [Rc = f(agglomeration–furnace dust, bauxide residue] shows the positive influence of bauxide residue, practically an addition of this between 4-16% ensures good values for the compression resistance of the pellets, even if the content of agglomeration–furnace dust does not exceed 50%. Good values of compression resistance are also obtained for the level curves corresponding to some participations in recipes with 10-50% agglomeration powder–furnaces and smaller amounts of bauxide residue (2-6%).

In the case of use of anti–corrosive sludge together with agglomeration–furnace dust – Figure 16 [Rc = f(agglomeration–furnace dust, anti–corrosive sludge)] –, it is found that higher values for compression resistance are obtained at their lower proportions in the pelletization charge. Very good quality pellets can be obtained with agglomeration–furnace dust within a maximum of 50% and for 10-15% anti-corrosive sludge, or even lower (5-10%).



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

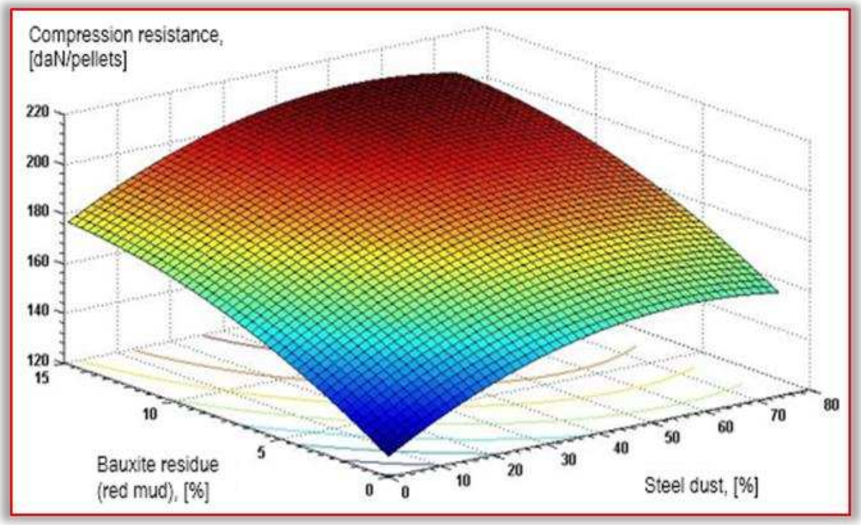
Figure 13

Compression resistance of hardened pellet's with the proportions of agglomeration–furnace dust and steel dust used in the experimental recipes – [Rc = f(agglomeration–furnace dust, steel dust)]

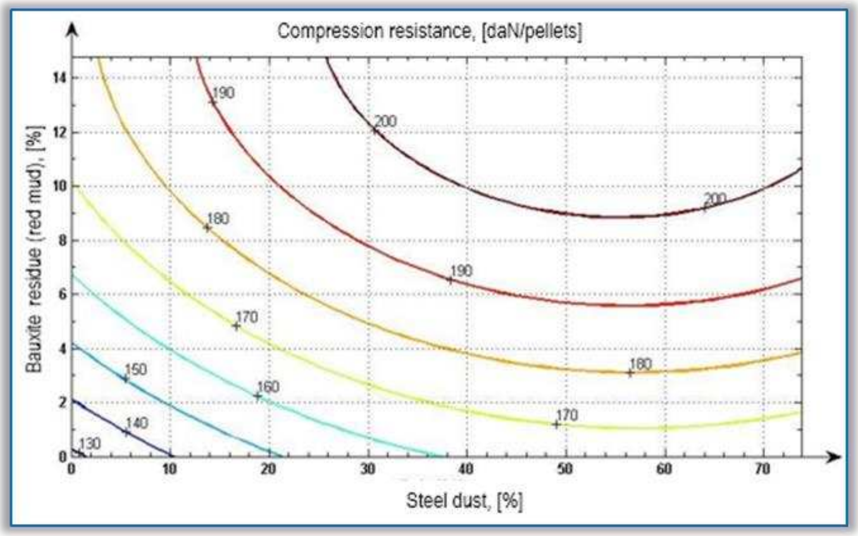
Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.0074$; $a_{(2)} = -0.0137$; $a_{(3)} = -0.0085$; $a_{(4)} = 0.8945$; $a_{(5)} = 0.9330$; $a_{(6)} = 166.7903$

Coefficient of multiple correlation: $R^2 = 0.7482$ (relative high grade of correlation)



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

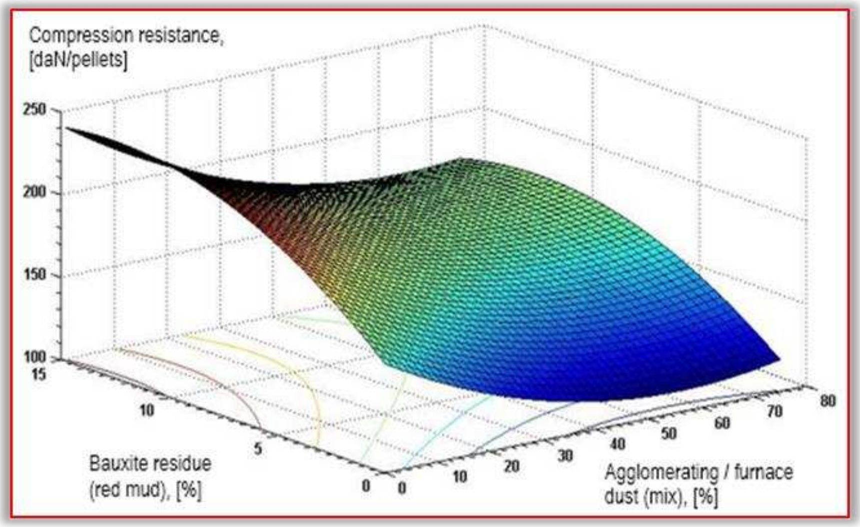
Figure 14

Compression resistance of hardened pellet's with the proportions of steel dust and bauxite residue / red mud used in the experimental recipes – [Rc = f(steel dust, bauxite residue / red mud)]

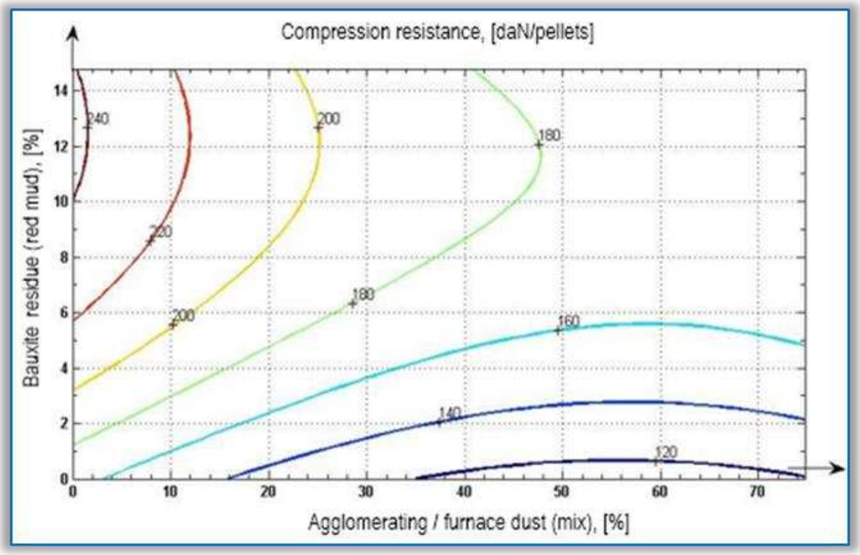
Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.0109$; $a_{(2)} = -0.1750$; $a_{(3)} = -0.0055$; $a_{(4)} = 1.2592$; $a_{(5)} = 5.8883$; $a_{(6)} = 128.1288$

Coefficient of multiple correlation: $R^2 = 0.9389$ (high grade of correlation)



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

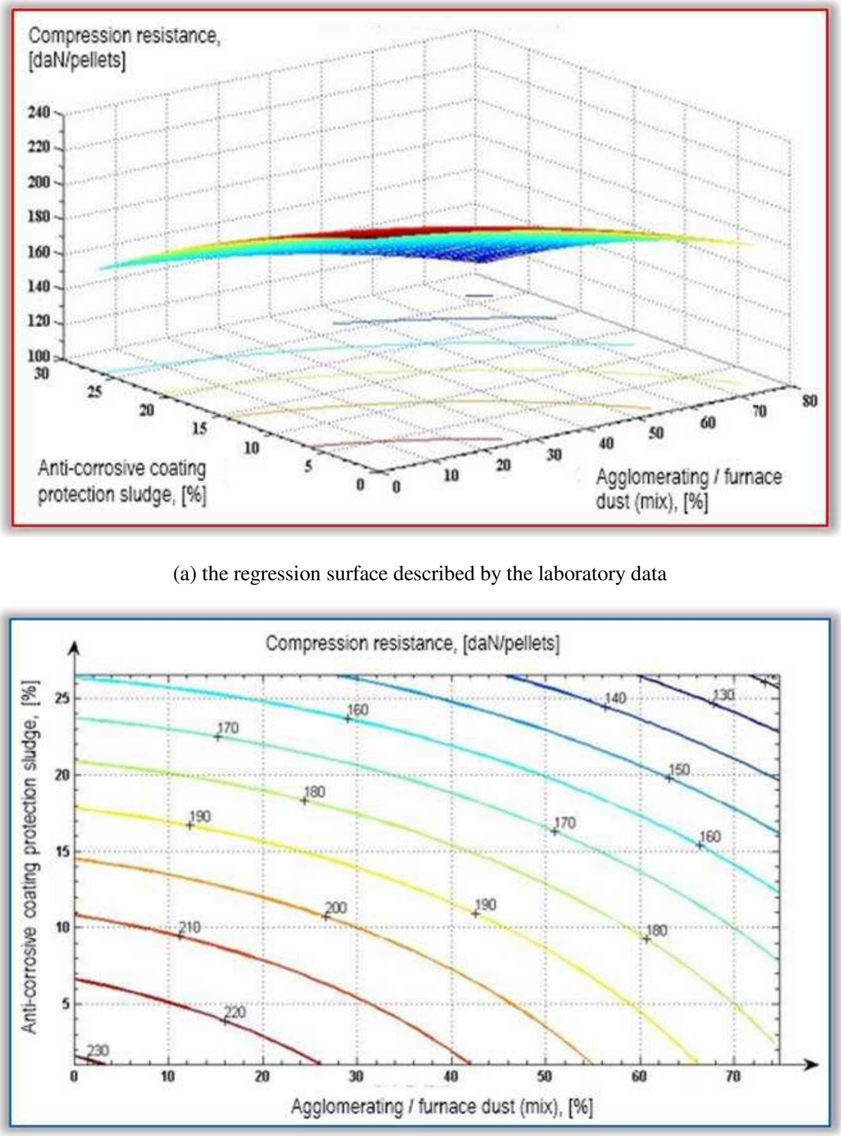
Figure 15

Compression resistance of hardened pellet's with the proportions of agglomeration–furnace dust and bauxite residue used in recipes – [Rc = f(agglomeration–furnace dust, bauxite residue)]

Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = 0.0173$; $a_{(2)} = -0.4860$; $a_{(3)} = -0.0193$; $a_{(4)} = -1.9154$; $a_{(5)} = 12.2658$; $a_{(6)} = 165.8485$

Coefficient of multiple correlation: $R^2 = 0.8652$ (high grade of correlation)



(a) the regression surface described by the laboratory data

(b) the level curves of compression resistance, in 2D coordinates

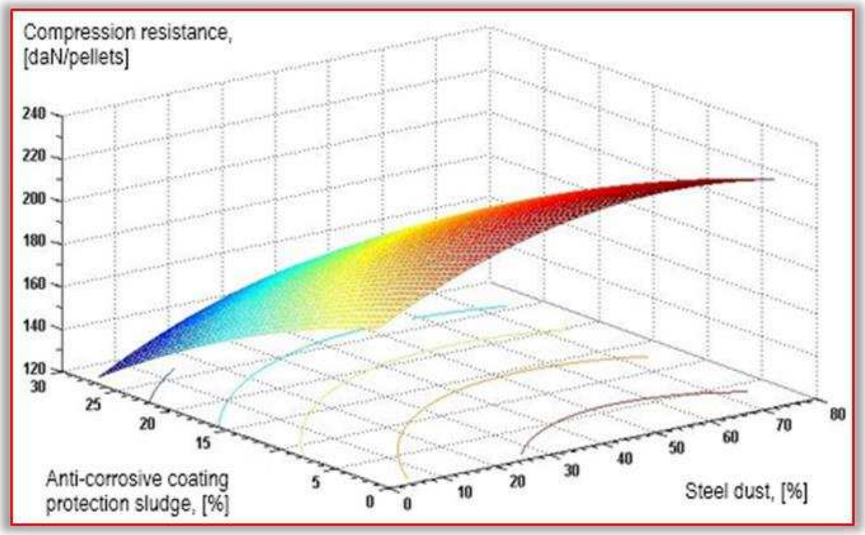
Figure 16

Compression resistance of hardened pellets with the proportions of agglomeration–furnace dust and anti–corrosive sludge used in recipes – [Rc=f(agglomeration–furnace dust, anti–corrosive sludge)]

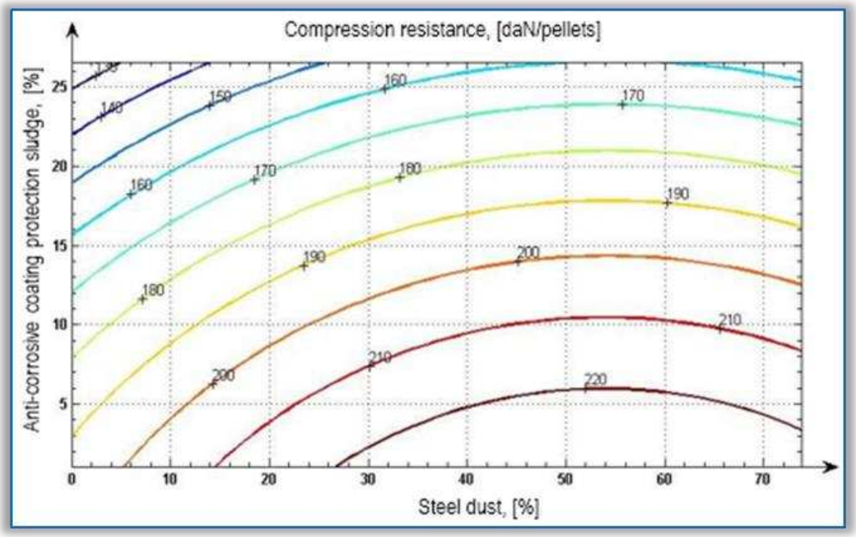
Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.0050$; $a_{(2)} = -0.0430$; $a_{(3)} = 0.0037$; $a_{(4)} = -0.2944$; $a_{(5)} = -1.6302$; $a_{(6)} = 232.8057$

Coefficient of multiple correlation: $R^2 = 0.8742$ (high grade of correlation)



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

Figure 17

Compression resistance of hardened pellet's with the proportions of steel dust and anti–corrosive sludge used in the recipes – [Rc=f(steel dust, anti–corrosive sludge)]

Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.0124$; $a_{(2)} = -0.0420$; $a_{(3)} = 0.0007$; $a_{(4)} = 1.3304$; $a_{(5)} = -1.5737$; $a_{(6)} = 194.9876$

Coefficient of multiple correlation: $R^2 = 0.8967$ (high grade of correlation)

The increase in the proportion of steel dust to 30-70% of the pelletizing charge has a positive influence on the resistance to compression of the pellets, but the anticorrosive sludge have a negative influence at increases in quantities, preferably a use of only 5-10%, as is presented in the Figure 17 [Rc = f(steel dust, anti–corrosive sludge)].

Our analysis on the compression resistance of the pellets provided by the additions (bentonite, lime and graphite) are graphically represented in Figures 18-20, based on the results obtained in our experiments. Also, we shows the lime dependence with the two main components (steel dust and agglomeration–furnace dust), in Figure 21 [Rc = f(steel dust, lime)] and Figure 22 [Rc = f(agglomeration–furnace dust, lime)]. With regard to these cases of analysis, the following technological remarks are required:

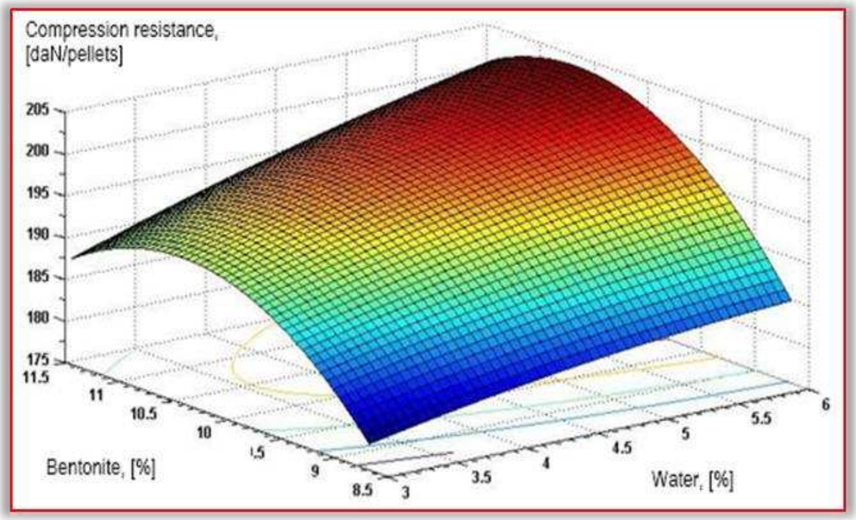
- Increasing the proportion of water to 4-6% and bentonite to about 9% leads to an increase in compression resistance, as appropriate conditions have been provided to obtain “green” pellets with high strength, from powdery materials, which is also found in the burned pellets (Figure 18 [Rc=f(water, bentonite)]).
- The proportions of the water and lime additions of are well correlated in most cases, with the exception of the addition of lime below 1% and water below 4%, respectively the addition of lime above 5-6%, where the compression resistance is below the required limit. The field of work is very extensive, virtually unlimited, and from this point of view there are no particular problems (Figure 19 [Rc = f(water, lime)]).

The proportions of lime and bentonite are well chosen, so that they have close influence and can substitute each other, ensuring for compression resistance values >180 daN/pellet (Figure 20 [Rc = f(lime, bentonite)]).

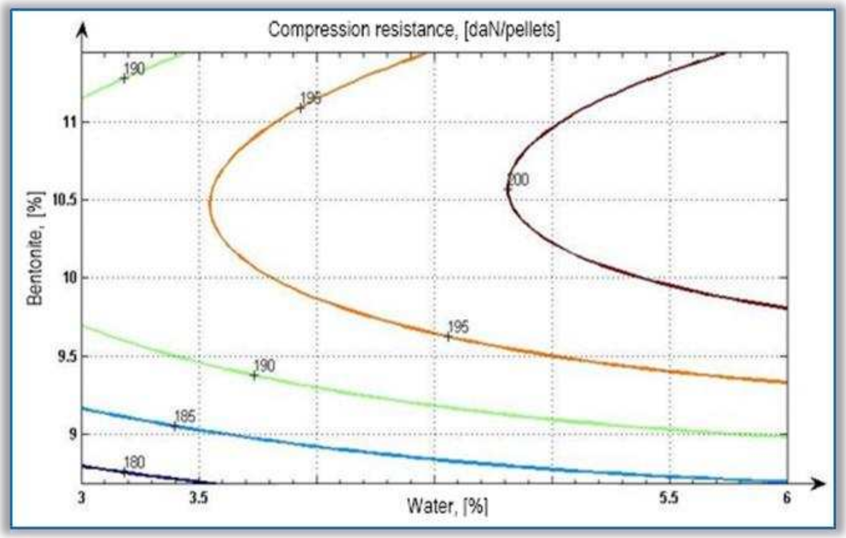
The addition of graphite is insignificant, and the quantity used in recipes, does not influence the process of pelletization nor the quality of the burned pellet. For this reason, graphic dependencies are not shown among these results.

Increased resistance to compression of the pellets, slightly decreases with the increase in the proportion of agglomeration–furnace dust and instead increases with the addition of lime, according to the Figure 21 [Rc=f(agglomeration–furnace dust, lime)]. The agglomeration–furnace dust also has a positive effect, in that, in the combustion process, it ensures the reduction of iron oxides, increasing the degree of metalization of the pellets.

Finally, the chart presented in Figure 22 [Rc=f(steel dust, lime)] shows that, in terms of the correlation of steel dust and the addition of lime, practically regardless of their proportion and the limits of the addition of lime, up to 6%, consistent pellets are obtained for use in steelmaking aggregates.



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

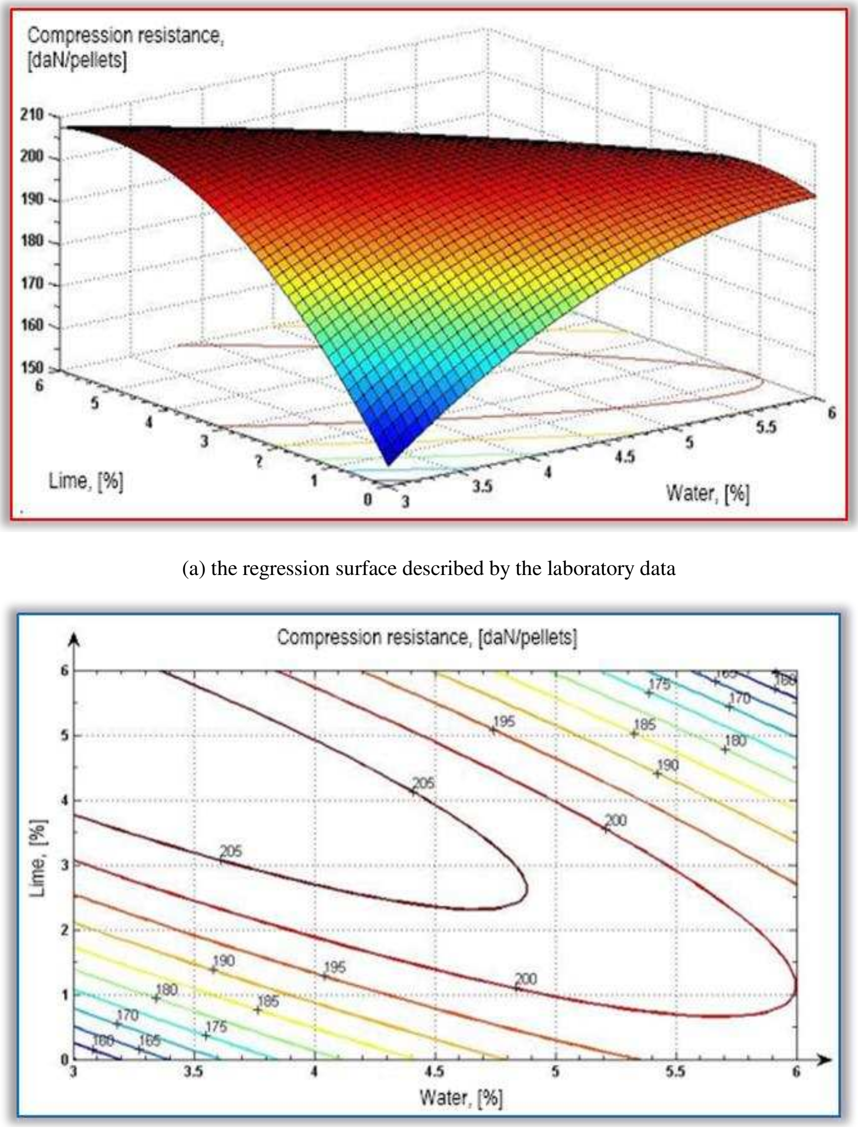
Figure 18

Compression resistance of hardened pellet's with the proportions of bentonite and water used in the experimental recipes – [Rc = f(bentonite, water)]

Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.4072$; $a_{(2)} = -1.1150$; $a_{(3)} = -4.7305$; $a_{(4)} = 0.8048$; $a_{(5)} = 96.2018$; $a_{(6)} = -314.4748$

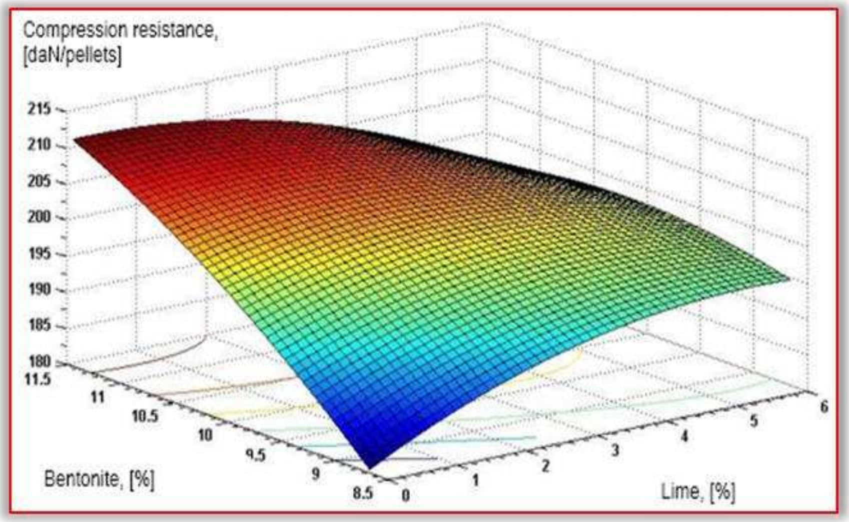
Coefficient of multiple correlation: $R^2 = 0.7922$ (high grade of correlation)



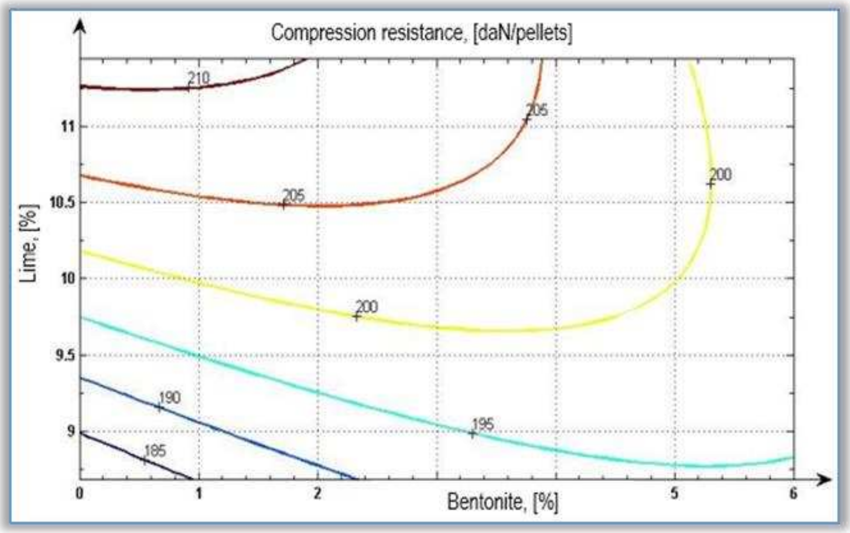
(a) the regression surface described by the laboratory data

(b) the level curves of compression resistance, in 2D coordinates

Figure 19
Compression resistance of hardened pellet's with the proportions of lime and water used in the experimental recipes – [Rc = f(lime, water)]
Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:
$a_{(1)} = -4.4814$; $a_{(2)} = -2.0376$; $a_{(3)} = -5.4652$; $a_{(4)} = 54.5467$; $a_{(5)} = 37.3916$; $a_{(6)} = 31.4430$
Coefficient of multiple correlation: $R^2 = 0.6056$ (relative high grade of correlation)



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

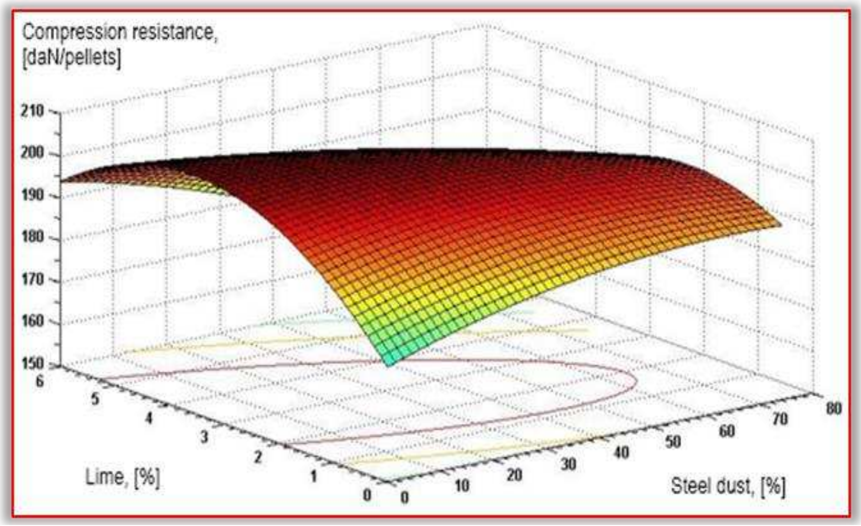
Figure 20

Compression resistance of hardened pellet's with the proportions of lime and bentonite used in the experimental recipes – [Rc = f(lime, bentonite)]

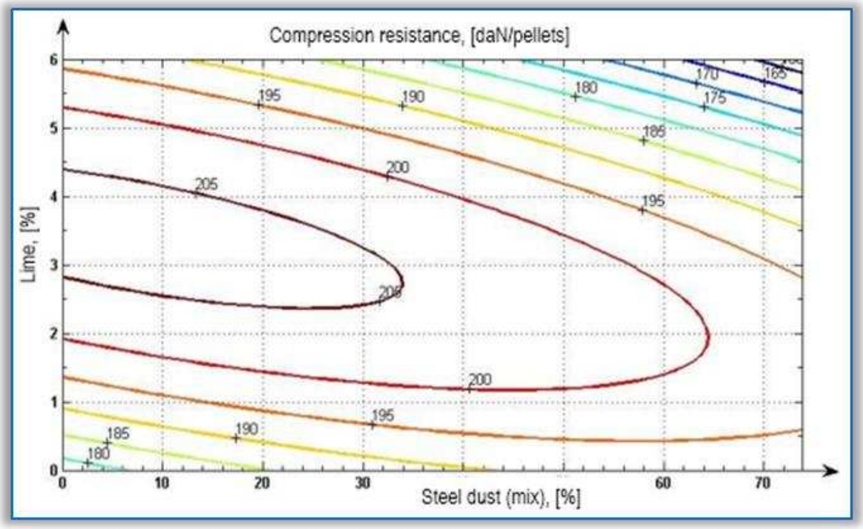
Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.4720$; $a_{(2)} = -1.4200$; $a_{(3)} = -1.7941$; $a_{(4)} = 20.7279$; $a_{(5)} = 39.7588$; $a_{(6)} = -57.6807$

Coefficient of multiple correlation: $R^2 = 0.8698$ (high grade of correlation)



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

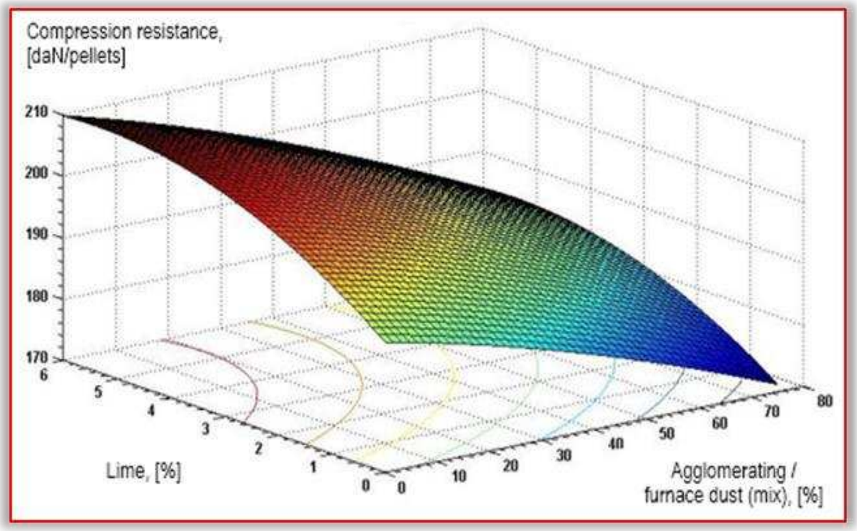
Figure 21

Compression resistance of hardened pellet's with the proportions of steel dust and lime used in the experimental recipes – [Rc = f(steel dust, lime)]

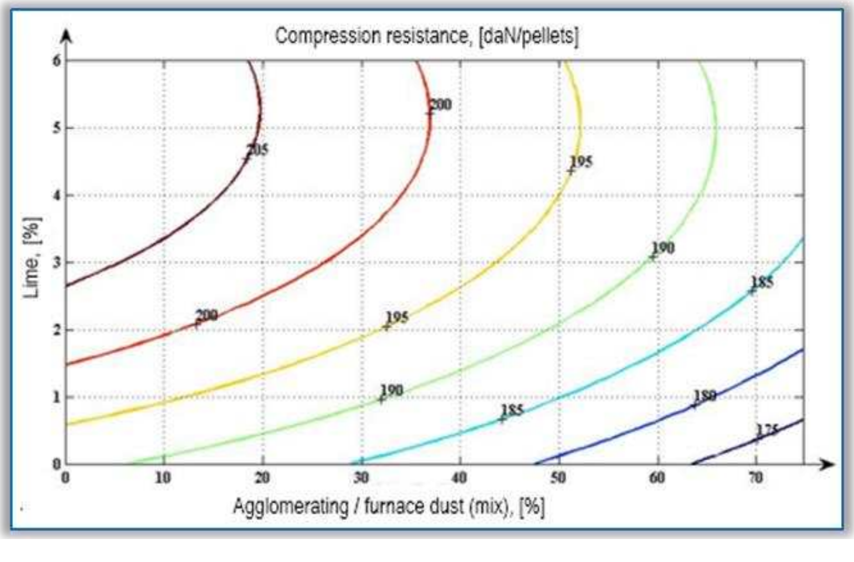
Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.0012$; $a_{(2)} = -0.6413$; $a_{(3)} = -0.0085$; $a_{(4)} = -0.1783$; $a_{(5)} = 6.9270$; $a_{(6)} = 191.1665$

Coefficient of multiple correlation: $R^2 = 0.7759$ (high grade of correlation)



(a) the regression surface described by the laboratory data



(b) the level curves of compression resistance, in 2D coordinates

Figure 22

Compression resistance of hardened pellet's with the proportions of agglomeration–furnace dust and lime used in the experimental recipes – [Rc = f(agglomeration–furnace dust, lime)]

Equation type: $z = a_{(1)}x^2 + a_{(2)}y^2 + a_{(3)}xy + a_{(4)}x + a_{(5)}y + a_{(6)}$, in which the coefficients are:

$a_{(1)} = -0.0034$; $a_{(2)} = -2.2408$; $a_{(3)} = -0.1161$; $a_{(4)} = 0.4437$; $a_{(5)} = 16.1699$; $a_{(6)} = 177.2020$

Coefficient of multiple correlation: $R^2 = 0.8171$ (high grade of correlation)

Conclusions

In this paper the research and relevant results are presented, in regard to the formation of pellets, using waste from ferrous industry (steel dust, agglomerating–furnace dust) and mining and mineral processing sectors (red mud/bauxide residue, anti-corrosive sludge). For this, we considered the existing ferrous powdery waste in the Hunedoara County (Hunedoara and Calan) area and several powdery wastes from Bihor County (Oradea) area. In addition, within the pellets recipes, graphite is used, as the reducing agent and respectively, bentonite and lime are used as binders.

Our research work has considered the following technological problems:

- The wetting capacity of processed materials, in the pelletizor scale
- The quality of “green” pellets according to the processed materials, their chemical composition and granulation, the addition of water and binders
- The quality of the hardened pellets according to the processed solid materials, the chemical composition and their obtained dimensions, the addition of water and the binders quantity
- The compression resistance of the burned pellets according to the solid materials processed, the addition of water and binders

On the basis of the above, we studied the possibilities of recovery of powdery ferrous waste in the steel industry. Given that at the national level, this activity, unfortunately, has a very limited scope. Our strategies should focus on development of recycling/valorisation capacities of iron containing raw materials contained in waste, as well as on installations and technologies using waste in the production process. In this sense, analyzing the results obtained in our laboratory, it follows that various waste processed in the form of pellets, according to the experimental recipes #1-6, can be further utilized in industrial applications.

Acknowledgment

The study was carried out on the basis of extensive research over the last 20 years, undertaken within the Hunedoara Faculty of Engineering, following the results of numerous researches [7, 9, 10, 13-26] that concerned a large number of researchers in this institution. This research presented now is part of the research on the doctoral thesis [13], as well as a number of researches undertaken after this period in order to recover by pelletization of materials containing valuable or deficitary elements from different kinds of powdery waste stored in areas with risk factor.

References

[1] Romanian Government, *Romania’s Sustainable Development Strategy 2030 – Implementing the 2030 Agenda for Sustainable Development*, Department of Sustainable Development (2018) http://dezvoltaredurabila.gov.ro/

[2] European Commission, *The Environmental Implementation Review 2019 – Country Report: Romania*, Directorate–General for Environment (2019)

[3] Blengini, G. A., Mathieux, F., Mancini, L., Nyberg, M., Viegas, H. M., *Recovery of critical and other raw materials from mining waste and landfills – State of play on existing practices*, Luxembourg: Publications Office of the European Union (2019)

[4] Ministry of Environment, Waters and Forests, Romania, *National Strategy for the Management of Contaminated Sites* (2015) http://www.mmediu.ro

[5] National Environmental Protection Agency, *Present status for Soil Rehabilitation in Romania – Priorities, Workshop on Contaminated site caracterisation*, 7th Framework Programme Theme 6 "Environment", Soil and Subsoil Protection Office (2013)

[6] Romanian Government, National Strategy and National Plan for the management of contaminated sites in Romania. *Chapter III: Current situation of potentially contaminated/contaminated sites in Romania* (2013)

[7] T. Heput, E. Ardelean, N. Constantin, A. Socalici, M. Ardelean, R. Buzduga, *Recovery of small and powder ferrous waste*, Politehnica Publishing House, Timisoara, (2011) pp. 7-11

[8] H. W. Campbell, Sludge management–future issues and trends, *Water science and technology*, 41(8), (2000) pp. 1-8

[9] I. Butnariu, N. Constantin, C. Dobrescu, T. Heput, Research on the recycling of pulverulent waste from the ferrous and non-ferrous industry in order tu reduced the pollution. *Revista de Chimie*, 69(5) (2018) pp. 1066-1070

[10] V. A Socalici, *Contribution on improvement the quality of steel*, University Politehnica Timisoara (2016) habilitation thesis

[11] F. Su, H. O. Lampinen, R. Robinson, Recycling of sludge and dust to the BOF converter by cold bonded pelletizing, *ISIJ international*, 44(4) (2004) pp. 770-776

[12] S. Kumar, R. Kumar, A. Bandopadhyay, Innovative methodologies for the utilisation of wastes from metallurgical and allied industries, *Resources, Conservation and Recycling*, 48(4), (2006) pp. 301-314

[13] S. Serban, *Research on waste recovery containing iron and steel alloying elements*, University Politehnica Timisoara, (2015) doctoral thesis

[14] T. Heput, A. Socalici, E. Ardelean, M. Ardelean, Environment ecological process in Hunedoara area through reinsertion in economic circuit of scrap and pulverous waste. *Annals of the Faculty of Engineering Hunedoara – Journal of Engineering*, VII(3) (2009) pp. 293-298

[15] T. Heput, I. Kiss, V. Puţan, Researches regarding the implementation in the industrial practices of the accounting technologies of the ferrous pulverous wastes, stored in the regional ponds, *Annals of the Faculty of Engineering Hunedoara*, 1(2) (2003) pp. 71-75

[16] S. Serban, T. Heput, I. Kiss, Recycling experiments on pulverous wastes resulted from ferrous industry, mining and energetic sectors, *Acta Technica Corviniensis – Bulletin of Engineering*, X, 2, (2017) pp. 139-146

[17] S. Serban, T. Heput, Ferrous wastes recovery possibilities in the area of steel industry – experiments in the laboratory phase on the briquettes production from fine and pulverous wastes, *Acta Technica Corviniensis – Bulletin of Engineering*, VIII, 3, (2015), pp. 49-56

[18] S. Serban, T. Heput, I. Kiss, Recovery possibilities through pelletizing of the pulverous wastes stored in the regional ponds, *VIth International Conference Industrial Engineering & Environmental Protection (IIZS 2016)*, Zrenjanin, Serbia, 2016, pp. 116-121

[19] S. Serban, Reintroduction of iron–containing waste and steel alloying elements into the economic circuit, *Acta Technica Corviniensis – Bulletin of Engineering*, XI, 4, (2018), pp. 127-134

[20] M. L. Strugariu, T. Heput, A. Socalici, Recovery of sludge resulting from corrosion protection operations, *Metalurgia International*, 18(8) (2013), pp. 161-166

[21] A. V. Socalici, E. Ardelean, M. L. Strugariu, Research on sustainable use of powdery waste, *Environmental Engineering and Management Journal (EEMJ)*, 15(1) (2016) pp. 207-212

[22] D. A. Popescu, Study on the quality of industrial waste deposited in ponds, *Annals of Faculty of Engineering Hunedoara – International Journal of Engineering*, XII, 4 (2014), pp. 315-321

[23] D. A. Popescu, *Research into the recovery of pulverous and small waste from the metallurgical industry*, University Politehnica Timisoara (2018) doctoral thesis

[24] A. S. Todoruţ, *Management research and recovery of small and powdery wastes, resulted from materials industry, for sustainable development of Hunedoara*, University Politehnica Timisoara, (2013) doctoral thesis

[25] E. M. Crişan, *Research into the recovery of iron and carbon–containing pulverous and small waste in steel*, University Politehnica Timisoara, (2018), doctoral thesis

[26] E. M. Crişan, T. Heput, Research on the influence of basic additives on the compressive strength of pellets. *Annals of Faculty of Engineering Hunedoara – International Journal of Engineering*, 9(3) (2011) pp. 449-454

A Genetic Algorithm for the Minimum Vertex Cover Problem with Interval-Valued Fitness

Benedek Nagy[1], **Péter Szokol**[2]

[1] Department of Mathematics, Faculty of Arts and Sciences, Eastern Mediterranean University, Famagusta, North Cyprus, via Mersin-10, Turkey, benedek.nagy@emu.edu.tr

[2] EPAM Systems, Bókay János u. 44, 1083 Budapest, Hungary, peter_karoly_szokol@epam.com

Abstract: This paper presents a new genetic algorithm for the minimum vertex cover problem. It uses interval-valued fitness and greedy error correction to obtain phenotypes (candidate solutions). By the interval-valued fitness the fitness of the candidate solution is measured not only for the whole graph, but for some of its disjoint subgraphs. A new candidate solution is obtained from those subgraphs that have the best performance among the subgraphs of the candidates with the same set of vertices. The interval-valued fitness accelerates the search effectively for graphs with a great deal of nodes and relatively small numbers of edges. In the presented algorithm we prefer to distinguish genotypes and phenotypes and do not use Lamarckian inheritance. Phenotypes are easily generated by greedy error correction from the genotypes and, in this way, a larger variety of genomes can be used during the process.

Keywords: genetic algorithm; minimum vertex cover; interval-valued fitness; phenotypes; memetic algorithm

1 Introduction

A vertex cover of a graph is a set of vertices, such that each edge of the graph is incident to at least one vertex of the set. Minimum vertex cover means that the size of the vertex cover set is minimal. Finding the minimum vertex cover is a classical NP-complete problem. As such, genetic algorithm (GA) seems an ideal method for the minimum vertex cover problem to obtain reasonable solution in reasonable time. Genetic algorithms are based on the idea of Charles Darwin's evolution. GA is a heuristic search algorithm, it finds a relatively good solution in a short amount of time, and it can be stopped any time, it will always have a, relatively, good solution [1, 2, 3].

Every GA has a population of candidate solutions. For the sake of simplicity, from now on, we call them solutions, instead of candidate solutions. Each of these solutions have a genome, the entirety of their hereditary information. The algorithm must calculate the fitness of the solutions and the next generation can inherit the genomes of the best solutions. A child can inherit genes from one or more parents. There is also a small chance for mutation. The algorithm measures the fitness of each solution of the new generation again, and the cycle continues until a given condition is met.

In the literature there are various approaches to use GA and related methods for obtaining reasonable solutions for the minimum vertex cover problem. Heuristic approximation is used based on the share degree distribution of the vertices [4]. The performances of random local search algorithm and the basic (1+1) evolutionary algorithm are compared in various subclasses of graphs in [5]. Various heuristic algorithms for minimum vertex cover are compared in [6], including hierarchical Bayesian optimization algorithm, branch-and-bound problem solver, simple genetic algorithm and the parallel simulated annealing to show that evolutionary, and so, genetic algorithms are reasonable choices to solve the problem. In [7], the initial population of the GA is created in mathematical manner (uniformly distributed initial population) instead of a random population; moreover, the minimum vertex-cover problem is converted into constrained combinatorial optimization problem. Local optimization technique, in fact, hill climbing was used as Lamarckian evolution in [8] to obtain a hybrid GA. The genetic algorithms in these papers used mutation to avoid falling into a local optimum of the problem. The ideas mentioned in them inspired also us to make experiments with our algorithm and improve it. Our novel idea is to use interval-valued fitness, however, we have combined it with various other well-known GA methods, such as, e.g., with elitism and local (error) correction.

GAs give a lot of freedom to the programmer. Creating the structure of the genome is the first step. The most obvious way to do this for the minimum vertex cover problem, is to create a gene for every node. The gene can have two values: true or false. True means that the corresponding node is in the vertex cover set, false means that the node is not in the vertex cover set. We distinguish genotypes and phenotypes [9]. It results a better variety of the gene pool. It is possible that a vertex cover candidate set is not covering the graph. In this case, the genotype would not be a vertex cover, but the algorithm can add nodes to the candidate vertex cover, and thus, the phenotype becomes one. With the fast, and effective greedy algorithm, the phenotype improves significantly. Creating the phenotype can be viewed as a part of the fitness function. The fitness function assigns fitness values to the genomes. Fitness value indicates how good the solution is, i.e., how close it is to an ideal or optimal solution. Inheritance is another main part of GAs, and there are lots of opportunities here as well. We present a new idea, the interval-valued fitness method. We apply to the GAs the concept of intervals. It means that the child inherits gene sequences from one or more parents. Interval-values are finite unions of components over the unit interval. They are used to represent various fuzzy and many-valued logics [10, 11] and they are also used to introduce a new computing paradigm [12, 13] that is applied to solve various computationally hard problems [14, 15, 16] and classes of problems [14, 17], theoretically, based on its inner parallelism. Here we use this concept at the fitness function of the GA. If a parent phenotype has a good fitness value on a component of the interval, then the genes of the parent on this component will be added to the child's genome. This process is very useful if it is easy to distinguish various characteristics of a solution, and these characteristics are independent, or weakly dependent on each other. For the vertex cover problem, to find and distinguish various characteristics of a graph is not very straightforward, but we present some valuable ideas. It is expected that for graphs where there are partitions such that most of the edges are inside partitions the algorithm could efficiently use its interval-valued fitness to compose reasonable solutions on the partitions to a general solution for the whole graph.

The rest of the paper is organized as follows: First we introduce the problem, its mathematical representation, and the structure of a genome in Section 2. In Section 3, we describe four ideas for determining the partitions of a graph. We present a GA for determining the partitions of a graph in Section 4 and in Section 5, we explain the greedy error correction, a method to repair invalid vertex covers, obtaining the phenotypes from genotypes. Next, we describe our genetic algorithm for the minimum vertex cover in Section 6. Section 7 provides an example, Section 8 shows the Experimental Results and finally, Section 9 discusses the Conclusions.

2 Minimum Vertex Cover, the Genome of the GA and the Model of the Problem

In this section the minimum vertex cover problem and the structure of the genome for our GA is presented. Let a simple graph *G* be given with *n* vertices consisting of a set of vertices *V*. (The order of the vertices in *V* is arbitrarily fixed.) The aim is to find the minimum vertex cover *C*, a set of vertices, where *C* is a subset of *V* such that each edge of *G* is incident to at least one vertex in *C* (that is *C* is a so-called vertex cover of *G*) and the size of *C* is minimal (among all vertex covers of *G*). *G* can be connected or disconnected, and *G* may have isolated vertices. An isolated vertex is not an endpoint of any edge, therefore the isolated vertex is not in any minimal vertex cover *C*. For an example of a simple graph and its minimum vertex cover see Figure 1. This graph *G* has 6 vertices, two of those are indicated by circle, they form a vertex cover *C* for this graph. It is also easy to see that there is no sole vertex in *G* which forms a vertex cover alone. Thus, *C* is, in fact, a minimum vertex cover. The minimum vertex cover problem is a classical optimization problem, and it is one of the 21 NP-complete problems listed by Karp [18]. Since this is one of the most known NP-complete problems, it plays an important role in computational complexity theory. On the other hand, since generally it is intractable, it is a good target for evolutionary computing and GA to find approximate solutions.

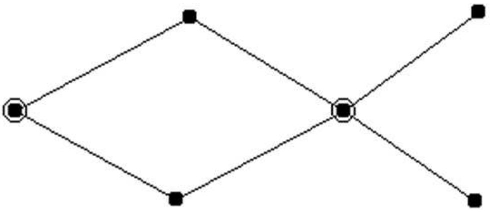


Figure 1
A graph, with six vertices and its minimum vertex cover

Now, let us turn to our approach. In the rest of this section, some details are shown how this problem can be modeled and represented by genes such that evolutionary steps can reasonably solve it. Also some essential details of our approach are described. In our approach the vertex covers are computed by greedy error corrections from the genome in deterministic manner (see Section 5), in this way, the vertex cover is, in fact, the phenotype corresponding to the given genome. The population of the GA consists of a set of genomes $\langle F, I, K \rangle$, where:

- F is the fitness value of the genome is in fact computed for the phenotype, the vertex cover C implied by the genome. It is calculated by the fitness function, and it represents the overall fitness of the solution. F represents the fitness of the phenotype, meaning that first we use the greedy error correction on the solution of a genome, then we compute the fitness function on this modified solution. The greedy error correction guarantees that the phenotype is a vertex cover. $F = |C|$ (the size of the vertex cover).
  Note that F is not a fitness function in the strict sense, since larger value of F means weaker solutions; in this sense F is more related to error-measures.

- $I = (I_1,I_2,...,I_m)$ is an ordered list (set) of interval fitness values of the phenotype solution C. Let $P = (P_1,P_2,...,P_m)$ denote the partitions of V. P is the same for every genome, and every generation. If $P_j = \{a_1,a_2,...,a_k\}$, then I_j is the component fitness of the solution on the subgraph of G containing only vertices of P_j and all the edges of G which contain only such vertices. The component fitness on a subgraph means the number of vertices E of the subgraph in the vertex cover phenotype, where E is also a member of C, e.g., $I_j = | P_j \cap C |$.

- $K = (K_1,K_2,...,K_n)$ is a binary vector of the vertices in the genome. If $K_i = 1$ (i.e., true), then vertex V_i is in the genome for the cover set C proposed by the phenotype of this genome.

For an example, see Figure 2. As it is shown, the partitions of the graph are also described by our data: there are three partitions in this example. The set of circled vertices is a vertex cover C, and its size is 8. The interval fitness is assigned to the partitions, and actually, the sum 2 + 3 + 3 = 8 gives the fitness of the vertex cover. Since the genotype, in this example, for simplicity and luckily, is the same as the phenotype, we do not need to apply the error correction, thus, in this case, K and C are identical. The binary genome vector K has 1's in exactly those 8 position.

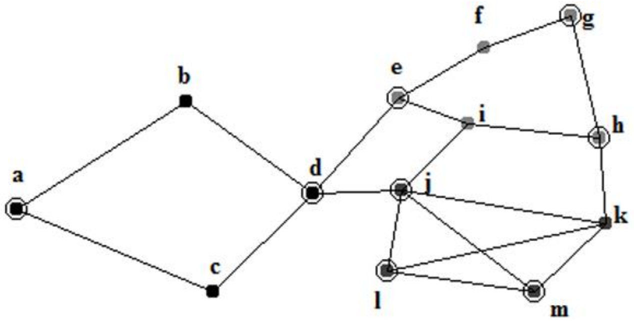


Figure 2
A graph where P = ({a,b,c,d}, {e,f,g,h,i}, {j,k,l,m}), and one of the possible solutions, where F = 8 (= $|C|$), I = (2,3,3), and K = (1,0,0,1,1,0,1,1,0,1,0,1,1)

3 Initialization: Determining the Partitions

We use fitness values on interval components and it is a crucial point to divide the graph to partitions. The interval fitness function works more efficiently if every P_j contains closely related vertices and the partitions are not connected so strongly. There are several methods to determine a feasible P.

- Random partitions
- Divide by the time of creation of the vertices
- Manually determine the partition of each vertex
- Spectral partitioning [19]
- GA, with a fitness function that calculates fitness by the dissimilarity of the neighbors of each vertex

The partitioning does not have to be optimal. But it has to be fast. It can help even if the partitioning is random, since as it will be detailed later, for every generation, only one genome will be created with the interval fitness method. For example, if the population size is constantly 200, then 199 children will evolve normally, only 1 child (super-child) will be created from the best intervals. Even if the partitioning is random, there is a good chance that the super-child will be successful. If that happens, the quality of the gene pool will significantly increase. Dividing by the time of creation is an idea only slightly better than the random partitioning. It is based on the assumption that there is a connection between the vertex structure and the time of creation of the vertices. According to our tests, spectral partitioning [19] takes much time. Even with a simple graph of 500 vertices, only finding the eigenvalues took 3 seconds. Creating the partitions takes precious time from the actual work, so it has to be fast. GA, combined with an idea similar to the k-means algorithm [20], seems to be the best choice, since this sub-problem is hard, but the solution does not have to be optimal, only close to optimal. The GA can stop after 1 second, and it still gives a relatively good solution.

4 Genetic Algorithm for Partitioning

The first step for both the genetic algorithm and the random method is to determine the number of partitions. 10 partitions for a graph with 10 vertices would result the same as there would not be groups at all, but little number of partitions for a big graph is not ideal either. After experimenting with various functions to determine the optimal number of partitions, we found the following function to work the best:

$$|P| = \frac{|V|^{0.6}}{3}$$

The GA for partitioning has to be fast, so the genome structure is not as straightforward as the genome structure of the main GA.

Let $R = \{R_1,...,R_k\}$ be the genome of an individual, where $k = |P|$, it is the number of partitions we plan to create. Every gene R_i denotes a randomly chosen vertex from the vertices of the graph, it is the starting vertex of the partition.

The fitness function first simulates a deterministic vertex-conqueror game for every individual. The rules of the game are:

Every partition P_i has a set of conquered vertices (CV), initially they contain only the randomly chosen starting vertex R_i. (If a partition has a starting vertex that is already taken by another partition, i.e., the same vertex is chosen for two different partitions, then the new partition conquers the previous partition, and the previous partition leaves the game with no conquered vertices, thus we obtain one less partition).

In every round, each vertex of the *CV* of each partition expands, meaning that they conquer every neighbor vertices, except those already conquered by any partition.

When no partition can expand any more, the first partition conquers every non-conquered vertices if there are any left (may happen in case of not connected graph). The game ends here.

The fitness value is as follows:

Let *g* be the number of edges of the graph that connect vertices belonging to the same partition, and let *w* be the number of edges that connect vertices belonging to different partitions.

Let $a = \frac{g}{g+w}$, this is a number between 0 and 1, as *g* and *w* are both positive numbers.

Let,

$$cmax = \max_{i=1..|P|} \{CV_i\}$$

Let,

$$cmin = \min_{i=1..|P|} \{CV_i\}$$

Let,

$$b = \frac{cmin}{cmax}$$

a number between 0 and 1, as $cmin \leq cmax$, and both are positive numbers

The fitness value is: (ab^2). The aim is to have (almost) equal size partitions and it is better if the most of the edges are inside partitions.

For every new generation, the GA for partitioning uses elitism, roulette wheel selection, one-point crossover, and mutation. This algorithm stops after one second, meaning that using the intervals adds only one second penalty (i.e., cost) to the main algorithm.

5 Greedy Error Correction

In this paper, we explain how phenotypes are produced from genotypes with a deterministic greedy algorithm.

There is a problem with the simple GA approach: most of the time, the genotypes of the solutions are not even covering the graph. Finding only a cover set with pure natural selection and random mutations is difficult for a simple GA, but the aim is to find the minimal cover set, so the task is even more complicated.

Fortunately, there are easy methods to create a cover set from an imperfect cover set.

The simplest and fastest method is to check every edge. If an edge is not incident to any vertex in *C* (unstable edges), then add one of the vertices of this edge to *C*. Some nodes may be redundant; these unnecessary nodes can be removed [6].

We analyze the usage of a greedy method, first it checks every vertex that is not in *C* (unselected vertices), but adds them to *C* only if they have the highest number of unstable edges incident to them. In the next cycle, it checks every unselected vertices again, until there are no more unstable edges. This method is slower, because it runs for every genome in every generation, but our tests showed that the algorithm gets significantly better with the greedy error correction.

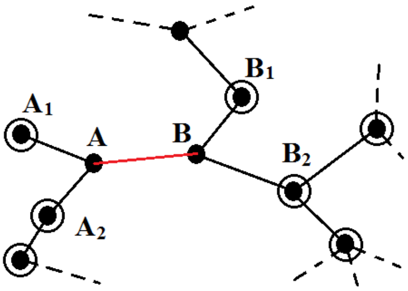


Figure 3
A subgraph with the last cycle of the greedy algorithm where each vertex has a maximum of one unstable edge

This correction works if the genome selects too few vertices. It is possible to correct those genomes that selects too many. Invert correction: if the neighbors of a selected vertex *X* are all selected vertices, then *X* can be removed from *C*.

Combining these ideas, it is possible to further improve these methods. At the last cycle of the greedy correction, there are no unselected vertices with multiple unstable edges. In this cycle, the choice of which vertex of an unstable edge should be selected is arbitrary. In some case however, this choice has an impact on the final solution, because of the final invert correction.

Consider the subgraph shown in Figure 3.

Between vertices A and B, there is an unstable edge. Since the greedy correction is at its last cycle, the algorithm can be sure that there are no vertices with more than one unstable edges. Suppose the algorithm first discovers vertex A. Vertex A has one unstable edge. The other vertex of this unstable edge is vertex B. It is granted that B has no other unstable edges, only this one.

Let us introduce a new concept, the compatible neighbor.

Definition 1 (Compatible neighbor) At the last cycle of the greedy correction, if B and A are neighbor vertices in a graph, A is selected, and beyond B, A has only selected neighbors, then A is a compatible neighbor of B.

The algorithm will decide as follows.

If B has more compatible neighbors than A, then select B, else select A.

Compatible neighbors are important, because the invert correction will deselect them if all their neighbors are selected. Since A has two compatible neighbors (A_1 and A_2), and B has one compatible neighbor (B_2), the algorithm will select A, in our example.

Lamarckian inheritance is the idea that characteristics developed during an organism's lifetime (the greedy error correction, in our case), can be inherited [21]. The question is if the algorithm should follow this idea, and pass through the genes reflecting the solution after the correction, or not.

Since the greedy error correction is not a stochastic method, it creates the same phenotypes for a given genotype. If the child inherits every gene of a parent's genotype perfectly, it will develop the same phenotype as the parent did. Furthermore, it is possible that the algorithm creates the same phenotype for different genotypes. That is why, using the Lamarckian inheritance could lead to a less varied gene pool, which is not desirable. By inheriting the original genes instead of the information in the phenotype, we increase the speed of converging to a near-optimal solution and increase the population diversity. Both of these features are important for a good GA [22]. On the other hand, the greedy error correction usually increases the number of vertices in C, therefore we expect better results in the new generation if these vertices in the genotype will be selected only if necessary.

In memetic algorithms some local search is done at the new individuals after the genetic operations [23, 24]. Our greedy error correction has a similar feature, but there is an important difference: we do not modify the genome (allowing to inherit its original version), only the phenotypes are vertex covers surely.

6 The Minimum Vertex Cover Algorithm

In this section, we describe how the new generation is computed in our GA.

The first 3 genomes of the next generation are 3 genomes of the previous population with the best F values. This is called elitism, it ensures that the best solutions will always stay in the population, with elitism, there is no chance of losing quality due to mutation.

The next genome of the next population is the super-child S. For every i = 1,2, ...,m, where $m = |P|$, let genome B_i be the genome of the previous generation with the best value I_i. $K_j^S = K_j^{B_i}$ for every $j = P_{i_1}, P_{i_2},..., P_{i_q}$, where K^S is the vector K of the super-child, K^{B_i} is the vector (list) K of B_i, P_{i_x} is the x-th member of P_i, and $q = |P_i|$.

In other words, the super-child inherits the best genes of the population for every interval, from one, two, or more parents.

For the rest of the population, our GA uses roulette wheel selection to select two parents, then one-point crossover to create two children from the two parents. Each gene of each child created with the one-point crossover, has a chance to mutate. By default, this chance is $1/|V|$ (where $|V|$ is the number of vertices), but it can be changed. If a gene in K mutates, then its value changes from false to true, or from true to false. For example: K = (0,0,0,1,1). If the second and fourth values mutate, the new value of K is: K = (0,1,0,0,1).

The new generation has the same size of population as the previous one had. The old generation is completely replaced by the new one (created in these steps).

Naturally, after the next generation is ready, the algorithm updates the values F and I of each genome. The algorithm stops after a given time.

7 Example

In this section, we illustrate the work of our algorithm in a small example. There is a population of 3 solutions for a graph, and the partitions of the graph is known ({1,2,3,4},{5,6,7}). The order of the vertices of the graph is shown in Figure 4. The first generation is also known. We use black and grey color for the partitions (intervals).

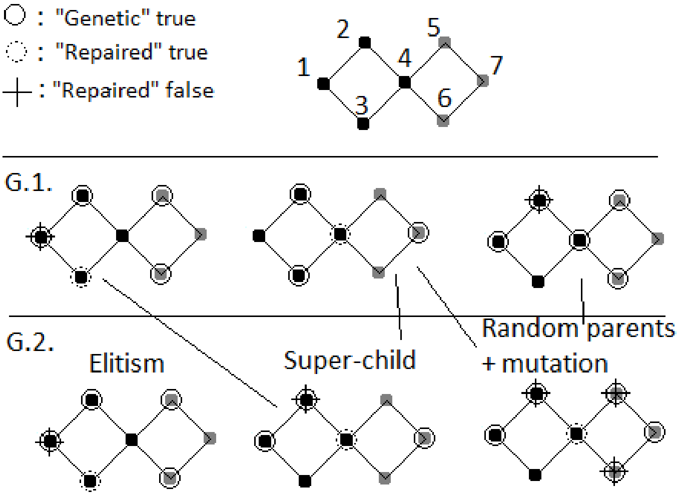


Figure 4
A simplified illustration of the algorithm

To repair the first solution, the algorithm has to complete the graph cover. The algorithm is greedy, but there is only one edge without cover vertices. The edge is between vertex 3 and 4. Selecting vertex 3 would result one unnecessary cover vertex, selecting vertex 4 would result one unnecessary cover vertex too, the algorithm selects vertex 3. Vertex 1 becomes unnecessary. Therefore, this cover has 4 vertices.

Solution 2 has edges without cover vertices too. Since the algorithm is greedy, vertex 4 is selected, because it will cover 2 uncovered edges. Solution 3 is a cover set, but it has one unnecessary vertex.

The size of the final cover set of each solution in the first generation is 4. But on the black interval solution 1 and 3 are better, and on the grey interval solution 2 is better.

The first solution of the new generation is the first best solution of the previous generation. Since they all had the same fitness value, solution 1 is now chosen as the elite.

The second solution is the super child, with the best intervals. It inherits the genes of the black interval of solution 1, and the genes of the grey interval of solution 2. Vertex 4 is selected because it will cover 3 uncovered edges. Vertex 2 is unnecessary. Therefore, this cover has 3 vertices.

The third solution inherits genes from two parents randomly, then mutates. It has one edge without cover vertices. The algorithm selects vertex 4, because it will result 3 unnecessary vertices. This cover has 3 vertices too.

We have tested our algorithm on large graphs as we describe it in the next section.

8 Experimental Results

We compare our algorithm with a fast greedy algorithm, a normal GA with simple error correction, "TVCA", "Darwin" and “NOVCA”.

TVCA of Ashay Dharwadker is a great algorithm for the vertex cover problem [25]. It always finds the cover set with size of a given number *k*, if there is any, but it may take a long time. TVCA first fills *C* (the cover set) with the members of *V* (vertices of the graph). The algorithm then finds the elements it can take out from *C*. But this is not the only difference between the two algorithms: TVCA is a deterministic algorithm. We used the original TVCA demonstration program, however, in this way, we could not manage to make tests in an automatic way. Thus, we tested manually some of our graphs (and not all of them).

Darwin is a programming environment developed for ETH Zurich [26]. Amongst many functions, Darwin is able to find vertex covers of a graph. This function is implemented by Gaston H. Gonnet, and it is based on another very good approach, the fixed-parameter algorithm [27]. Unfortunately, Darwin is designed for small graphs, so we could not test all our graphs with it. Actually, with Darwin, we could only test graphs where MGA, Darwin and NOVCA all found the same result almost immediately.

NOVCA is a deterministic polynomial time algorithm that we implemented based on the pseudo code described in its paper [28].

Our algorithm (**MGA**) is fast and accurate in almost any cases, except with near-complete graphs, where almost every node is connected to almost every node, like the Witzel Graph [25].

We have randomly generated 100 graphs with 500 vertices and 2500 edges, then tested GA, greedy, NOVCA, and MGA on each graph. For the results, see Figure 5 and Table 1. For 10 graphs, we tested TVCA as well, it found the vertex cover size of the MGA algorithm in 9 seconds on average.

Table 1

Test data for relatively small graphs: Vertices: 500, Edges: 2500 (average of 100 random graphs)

Algorithm	1 second	2 seconds	5 seconds	9 seconds
TVCA	n/a	n/a	n/a	same as MGA*
Greedy	360.24	n/a	n/a	n/a
GA	361.86	359.82	357.91	n/a
NOVCA	351.63	n/a	n/a	n/a
MGA	356.69	352.11	350.52	n/a

* TVCA was manually ran on 10 graphs and compared to MGA.

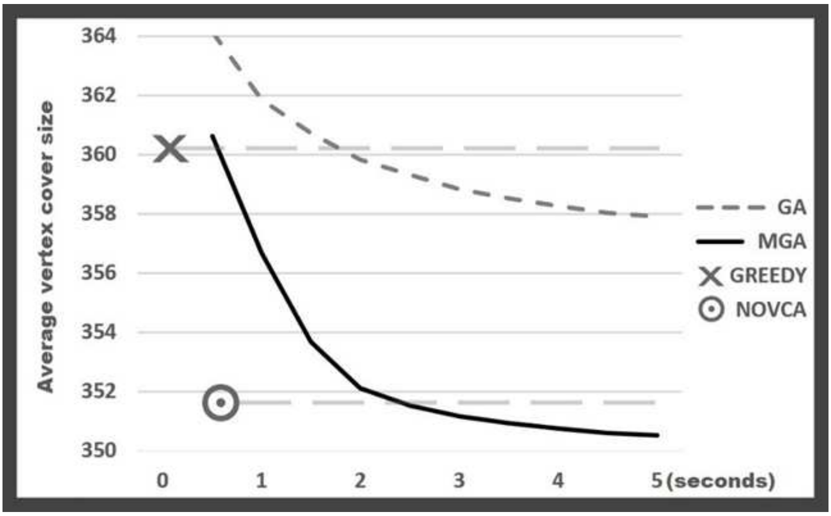


Figure 5

The average results of 100 random graphs with 500 vertices and 2500 edges

We also created an algorithm to generate "clustered graphs", where we assigned each vertex to a cluster, and increased the chance of edges between vertices from the same cluster. The reason for this is that many graphs in nature are clustered, and also we expected that our algorithm would have an advantage with these. However, for more realistic results, the number of clusters is not the same as the number the partitioning algorithm in MGA uses.

We have repeated the above test with the same parameters on randomly generated clustered graphs. We tested TVCA too, and this is where we reached its limit, because 8 out of 10 cases, in 2 minutes TVCA was not able to find the same result that MGA found in 5 seconds. For detailed results, see Figure 6 and Table 2.

Table 2

Test data for clustered graphs: Vertices: 500, Edges: 2500 (average of 100 random graphs)

Algorithm	1 second	2 seconds	5 seconds	120 seconds
TVCA	n/a	n/a	n/a	found MGA in 2 out of 10 cases
Greedy	562.80	n/a	n/a	n/a
GA	364.40	362.11	360.16	n/a
NOVCA	356.93	n/a	n/a	n/a
MGA	360.66	354.33	353.04	n/a

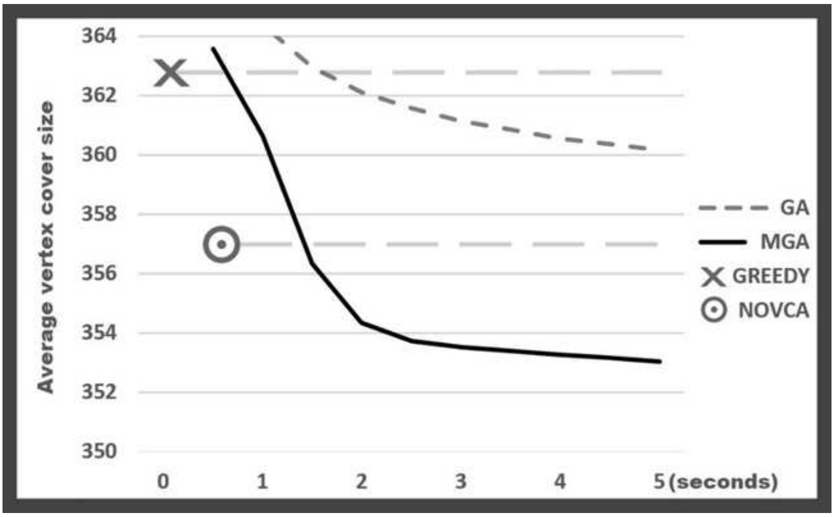


Figure 6

The average results of 100 random clustered graphs with 500 vertices and 2500 edges

We repeated the test on 100 random (not clustered) graphs with 2000 vertices and 10000 edges. For the results, see Figure 7 and Table 3. For 10 of the graph, we have tested TVCA as well, but it never found the result of MGA/NOVCA in 10 minutes.

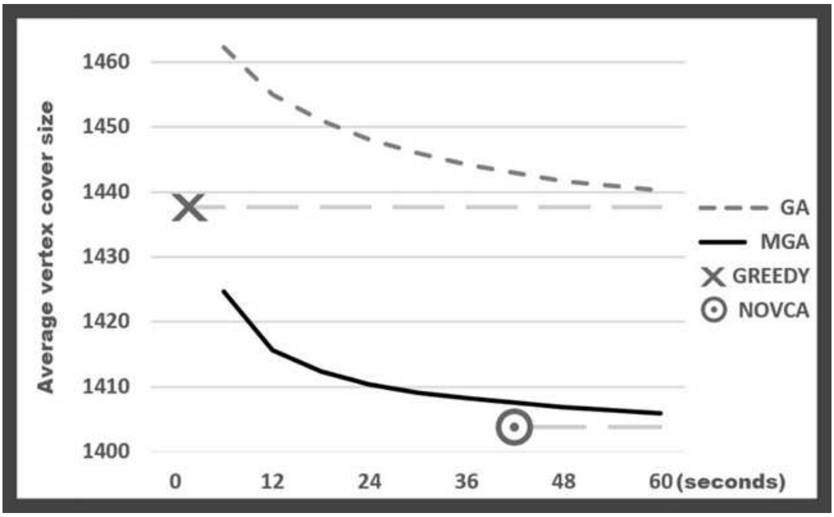


Figure 7

The average results of 100 random graphs with 2000 vertices and 10000 edges

Table 3

Test data for large random graphs: Vertices: 2000, Edges: 10000 (average of 100 random graphs)

Algorithm	1 second	12 seconds	42 seconds	60 seconds	600 sec
TVCA	n/a	n/a	n/a	n/a	n/a
Greedy	1437.74	n/a	n/a	n/a	n/a
GA	n/a	1454.99	1442.91	n/a	n/a
NOVCA	n/a	n/a	1404.02	n/a	n/a
MGA	n/a	1415.51	1407.52	1405.95	n/a

Further, 100 random clustered graphs with 2000 vertices and 10000 edges were also tested. For the results, see Figure 8 and Table 4. For 10 of the graph, we tested TVCA as well, but it did not manage to find the result of MGA/NOVCA in 10 minutes.

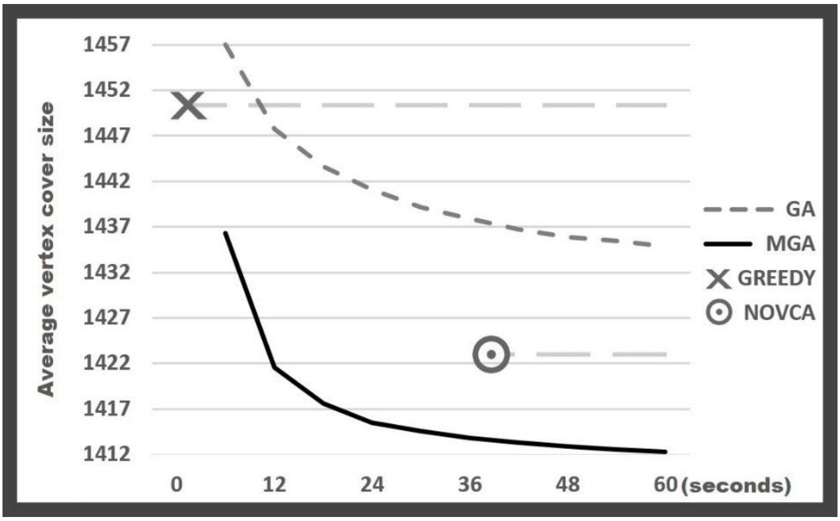


Figure 8

The average results of 100 random clustered graphs with 2000 vertices and 10000 edges

Table 4

Test data for large clustered graphs: Vertices: 2000, Edges: 10000 (average of 100 random graphs)

Algorithm	1 second	12 seconds	39 seconds	60 seconds
TVCA	n/a	n/a	n/a	n/a
Greedy	1450.57	n/a	n/a	n/a
GA	n/a	1447.74	1437.86	1434.89
NOVCA	n/a	n/a	1423.03	n/a
MGA	n/a	1421.52	1413.79	1412.27

Table 5
Test on Witzel Graph (Vertices: 450, Edges: 17827, M.V.C. Size: 420)

Algorithm	2 second	43 seconds
TVCA	n/a	420
Greedy	428	n/a
GA	n/a	426
NOVCA	424	n/a
MGA	n/a	426

On the Witzel Graph with 450 vertices and 17827 edges, TVCA finds the minimum vertex cover of 420 vertices in 43 seconds. MGA finds a vertex cover of 426 vertices in the same amount of time. NOVCA almost immediately finds a vertex cover of 424 vertices (see Table 5).

Conclusions

Our modified GA yielded, minimally, the same solution as a greedy algorithm, but more often, our algorithm was better. The reason is, that the first generation has a blank genome and the greedy error correction works in a similar way, as a greedy algorithm. The modified GA is a big improvement for the GA. In case of large graphs, it converges much faster with the interval fitness method, which can be viewed as a problem reduction approach. At the same time, the algorithm focuses on the whole problem. With the error correction, almost all members of the population had a better fitness and the genetic variation is much better too. But the biggest advantage of the error correction, is obviously the fact that it mixes the fast and efficient greedy method, with the heuristic GA, taking advantage of both approaches. Distinguishing of genotypes and phenotypes, combined with the intervals also leads to good efficiency. By inheriting only, the backbone (genotype) of the covering sets, it is easy to create efficient hybrids of solutions.

In the case of a difficult graph, the basic GA is not as efficient as a simple greedy algorithm, not even with more time given. As Figure 7 shows, even after 60 seconds, the GA did not find the solution that greedy found almost immediately. With greedy error correction, it is possible that the genome of a solution is blank, meaning every element of set K has a false value. Since the greedy error correction step itself is a good method for the problem as well, it is a good thing if the GA cooperates with the greedy method, by only selecting genes that improve the fitness. For this, the GA has to start with at least one blank genome in the population.

GAs tend to converge towards local optimum. To avoid this, it is worth to make experiments with multiple populations. Since members of different populations have no chance to mix their genetic information, multiple populations develop independently, and it is possible that each population converges towards different local optima. It also means that the different populations can use different processors. Another idea to solve the problem of local optimum is to restart the algorithm after a given time or generation. Obviously, the algorithm has to save the best solution, but the new population must not remember that genome, it has to start with random genomes again.

In the Experimental Results section, we see that the Greedy algorithms is the fastest, but it is also, the most inaccurate and it has no chance to improve with more time. TVCA [25] and Darwin [26] are very good in finding a small vertex cover, but is very slow and is not usable with large graphs. As Table 2 shows, TVCA needed 2 minutes to have a chance of finding a result MGA found in 5 seconds. NOVCA [28] is similar to Greedy, but it yields much better results at the cost of being a bit slower. Our algorithm, MGA, yields similar results to NOVCA, but has several advantages. In some graphs (Witzel, and fully random graphs) it may underperform a bit, but on other graphs (especially in clustered graphs) – MGA outperforms NOVCA. MGA also has the advantage of providing early results, and can run longer to find even smaller vertex covers. See Figure 8, where MGA found a better result in 12 seconds than NOVCA in 39, and then continued to improve.

Thus, it seems that the idea to using interval-valued fitness is a worthwhile consideration. We note that another type of combination of GA, with interval-values is found in [29].

References

[1] Álmos Attila, Horváth Gábor, Várkonyiné Kóczy Annamária, Győri Sándor: Genetikus algoritmusok, Typotex Kiadó (2003) [Genetic Algorithms, in Hungarian]

[2] David E Goldberg: Genetic Algorithms in Search Optimization and Machine Learning., Addison Wesley (1989)

[3] Benedek Nagy, Elisa Valentina Moisi: Binary tomography on the triangular grid with 3 alternative directions - a genetic approach, ICPR 2014: 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 1079-1084, IEEE Computer Society (2014)

[4] Shaohua Li, Jianxin Wang, Jianer Chen, Zhijian Wang: An Approximation Algorithm for Minimum Vertex Cover on General Graphs, Proceedings of the Third International Symposium on Electronic Commerce and Security Workshops (ISECS 10) Guangzhou, P. R. China, 29-31 July, pp. 249-252 (2010)

[5] P. Oliveto, J. He, X. Yao: Analysis of the (1+1)-EA for Finding Approximate Solutions to Vertex Cover Problems, IEEE Transactions on Evolutionary Computation, 13(5), 1006-1029, October (2009)

[6] Martin Pelikan, Rajiv Kalapala, Alexander K. Hartmann: Hybrid Evolutionary Algorithms on Minimum Vertex Cover for Random Graphs, GECCO '07 Proceedings of the 9th Annual Conference on Genetic and evolutionary computation ACM New York, NY, USA, pp. 547-554 (2007)

[7] Ali Karci, Ahmet Arslan: Bidirectional evolutionary heuristic for the minimum vertex-cover problem, Computers and Electrical Engineering, 29/1, 111-120, Elsevier (2003)

[8] Ketan Kotecha, Nilesh Gambhava: A Hybrid Genetic Algorithm for Minimum Vertex Cover Problem, In Prasad, B. (Ed.), The First Indian International Conference on Artificial Intelligence, 904-913 (2003)

[9] D. B. Fogel: Phenotypes, genotypes, and operators in evolutionary computation, in Proc. IEEE Int. Conf. Evolutionary Computation (ICEC'95) New York: IEEE Press, 193-198 (1995)

[10] Benedek Nagy: A general fuzzy logic using intervals, 6th International Symposium of Hungarian Researchers on Computational Intelligence, Budapest, Hungary, 613-624 (2005)

[11] Benedek Nagy: Reasoning by Intervals, Diagrams 2006, Fourth International Conference on the Theory and Application of Diagrams, Stanford, CA, USA, 145-147 (Lecture Notes in Computer Science - Lecture Notes in Artificial Intelligence, LNCS-LNAI 4045)

[12] Benedek Nagy: An interval-valued computing device (2005) CiE 2005, "Computability in Europe": New Computational Paradigms, Amsterdam, Netherlands, 166-177

[13] Benedek Nagy: Effective Computing by Interval-values, INES 2010, 14th IEEE International Conference on Intelligent Engineering Systems, Las Palmas of Gran Canaria, Spain, 91-96 (2010)

[14] Benedek Nagy, Sándor Vályi: Interval-valued computations and their connection with PSPACE, Theoretical Computer Science - TCS 394/3, 208-222 (2008)

[15] Benedek Nagy, Sándor Vályi: Prime factorization by interval-valued computing, Publicationes Mathematicae Debrecen 79/3-4 (2011) 539-551

[16] Benedek Nagy, Sándor Vályi: Computing discrete logarithm by interval-valued paradigm, (Benedikt Loewe, Glynn Winskel, eds.), Proceedings 8th Workshop on Developments in Computational Models - DCM 2012, Cambridge, England, Electronic Proceedings in Theoretical Computer Science - EPTCS 143 (2014) 76-86

[17] Benedek Nagy, Sándor Vályi: An Extension of Interval-Valued Computing Equivalent to Red-Green Turing Machines, MCU 2018: 8th International Conference on Machines, Computations, and Universality, LNCS 10881 (2018) 137-152

[18] R. M. Karp: Reducibility among combinatorial problems. In: (R. E. Miller and J. W. Thatcher, Eds.) Complexity of Computer Computations, Proc. Sympos. IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., Plenum Press, New York, 1972, pp. 85-103

[19] F. McSherry: Spectral partitioning of random graphs, Foundations of Computer Science, 2001, Proceedings. 42nd IEEE Symposium on 8-11, 529-537, Oct. (2001)

[20] J. A. Hartigan, M. A. Wong: Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society, Series C (Applied Statistics) 28 (1), 100-108 (1979)

[21] G. M. Morris, D. S. Goodsel, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, Journal of Computational Chemistry 19(14), 1639-1662 (1998)

[22] Xin Yao: An empirical study of genetic operators in genetic algorithms, Microprocessing and Microprogramming, 38/1-5, 707-714, Elsevier (1993)

[23] V. Di Gesu, G. Lo Bosco, F. Millonzi, C. Valenti: A memetic algorithm for binary image reconstruction, Lecture Notes in Computer Science 4958, 384-395 (2008)

[24] Benedek Nagy, Elisa Valentina Moisi: Memetic algorithms for reconstruction of binary images on triangular grids with 3 and 6 projections, Applied Soft Computing 52, 549-565 (2017)

[25] Ashay Dharwadker: The Vertex Cover Algorithm, CreateSpace. http://www:dharwadker:org/vertex cover/ (2011)

[26] G. H. Gonnet, M. T. Hallett, C. Korostensky, L. Bernardin: Darwin v. 2.0: an interpreted computer language for the biosciences, Bioinformatics 16: 101-103, http://www.cbrg.ethz.ch (2000)

[27] R. Balasubramanian, M. R. Fellows, V. Raman: An improved fixed-parameter algorithm for vertex cover, Information Processing Letters 65, 163-168 (1998)

[28] Sanjaya Gajurel, Roger Bielefeld: A Simple NOVCA: Near Optimal Vertex Cover Algorithm, Procedia Computer Science, Volume 9, 747-753, Elsevier (2012)

[29] Lajos Zámbó, Benedek Nagy: Optimization of the Painting Problem by a Genetic Approach using Interval-values, CINTI 2011: 12th IEEE Int. Symp. Computational Intelligence and Informatics, 21-22 November, 2011 Budapest, Hungary, 127-132 (2011)

Bus Transport Process Networks with Arbitrary Launching Times

Zsolt Ercsey[1], Albert Nagy[2], József Tick[3], Zoltán Kovács[4]

[1] Department of System and Software Technology, Faculty of Engineering and Information Technology, University of Pécs, Boszorkány u. 2, 7624 Pécs, Hungary, e-mail: ercsey@mik.pte.hu

[2] Doctoral School of Applied Informatics and Applied Mathematics, Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary, e-mail: albert.nagy@me.com

[3] John von Neumann Faculty of Informatics, Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary, e-mail: tick@uni-obuda.hu

[4] Optin Ltd. Oroszlán u. 4, 6720 Szeged, Hungary,
e-mail: zoltan.kovacs@optin.hu

Abstract: The current paper is about a process network synthesis solution for a bus transport problem, where arbitrary launching data are given in an available timetable. Here, the bus transport problem is presented as an application of the p-graph methodology and is investigated from structural point of view. Focusing on the synthesis step, the bus transport process network is generated and detailed. Based on the maximal structure of the problem a mathematical programming model is generated which has the advantage of a smaller number of variables and constraints than the conventional mathematical models for similar problems. The solution of the mathematical programming model, results in the optimal solution of the bus transport problem. From the traffic point of view, an example of medium difficulty is solved, by a publicly available solver.

Keywords: optimization; mathematical programming model; process network; p-graph; synthesis; bus transport

1 Introduction

Sustainable urban mobility demands significant attention for effective operations. Optimization of public transportation contributes to an attractive and healthy urban environment, while further improving overall competitiveness. The industry has seen major technological advancements over the last decade: Buses represent more than just a way to get from one destination to another. Emerging new technologies, automation of processes, car sharing and other factors influencing public transportation habits also rapidly and continuously alter situations. It is of high importance that public transport service companies exploit their resources at the maximum level, while focusing on the most effective service availability for their passengers.

The rapid adoption of telematics systems at bus operators has extended the possibilities of the historically available user generated data which improved the monitoring of services, further motivating the operators to enhance daily operations. These data can be used both at the strategic planning and the tactical planning, considering frequency setting; robust timetable design as well as vehicle and crew scheduling. Optimizing this as a whole is an immensely complex issue, even its subtasks may lead to NP hard problems. Until recent past, solution of even small subproblems often met computational difficulties. The dynamic development of the technology, and the continuous improvement of the methodologies and models, together with the applied solution algorithms, make it now possible to solve real size, real case problems.

In this paper operational optimization issues are considered, where bus transport service companies usually have more remarkable influence. Since operational costs cover the main part of the expenditures of these service companies, any savings in this regards have great leverage. It is under their direct control, how their bus fleet is scheduled, how their drivers are assigned to the various work shifts both in the short and in the long run. In general, this problem involves legal perspective, employee interests, as well as individual claims. The constraints correspond to international and national regulations as well as company specific standards in terms of working hours, driving times, breaks or rest periods, depot considerations, labor availabilities etc. It is obvious, that each bus transport service company has its own specific expectations and constraints due to its particular business situation, which had to be handled together with common considerations. The decision support system suggested here is based on the p-graph methodology. The current method was developed and presented for the Budapest Transport Corporation, from where the published example originates. The purpose of the work was to effectively solve a vehicle and driver scheduling task that can be injected into the company's available information system. The paper is structured as follows. In Section 2 a literature review on vehicle scheduling problems and on the p-graph methodology is followed by the problem definition. In Section 3 the proposed solution framework is presented and detailed, namely (i) the maximal structure of the bus transport synthesis problem with arbitrary launching times is presented, the key elements of the *i*-th launch is highlighted and discussed, (ii) the mathematical programming model that suits the particular maximal structure is presented and (iii) is followed by an example. Section 4 presents conclusions for the most important results. Based on the synthesis step, the presented method has the advantage of being limited in the number of variables and constraints and therefore publicly available solvers are capable of generating results of real size problems.

2 Materials and Methods

2.1 Vehicle Scheduling Problems

Public transport scheduling problems are very complex, most of them are NP-complete when examining them from a theoretical point of view. Frequently, the proposed models are suggested for selected situations considering rules relating to timetabled and overhead trips, bus types and capacities. Commonly, it is not possible to include other conditions that arose in other real application environments. Several papers distinguish the problems by the number of device types and the number of depots. For example, the Single Depot Vehicle Scheduling Problem (SDVSP), where the vehicles belong to a single vehicle type and are located in the same physical location, and the Multiple Depot Vehicle Scheduling Problem (MDVSP), which considers different scheduled trips (and their vehicles) with different special needs. Alternative-fuel vehicles are also gaining advantage, which implies further research, thus assigning the available fleet of vehicles to service the given set of trips with specific start and end times.

The heart of these methods is the mathematical programming model, e.g. MILP or MINLP, which supposedly ensures the optimality of the resultant solution. Their algorithmic generation is a common step in most of the papers and is rarely detailed. Since the crucial difference is in the underlying mathematical programming models, these are elaborated and illustrated. Nevertheless, it is often difficult to generate and solve the complete models.

Bodin et al. [1] summarizes routing and scheduling problems of vehicles and crews. It classifies and categorizes routing and scheduling problems, as well as reviews algorithmic techniques and solution methodologies. The applicability of the results for bus transport problems is limited, however. Time-space network flow model is suggested in Kliewer et al. [2] involving multiple depots for vehicles and different vehicle types for bus scheduling. Integer programming methods with heuristic solution techniques are proposed by Dávid and Krész [3] and Tóth and Krész [4]. More recently a branch-and price based solution is suggested by Horváth and Kis [5] and a set partitioning-based mathematical model, where most vehicle-specific activities can be integrated based on the desired constraints is presented in [6]. This model is solved using a column generation approach. Békési and Nagy [7] presented a model to automatically calculate approximately optimal vehicle and driver schedules. For a given list of trips and considering company specific requirements and parameters in compliance with regulations.

2.2 P-Graph Methodology and Approaches

The solution of vehicle scheduling problems are mainly based on the mathematical programming models as discussed in the previous section. During the generation of the mathematical programming models, directed graphs are often used to support certain steps. These directed graphs themselves usually correspond to the execution sequences only, in other words, they relate to the precedence order. All other pieces of information are hidden within the attributes of the nodes and edges of the graphs. These pieces of hidden information then appear within the variables and constraints of the mathematical programming models.

Another possible approach is to exploit the unique features of the structure of the problem at the earliest possible stage of the solution; this approach is followed in the present paper also. Conventional directed graphs are suitable for representing and analyzing processes, but they are not suitable for the synthesis step, see Friedler *et al.* [8].

The p-graph methodology has its origin in the early 90s, when Friedler *et al.* [8] first applied specially constructed directed bigraphs to depict and solve chemical engineering processes. In a chemical process, operating units transform their input materials to output materials by transforming the quality and quantity of the materials under consideration. To unambiguously represent this transformation, materials and operating units are represented by two separate types of vertices of the p-graph, while the interconnections are described by the arcs of the graph, i.e. arcs represent whether a material is consumed or produced by an operating unit. With the focus on this representation, certain combinatorial properties were formalized, i.e. a set of axioms was constructed. The first axiom states, for example, that every product should be represented in the p-graph. This set of axioms express the necessary and sufficient combinatorial properties to which a feasible process structure should confirm. In other words, based on these axioms the set of potentially feasible solution structure can be generated and conversely no other structures need be considered when the potentially feasible solutions are sought. Further, the maximal structure is defined as the union of all feasible solution structures. Please note that this maximal structure is of outmost importance in the p-graph methodology. Generally speaking, the maximal structure is similar to a "super-structure," however this later does not have a formal definition. With the focus on the underlying structure generation of the problem which serves as the basis of the mathematical programming model to be solved, all feasible solution structures can be enumerated algorithmically and listed according to their related results, see Friedler *et al.* [9]. Therefore, should a feasible solution be found by any method for the synthesis problem considered, then it is certain that this very solution is a subset of the maximal structure and further, it can be enumerated by the p-graph methodology. And conversely, no other solutions should be considered but only those enumerated by the p-graph methodology.

Kovács *et al.* [10] explored fundamental structural properties of separation networks and studied the validity of the applied mathematical programming models specifically with respect to redundancy; classes of separation network synthesis problems were also studied extensively. They demonstrated that the synthesis step is critical when optimal separation networks are to be sought.

Sanmarti *et al.* [11] applied p-graphs to solve the scheduling of multipurpose batch plants. Süle et al [12] proposed the solution of supply chains with limitations as well as uncertainties of the renewable resources, i.e. the overall reliability of raw material availability is also considered when generating the n-most profitable supply chain alternatives. Tick *et al.* [13] applied p-graph methodology to business process modelling, giving both a nominal cost and an extended cost to each operating unit, while taking into consideration that only a limited number of operating units has the extended cost and the others have the nominal cost. A branch and bound based solution algorithm was presented together with a polynomial time dynamic programming algorithm for special problems. Vincze *et al.* [14] transformed CPM problems to p-graphs to manage the situation of alternative tasks and solutions within one step. Several corresponding mathematical programming models were given with illustrations how alternative cases appear in the structure. Later, Ercsey [15] solved a Hungarian clothing manufacturer's problem with p-graphs. Here, alternatives specified by mainly financial necessities as well as human resource constraints were managed within the proposed model.

An extension of the P-graph approach for multi-period process network synthesis is proposed by Tan and Aviso [16]. They offer a modification of the original p-graph approach to enable partial load operational lower limit for process units to be considered via the addition of fictitious streams. Tan *et al.* [17] combined p-graphs and Monte Carlo simulation approach to plan carbon management networks, generalized systems for minimizing emissions of CO_2. Their target was to identify robust and near optimal carbon management networks, which can be achieved based on the optimal and near optimal solutions generated by the p-graph methodology.

Cabezas *et al.* [18] considered sustainable process systems and supply chains in their review article. They thoroughly investigated the usability of the p-graph methodology as well as the connections towards other methods. They also illustrated the advantages of the potential application of p-graphs from the feasible structures point of view. Fan *et al.* [19] used p-graphs as a decision support tool to develop an integrated design of waste management systems in support of a Circular Economy. They considered a number of case studies of municipal solid waste compositions based on different country income levels, identified the most suitable treatment approach as well as the near optimal solutions in order to deal with the trade-offs between conflicting objectives that are difficult to monetize.

Recently, Bartos and Bertók [20] used p-graphs to determine the optimal load of automotive and electronic production lines where the assembly requires various components and significant number of human resources. Besides the algorithmic generation of the mathematical programming model and afterwards its resultant optimal solution, p-graphs here support the visual definition of the task to employee allocations. Bertók and Bartos [21] recently proposed an optimization method for energy allocation of renewable energy sources and storage considering different priorities of the producers, storage and consumers. They extended the original p-graph methodology to multi-period direction as well as to the direction where targets and intermediate entities cannot be accumulated.

König and Bertók [22] applied the p-graphs to freight transportation, and presented an algorithmic method which is capable to involve and enumerate all feasible scenarios, when analyzing the conditions of the contracts in case of uncertainty, i.e. which solution costs less or offer more flexibility. Bárány *et al.* [23] proposed a p-graph based method for minimizing cost and emission for vehicle assignment problems. There, the assignment problem is transformed into a p-graph, which provides the structural model and the basis for the solution. The approach includes the maximal structure generation, which further serves as the input for the generation and solution of the mathematical programming model. At the end, the optimal and a finite number of n-best sub-optimal networks in the ranked order are given.

Nagy *et al.* [24] already investigated bus transport problems within the p-graph methodology. There, the scenario considers various periods during the day, when the launching density of the buses can be considered to be the same; i.e. the scenario with a periodic timetable. With the focus on the synthesis step, the material type and the operating unit type vertices of the p-graph were specified and explained. Maximal structures for i) a one period and one bus without driver change problem, for ii) a one period and one bus with driver change problem and for iii) a two period and one bus without driver change problem were given together with a corresponding MILP / NLP mathematical programming model. From the maximal structure all feasible solution structures are generated. The corresponding mathematical programming model that suits to the periodic bus transport problem is explained in detail. This MILP / NLP model has to be solved for each generated solution structure, and the result prepares the generation of the launching table of the buses. In other words, bus transport problems with periods within the timetable are suggested to be solved in [24].

The current work expands the scenario of [24], namely when the timetable of the buses has no periods, but arbitrary individual launching times are defined is considered as the basic scenario of the present paper. This modification points towards a more general practical situation, nevertheless it has a fundamental effect on the p-graph elements, and therefore on the maximal structure, as well as on the related mathematical programming model that needs to be considered. Moreover, compared to the other methods cited in Section 2.1 where expansive mathematical programming models have to be solved often arduously, here the corresponding mathematical programming model can be solved even with publicly available solvers.

2.3 Problem Definition

Exploiting the above mentioned antecedents and results the following bus transport problem is discussed and solved hereinafter based on the p-graph methodology. Let the timetable of the buses with individual launching times be given, i.e. the starting times within the timetable can be arbitrarily given, and no periods can be determined within the timetable. The main goal is to determine the specific bus transport schedule that meets the given timetable. Obviously, besides this main goal, other viewpoints may also be considered and to be optimized, for example minimum fuel consumption of the buses considered, etc.

Let us consider the situation, where the departure station and the terminal station are given. Let the driving time of the turn also be given to each determined starting time. Obviously, the bus fleet is also given together with the number of different buses, the number of drivers etc. together with the parameters of the stance, secession. Obviously, labor standards of the drivers have to be considered and therefore, also be given in advance.

In summary, the following characteristics are considered:

- Bus route, containing information about the departure station and terminal station. Nametag, other stations etc. in connection with the bus route is also available.
- Bus turn, containing information about the bus route, and specific time information etc. The followings are considered hereinafter:
    - Bus launching times: P_1, P_2,…,P_N, (the launching time of the first bus: P_1, the launching time of the last bus: P_N)
    - Time required to perform the turn, T_1, T_2, …, T_N
    - The arrival time of the given bus, p_1, p_2, …, p_N, can be calculated from the launching time and the time required for the turn.
- Activities, containing information about rest, stance etc. both concerning duration, and whether the activity has to be considered within the working hours and driving hours.
    - Time for the rest period, in minutes, RT
    - Time for changeover, in minutes, AT
    - Time for driver change, in minutes, HT

  - Time for discharge, in minutes, LT
  - Time for stance, in minutes, ST
  - Time for entering the garage, in minutes, GT
- Labour standards, containing information about its type, working hours etc. The type can be normal or split:
  - Minimum working hours, in minutes, NWH
  - Maximum working hours, in minutes, XWH
  - Maximum hours until rest
- Fleet, containing information about the available buses
  - Number of buses, B
  - Number of drivers, D

Please note that the present problem definition does not depend on the bus line itself and therefore may consider multiple bus lines together. It is however important that the bus lines have the same departure and terminal stations and their routes and turns are explicitly specified and known as above. For example, both bus line *A* and bus line *B* be explicitly specified as above with a sole launch per bus line, with bus line *A* having an earlier launching time. Then, the data corresponding to the first launch, i.e. P_1, p_1 and T_1 will refer to the bus line *A*, while the second launch, i.e. P_2, p_2 and T_2 will refer to the bus line *B*.

3 Results

3.1 Solution Framework

The bus transport problem specified in the previous section is proposed to be solved by an algorithmic method as follows. First, the structural model, namely the maximal structure of the bus transport problem is generated and the parameters of the arcs and nodes are set. Please note that many characteristics of the problem can already be considered with the exact formulation of the structural model. Second, the corresponding mathematical programming model is generated and solved. The solution of the model serves as the solution of the bus transport problem with arbitrary bus launching times. Details are given in the subsequent chapters. The solution framework is given in Figure 1 with the emphasis on the synthesis step.

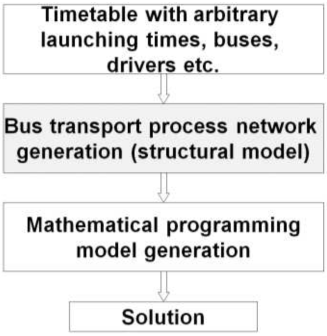

Timetable with arbitrary launching times, buses, drivers etc.
Bus transport process network generation (structural model)
Mathematical programming model generation
Solution

Figure 1
Solution framework of the bus transport synthesis problem with arbitrary launching times

3.2 Maximal Structure

The key elements of the maximal structure of the bus transport synthesis problem with arbitrary launching times are given in Figure 2. Please note that the subgraph represented by red color have to be considered within the maximal structure of the bus transport synthesis problem with arbitrary launching times as many times as many departure times are given within the timetable. Here, a bus and a driver after performs a turn. After the turn, the driver can go back to the garage or have a rest or have no rest. Drivers change is also acceptable. It is also visible, that after the completion of the turn the bus can perform further turns.

On this p-graph, buses and drivers are given as nodes of the raw material type, while garage and work shift end for the driver are given as nodes of the product type. Intermediate material type nodes represent departures and arrivals as follows: the arbitrary launches of the buses given in the timetable are represented by the nodes P_1, P_n, P_k, P_N, while the corresponding arrivals are represented by the nodes p_n, and p_N. Intermediate material type node with the label *M_n* represents the situation either without rest period after the completed turn, or with the rest period following the turn, or the situation when drivers change happen after the turn.

The operating unit type nodes represent the various activities, namely stance, turn, changeover, secession, no rest, rest and driver change. Since no rest, rest and driver change are parallel activities, should only one of them be within the solution of the *n*-th turn, the other two have to be omitted. Please observe that the size of the maximal structure given in Figure 2 depends on the number of individual launches; it is specified in Table 1 how the approach scales with the problem size.

It is important to mention that more complex bus driver scheduling problems may have the same or different maximal structures, depending on the problem. For example, real cases corresponding to the fact that only one or a maximum number of buses may be at the departure station at a time, since there is no more place for other buses for example, has no effect on the given maximal structure, this limitation has to be controlled within the subsequent mathematical programming model. Other cases, for example when driver change is permitted in between the departure and terminal stations of the turn corresponds to an altered maximal structure. This case is depicted in Figure 3, where the effect of this new condition is highlighted with violet color. Here, an additional operating unit type node representing the activity of the turn and the driver change together have the labels *TDC_i*. It is worth mentioning that since there are constraints on the possible working hours of the drivers, it is not realistic to perform a driver change at the last turn.

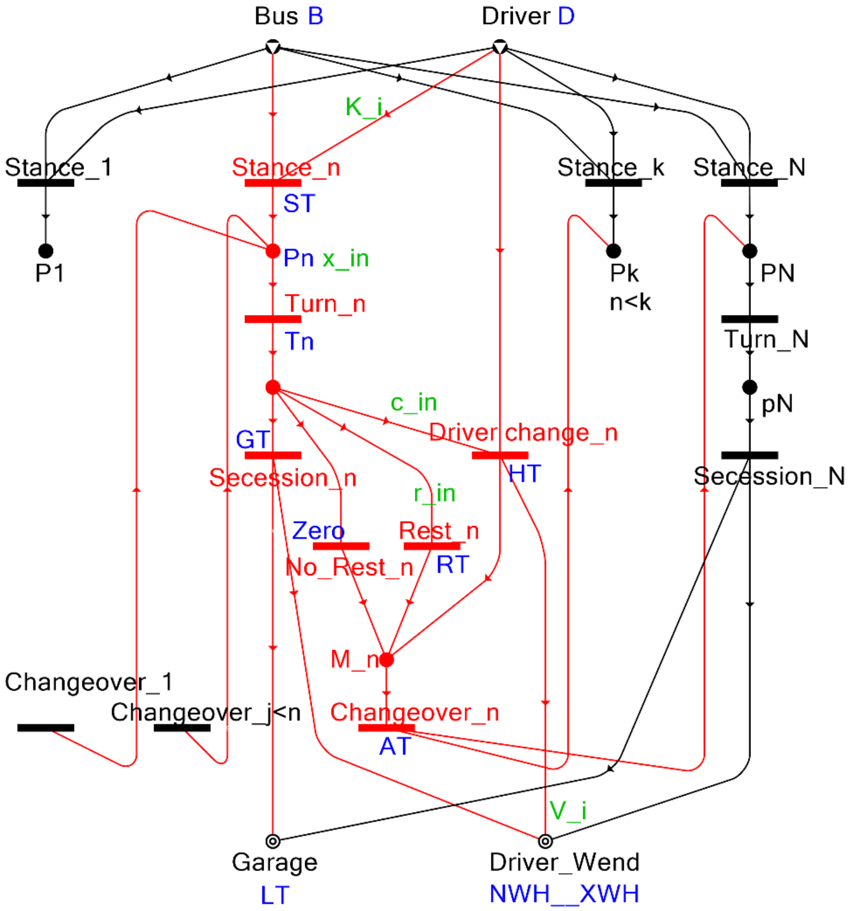


Figure 2

Key elements of the maximal structure of the bus transport synthesis problem with arbitrary launching times, the *n*-th launch is highlighted

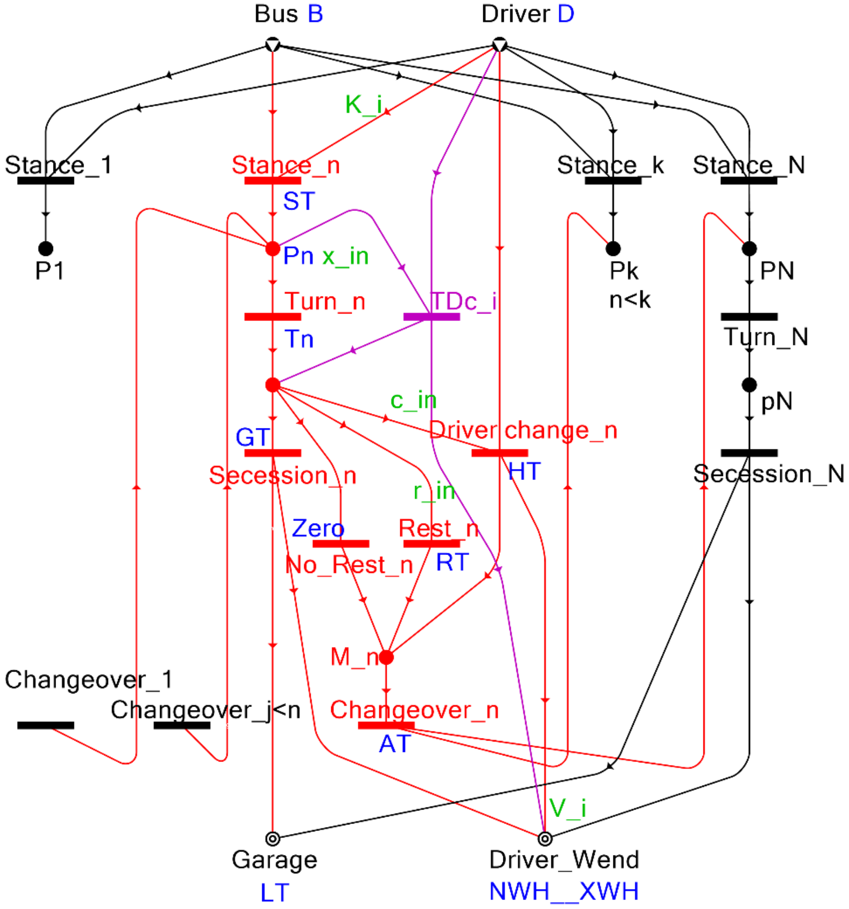


Figure 3

Key elements of the maximal structure of the bus transport synthesis problem with arbitrary launching times where driver change is permitted in between the departure and terminal stations of the turn, the n-th launch is highlighted

Table 1

Size of the maximal structure given in Figure 2, where n is the number of individual launches

Graph element	Size of the maximal structure
Edges	$\lvert E\rvert = 2n + 6n + \frac{n(n-1)}{2} + 9(n-1)$ $= \frac{1}{2}n^2 + \left(17 - \frac{1}{2}\right)n - 9$
Operating unit type nodes	$\lvert O\rvert = 7n - 4$
Material type nodes	$\lvert M\rvert = 3n - 1 + 4$

3.3 Mathematical Programming Model

In the previous chapter illustrated how the maximal structure of the bus transport process network, with arbitrary launches, is developed. Based on the maximal structure given in Figure 2 the following mathematical programming model is given. Please note that the parameters of the problem are highlighted with blue, and variables are highlighted with green color in Figure 2.

Let binary variable d_i denote whether the i-th driver takes part in the solution or not. Further, let binary variable x_{in} denote whether the i-th driver in the n-th bus turn departed or not. Let binary variable r_{in} denote whether the i-th driver after completion of the n-th bus turn takes its rest or not. Let binary variable c_{in} denote whether the i-th driver after completion of the n-th bus turn finishes his work shift and hands the bus over to another driver, i.e. a driver change activity is performed. Moreover, let variable K_i corresponds to the start, while V_i correspond to the end of the working hours of the i-th driver.

Constraints subject to the labor standards

The labor standards regarding the minimal and maximal length of the working hours have to be considered. Please see equation (1) and (2). It also has to be noted that the duration of the rest cannot be considered as working hours.

$$\forall i = \{1,2,\dots,D\} \quad V_i - K_i \le XWH + \sum_{n=1}^{N} r_{in} * RT \tag{1}$$

$$\forall i = \{1,2,\dots,D\} \quad V_i - K_i \ge NWH - \sum_{n=1}^{N} r_{in} * RT \tag{2}$$

Constraints subject to the start and end of the *i*-th driver's work shift

$$\forall i = \{1,2,\dots,D\}\, \forall n = \{1,2,\dots,\mathrm{N}\}$$
$$V_i \ge P_n + T_n + c_{in} * HT + (1 - c_{in}) * GT + LT - (1 - x_{in}) * M \tag{3}$$

Equation (3) corresponds to the fact that the end of the work shift of the i-th driver has to be later than the arrival time of the given turn plus discharge and entering the garage. This inequality is a dynamic one, it has to be regularly updated, obviously, the final constraint cannot be known in advance.

$$\forall i = \{1,2,\dots,D\} \quad \forall n = \{1,2,\dots,\mathrm{N}\} \quad K_i \le P_n - ST + (1 - x_{in}) * M \tag{4}$$

Equation (4) is similar to equation (3), but from the other direction, i.e. dynamic equation, regularly updated backwards. The launching time of the last bus turn is known. The start of the work shift of the i-th driver is set accordingly to the earliest bus turn performed by the driver.

Constraints subject to the bus turns

Equation (5) corresponds to the fact that the bus turns have to be performed by one of the drivers, while less than one is not sufficient and more drivers are superfluous.

$$\forall n = \{ 1,2,\dots,N\} \quad \sum_{i=1}^{D} x_{in} = 1 \tag{5}$$

Equation (6) corresponds to the fact that the bus driver cannot start a new bus turn before successfully finishing and fully returning from his previous bus turn.

$$\forall i = \{ 1,2,\dots,D\} \;\; \forall n = \{ 1,2,\dots,N-1\} \;\; \forall k = \{ n+1,n+2,\dots,N\}$$
$$x_{ik} * P_k \geq x_{in} * (P_n + T_n + AT) + r_{in} * RT + c_{in} * HT \tag{6}$$

Constraints subject to the rest period of the drivers

The binary variable r_{in} denotes whether the *i*-th driver after completion of the *n*-th bus turn takes its rest or not, namely it is 1 in case there is a rest period and 0 otherwise.

Should the working hours exceed 4 hours, then the driver has to have a rest period, it is specified by equation (7).

$$\forall i = \{ 1,2,\dots,D\} \, \forall n = \{ 1,2,\dots,N\}$$
$$P_n - K_i - 240 * (\textstyle\sum_{n=1}^{N}(r_{iN})\,) \leq 240 + (1 - x_{in}) * M \tag{7}$$

The rest period cannot be within the last working hour of the driver; it is specified by equation (8).

$$\forall i = \{ 1,2,\dots,D\} \quad \forall n = \{ 1,2,\dots,N\} \quad V_i - P_n * r_{in} \geq 60 \tag{8}$$

The rest period cannot start within the first working hour of the driver, it is specified by equation (9), where *M* is a large constant, *M* > 1440, for example 10000.

$$\forall i = \{ 1,2,\dots,D\} \quad \forall n = \{ 1,2,\dots,N\} \quad P_n * r_{in} - K_i \geq 60 + (r_{in} - 1) * M \tag{9}$$

Constraints subject to the resources

Should a driver change activity happen, the original driver hands over the bus to the new driver and finishes his work shift, it is specified by equation (10); while the constraints for the drivers is specified in equation (11).

$$\sum_{i=1}^{D} d_i - \sum_{i=1}^{D} \sum_{n=1}^{N} c_{in} \leq B \tag{10}$$

$$\forall i = \{1,2,\dots,D\} \quad \frac{1}{N} * \sum_{n=1}^{N} x_{in} \leq d_i \leq \sum_{n=1}^{N} x_{in} \tag{11}$$

Cost function

The cost function considered is to minimize all unwanted waiting time of the bus drivers.

$$\sum_{i=1}^{D} \left((V_i - K_i) - \sum_{n=1}^{N} r_{in} * RT \right) \rightarrow min$$

3.4 Example

As part of a daily work to optimally handle bus transport problems, from the traffic point of view a practical example was received from the Budapest Transport Corporation as a public transport service company. The example is based on a real situation and the problem is considered to be typical and of medium difficulty there. For the illustration of the present method only the key parameters of the problem are considered as follows. There is a single depot, the route length is 5.4 km, there are 13 stations along the route, there are 178 bus turns launched within a day, driver change is allowed at the departure station, there are 6 buses and 11 drivers available, rest period can be given only at the departure station. First, the maximal structure of the bus transport process network with arbitrary bus launching times problem was determined based on Figure 2. It is worth mentioning that the bus service company does not consider the situation practical where driver change is permitted in between the departure station and the terminal station practical for this line, therefore the maximal structure depicted on Figure 3 was not considered further. With exploiting the advantages of this structural representation, the corresponding mathematical programming model was generated as detailed in Section 3.3. The size of this mathematical programming model can be handled with relative ease, since based on the previous steps the number of variables in the resultant simplex table is limited. An MPS file was generated automatically, which is a column oriented text format storing linear programming problems and which is supported by a large number of solvers. For the particular case, CPLEX and FICO Xpress were used, as a publicly available NEOS solvers. The solvers were controlled with a time limit.

Please note that the mathematical programming model for the particular problem has 5885 binary variables, plus 22 non negative, real variables; furthermore, the model has 183296 constraints. Should more detailed regulations on the labor standards, work shifts etc. be considered, these appear as additional constraints within the model only, and does not have a significant effect on the scale of the problem. The minimum number of drivers necessary to solve this problem is 9, with less drivers the problem is infeasible. The total waiting time of the drivers is 17 hour and 57 minutes, namely the time when the drivers do not perform any useful activities neither rest; in other words, each driver has approximately 2 hours of unwanted waiting time, which is already a better solution than the current solution applied in the everyday life at the transportation company. The solution data is summarized in Table 2.

Table 2

Solution details of the example

Driver	Start	End	Working hours (minutes)
1	04:58	15:18	620
2	05:18	09:42	264

3	06:47	15:54	547
4	07:17	17:44	627
5	09:25	19:34	609
6	15:17	23:55	518
7	13:49	19:04	315
8	15:37	23:35	478
9	17:17	22:55	338

Conclusions

Herein, bus transport problems, with arbitrary launching times, were discussed. An approach based on the p-graph methodology was proposed to concentrate on the synthesis step of the problem. In contrast to other general purpose solution frameworks, this approach focuses on the pure structure of the system. The key elements of the maximal structure of the problem were detailed. A corresponding mathematical programming model was presented, that suits the particular maximal structure. This model had the advantage of being limited in the number of variables and constraints and therefore, publicly available solvers were capable of generating results and real solutions of industrial problems. A practical example with medium problem difficulty, available at a public transport company, illustrated that the solution method proposed is effective.

Acknowledgement

The authors greatly acknowledge the valuable contribution of Dávid Vizi and Optin Ltd.

References

[1] Bodin L, Golden B, Assad A and Ball M. Routing and Scheduling of Vehicles and Crews: The State of the Art, Computers and Operations Research, 10, 63-211, 1983

[2] Kliewer N, Mellouli T and Suhl L. A time-space network based exact optimization model for multi-depot bus scheduling, European Journal of Operational Research, 175, 1616-1627, 2006

[3] Dávid B and Krész M. Application Oriented Variable Fixing Methods for the Multiple Depot Vehicle Scheduling Problem, Acta Cybernetica, 21(1), 53-73, 2013

[4] Tóth A and Krész M. An efficient solution approach for real-world scheduling problems in urban bus transportation, Central European Journal of Operations Research, 21(1), 75-94, 2013

[5] Horváth M and Kis T. Computing strong lower and upper bounds for the integrated multiple-depot vehicle and crew scheduling problem with branch-and-price, Central European Journal of Operations Research, 27(1), 39-67, 2019

[6] Békési J, Dávid B and Krész M. Integrated Vehicle Scheduling and Vehicle Assignment, Acta Cybernetica, 23(3), 783-800, 2018

[7] Békési J and Nagy A (2020) Combined Vehicle and Driver Scheduling with Fuel Consumption and Parking Constraints: a Case Study. Acta Polytechnica Hungarica, Vol. 17, No. 7, 2020, DOI: 10.12700/APH.17.7.2020.7.3

[8] Friedler F, Tarjan K, Huang Y, Fan LT. Graph-theoretic approach to process synthesis: axioms and theorems. Chem Eng Sci 47(8), 1973-1988, 1992

[9] Friedler F, Varga J, Fan LT. Decision-mapping: a tool for consistent and complete decisions in process synthesis. Chem Eng Sci 50(11), 1755-1768, 1995

[10] Kovacs Z, Ercsey Z, Friedler F and Fan LT. Redundancy in a separation-network. Hungarian Journal of Industry and Chemistry 26(3), 213-219, 1998

[11] Sanmarti E, Puigjaner L, Holczinger T and Friedler F. Combinatorial framework for effective scheduling of multipurpose batch plants. Aiche Journal 48(11), 2557-2570, 2002

[12] Sule Z, Bertok B, Friedler F and Fan LT. Optimal design of supply chains by P-graph framework under uncertainties. Chem Eng 25: 453-458, 2011

[13] Tick J, Imreh C and Kovács Z. Business Process Modeling and the Robust PNS Problem. Acta Polytechnica Hungarica, 10(6), 193-204, 2013

[14] Vincze, N, Ercsey Z, Kovács T, Tick J, and Kovács Z. Process Network Solution of Extended CPM Problems with Alternatives, Acta Polytechnica Hungarica, 13(3), 101-117, 2016

[15] Ercsey Z. Process network solution of a clothing manufacturer's problem. Pollack Periodica 12(1), 59-67, 2017

[16] Tan RR and Aviso KB. An extended P-graph approach to process network synthesis for multi-period operations. Comput Chem Eng 85, 40-42, 2016

[17] Tan, RR, Aviso KB, and Foo, DCY. P-graph and Monte Carlo simulation Approach to planning carbon management networks, Computers & Chemical Engineering, 106, 872-882, 2017

[18] Cabezas H, Argoti A, Friedler F, Mizsey P and Pimentel J. Design and Engineering of Sustainable Process Systems and Supply Chains by the P-Graph Framework. Environmental Progress & Sustainable Energy 37(2) 624-636, 2018

[19] Fan YV, Klemeš JJ, Walmsley TG and Bertók B. Implementing Circular Economy in municipal solid waste treatment system using P-graph. Science of the Total Environment 701, 134652, 2020

[20] Bartos A and Bertok B. Production line balancing by P-graphs. Optim Eng 8(6), 1-18, 2019

[21] Bertók B and Bartos A. Renewable energy storage and distribution scheduling for microgrids by exploiting recent developments in process network synthesis. Journal of Cleaner Production 244, 118520, 2020

[22] König É and Bertók B. Process graph approach for two-stage decision making: Transportation contracts. Computers and Chemical Engineering 121, 1-11, 2019

[23] Bárány B, Bertók B, Kovács Z, Friedler F and Fan LT. Solving vehicle assignment problems by process-network synthesisto minimize cost and environmental impact of transportation. Clean Technol Environ Policy 13(4):637-642, 2011

[24] Nagy A, Ercsey Z, Tick J and Kovács, Z. Bus Transport Process Network Synthesis Acta Polytechnica Hungarica 16(7) 25-43, 2019

Possible Methods for Combining Tongue Contours of Dynamic MRI and Ultrasound Records

Réka Trencsényi[1] and László Czap[2]

[1]Department of Electrical and Electronic Engineering, Institute of Physics, Faculty of Science and Technology, University of Debrecen, Bem tér 18/a, 4026 Debrecen, Hungary, trencsenyi.reka@science.unideb.hu

[2]Institute of Automation and Infocommunication, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Egyetemváros, 3515 Miskolc, Hungary, czap@uni-miskolc.hu

Abstract: One of the trends of the current generation of machine speech, is articulatory speech synthesis, that is based on the processing of visual and geometric information, related to voice production. Accurate knowledge of the static and dynamic geometric parameters of the vocal organs, plays a fundamental role in the realization of speech synthesis. Appropriate sources of visual extraction of these data can be MRI and ultrasound (US) records made during speech, which can be described by different geometries. Harmonization of the geometries of MRI and US frames is not a trivial task. In this publication, we present one possible method for the transformation between the two sources. The starting point of the transformation process is formed by tongue contours obtained by automatic algorithms. Beyond this exact method, we also follow statistical procedures, by applying machine learning to interconnect MRI and US records.

Keywords: articulatory speech synthesis; tongue contour tracking; machine learning; dynamic MRI and US records; harmonization of MRI and US sources

1 Introduction

Speech synthesis is one of the most dynamically developing fields in speech research, with ever more complex technical and methodological challenges, which even today, forms an integral part of the human-machine relationship. In this regard, the communication role of the machine is crucial, since its basic designation is the implementation of text-to-speech transformation, i.e. the realistic imitation of the acoustic product forming during natural human speech. In the extended version of this, the model can be further refined by taking into account the supra-segmental elements of speech (rhythm of speech, voice level, pitch, tone, intonation, stress), which can have high importance in the domain of speech recognition, as well [1]. Currently, research is ongoing, in the field of speech synthesis, with focus on the creation and improvement of text-to-speech systems, that allow the spread of such applications as, e.g. passenger information systems, speaking smart devices, belletristic readers, screen readers, sound weather forecast or telephonic directory enquiry services. In the case of text-to-speech readers representing the traditional trend of researches, speech construction occurs by direct or indirect utilization of human voice samples. The success of these endeavors is proven by numerous publications of the literature [2-7] which report on speech synthesis based on different speech databases or corpuses, in the case of Hungarian, German, or multilingual synthesizers. In addition to the classical concepts, there are also such fields starting to evolve, which are less elaborate and many open problems are still expected to be solved. For instance, articulatory [8-9] or machine-learning-based speech synthesis [10-11] can be classified here.

Articulatory speech synthesis, instead of the application of human voice samples, tries to implement the imitation of the acoustic product by machine imaging of human voice production and articulation. One of the modern technological streams of this is the experimentation trending to the articulatory electromechanical speech generators needed for the production of speech of robots [12] [13]. The starting point of synthesis is the execution of articulatory-acoustic conversion that is built upon visual information relating to speech [14]. Consequently, different imaging procedures (e.g. Magnetic Resonance Imaging (MRI), Computer Tomography (CT), Ultrasound (US)) have essential roles, which supply new information channels in the process of scientific research. Accordingly, MRI or US records made during speech can be potential sources of visually supported extraction of the parameters describing human articulation. Since most actively the tongue takes part in voice production, it is expedient to monitor primarily the motion of the tongue as accurately as possible. In recent years, besides the mentioned MRI, CT, and US, popular tools of the investigations are electropalatography (EPG) or electromagnetic articulography (EMA). Applying the simpler accessible US, EPG, and EMA procedures, information about the dynamic features of speech can be obtained mostly along certain plane sections, although three-dimensional US technique is available, as well, which provides information in multiple planes [15]. Nevertheless, by dint of MRI and CT equipment demanding clinical conditions, three-dimensional morphological data can be acquired. Recently, several studies have dealt with elaboration and development of dynamic tongue contour tracking algorithms [16-18], which can form one of the keystones of research performed in the topic of articulatory speech synthesis. Dynamic scanning of the tongue contour is worth doing in the sagittal plane, where the up-down and forward-backwards motion of the tongue is visible in a two-dimensional section. The most convenient tools of the investigations can be US and MRI records, the advantage being the good spatial and temporal resolution, the ability for synchronization of the image and sound materials, and the protection of the speaker from harmful exposure. Designation of the tongue contour can be done manually or by automatic algorithms, though hundreds or even thousands of frames, creating a given record to justify the preference of dynamic programming against manual operations. The precision of tongue contour fitting is largely determined by the quality of the record and the type of contour tracking algorithm, thus, the ambition for refinement of image processing and tongue contour tracking is still a key task for research.

Beyond this, the application of machine learning algorithms designates an important direction, during which the machine produces output results from the set of certain input parameters, based on information gained from the environment, while it learns and improves performance. Machine learning algorithms try to imitate the behavior of the human brain, so the knowledge and realistic modelling of the operation of neural networks plays a key role. Biological neural networks realize a learning process based on different patterns, which can be mapped by creating appropriate algorithms, in the case of machine learning. In the field of speech synthesis, the set of input parameters of the machine can be formed by, for example, human voice samples or data retrieved from visual sources, which performs the training and the auditory product can be vocalized. Thus, the possibility of neural networks, trained by visual information, offers the linking of methods of articulatory speech synthesis and machine learning in a natural way. Opportunities are actually unlimited, and the procedures and their combinations are mostly, as of yet, not revealed fully.

Our work herein examines the transformational relationships between the geometries of US and MRI frames and the simultaneous application of tongue contour tracking and machine learning algorithms.

2 US and MRI Frames

2.1 Starting Points

Our current research focuses on the simultaneous analysis of US and MRI records made during speech, that can facilitate the visually supported complex retrieval of the static and dynamic parameters that describe human articulation. The MRI records were selected from the free-access multimedia package, on the website of the University of Southern California, the US records were available in the form of audiovisual materials created by the Micro system of the Lingual Articulation Research Group of the Hungarian Academy of Sciences and Eötvös Loránd University [19]. The dynamic moving images can be decomposed into static frames, as a result of that, the subsequent moments of speech generation can be studied step by step. Figure 1 presents a US and an MRI frame which visualize tongue positions corresponding to sound *k* arising from a female and a male speaker.

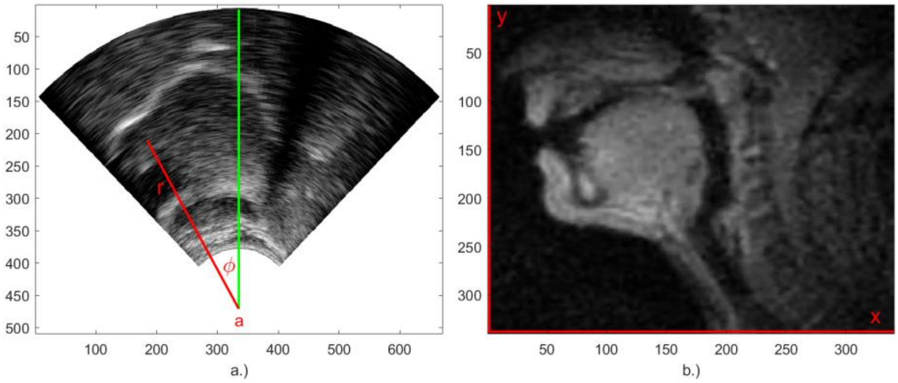


Figure 1

The side-view position of the tongue in a US (a) and an MRI (b) frame described by radial (a) and Cartesian coordinates (b)

The records display the region of the oral cavity in the sagittal plane dividing the human body into left- and right-hand parts, so in a two-dimensional section, the up-down and back-forth motions of the tongue become seeable. In the US record, the tongue contour appears as a bright band that is produced by the US waves reflected at the boundary of the tongue and the air above and the line of the edge of the tongue can be traced at the lower border of the bright band. Since the hyoid bone and the mandible partially shield the US waves, the US transducer is not capable of probing the region of the oral cavity entirely. This deficiency shows up in the form of a dark band emerging on the left and right sides of the image, at the front and rear parts of the tongue that hides the movement of the tongue root and the tongue tip, thus, in contrast to MRI records displaying the total region of the oral cavity, only partial information can be obtained about the shape and movement of the tongue. The fact can be noted as a further difference that the contour of the palate cannot be identified in the US frames, while in the MRI frames, the contour of the hard palate can be determined with sufficient accuracy, and also, the movement of the soft palate can be detected. In Figure 1, it can be observed that the US frames are spread in a zone covered by a sector of a circle, so the two-dimensional polar coordinates can be conveniently applied by the description of the position, of each pixel. These coordinates can be defined by radius $\mathbf{r}$ measured from center **a** of the circle and angle $\boldsymbol{\phi}$ relative to the vertical symmetry axis of the image. Thereby, the location of a pixel, taken in the plane of the frame, is determined by the pair of coordinates ($\mathbf{r}$,$\boldsymbol{\phi}$) unambiguously. In the case of the used US frames, the value of angle $\boldsymbol{\phi}$ can change between -45° and 45°. However, the most comfortable frame of reference needed for the treatment of MRI frames can be given by a two-dimensional Cartesian coordinate system, in which, the position of the designated point of the frame is fixed by the pair of coordinates (**x**, **y**). One of the aims of research is to harmonize the radial and rectangular arrangements of US and MRI records, by finding the appropriate geometric transformations.

Geometric transformations can be realized through the conversion of the relevant anatomic contours of US and MRI records. These curves can be obviously given by the tongue and palate contours, since in the dynamic description of articulation, the change of relative positions of the surface of the tongue and the palate plays an essential role in the region of the oral cavity. Hence, these examinations provide the most accurate data concerning tongue and palate contour required.

For determination of the contour of the edge of the tongue we developed and improved automatic tongue contour tracking algorithms, based on dynamic programming. The primary aim of tongue contour tracking, is the dynamic description of tongue positions, belonging to different speech sounds, and the investigation of tongue movements characterizing sound transitions created during co-articulation. Besides the qualitative analysis, the tongue contour can also be a good starting point for the quantitative study of speech, since the numeric values derived from tongue contour, can support the deeper understanding and development of articulatory models. Algorithms elaborated for detection of the tongue contour can be extremely diverse depending on the applied procedures. The edge of the tongue is drawn as a bright band in US records, while in MRI records it can be experienced as a contrast coming into existence between the dark domain of the air in the oral cavity and the bright domain of the tongue tissue, so contour tracking means the search for the pixels at the boundary of the dark and bright domains, determining the line of the edge of the tongue, in both cases. Using our approach, the application of our algorithm is preceded by the preprocessing of records, that tends to cancel the noise and discontinuities, resulting from imaging techniques. The most effective instruments of reducing the mentioned errors are edge-enhancement and averaging operations, that mathematically can be realized by convolution [20]. The found pixels of maximal brightness, adjusting to the uneven line of the edge of the tongue, produce a rough curve, the smoothing of that can be solved by a discrete cosine transformation. The images of Figure 2 show automatically fitted tongue contours in an MRI (a) and (b) and US (c) and (d) frames, respectively. In Figure 2a, the tongue position belonging to sound ***o*** can be observed, while Figure 2c renders the tongue position corresponding to sound *ɔ* by highlighting the smoothed tongue contour. In Figures 2b and 2d, the magnified details of the unsmoothed tongue contours drawn in frames 2a and 2c, can be seen.

Figures 2b and 2d can be created by a special transformation starting from Figures 2a and 2c.

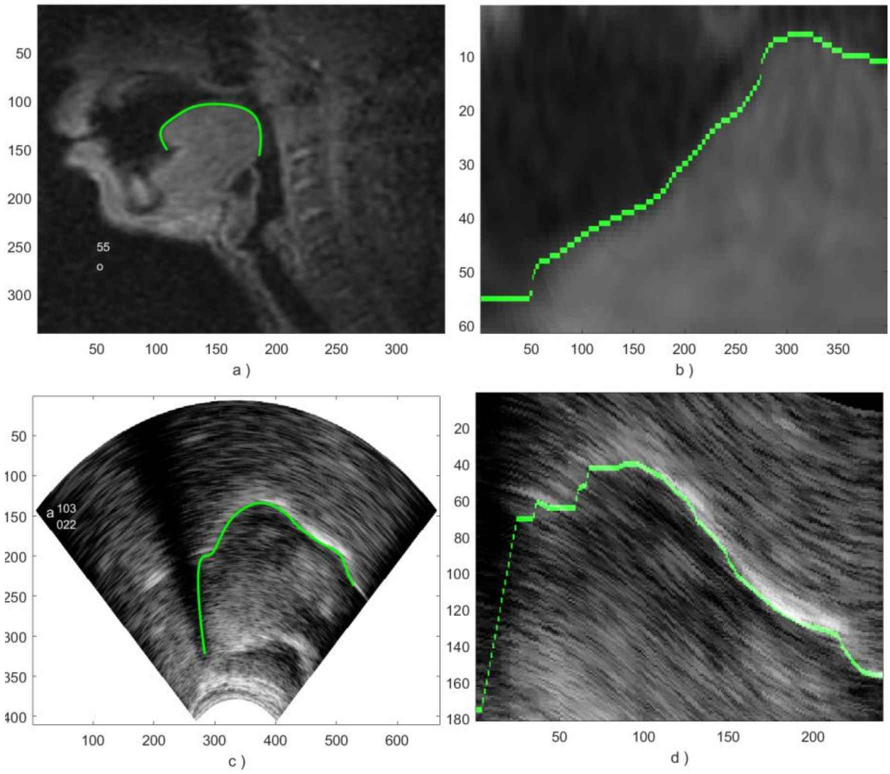


Figure 2

Automatically fitted tongue contours in MRI (a, b) and US (c, d) frames, showing also the magnified details of the unsmoothed tongue contours (b, d)

The substance of the transformation procedure is illustrated by dint of the US frame seen in Figure 3. As a first step, in Figure 3a of radial geometry, originating from the center of the circle, radial sections are formed in the range -45° and +45° defined by the record. Along these sections, the image is practically resampled. The sections produced in this manner are arranged in columns, resulting in such an image matrix, that most conveniently can be described in the Cartesian **x-y** plane. Figure 3b is generated on the track of shaping the matrix structure. Investigations show that sampling performed by 1/4° is the ideal, since this time, a change in the contour, greater than two pixels, does not occur between adjacent columns of the matrix. For the sake of clarity, the sections are depicted only by 5° that are demonstrated by the white lines in Figure 3. The procedure works in the case of MRI frames in a similar way, by applying the center and angular domain (usually wider than the range -45° – 45°) designated in the MRI frame properly.

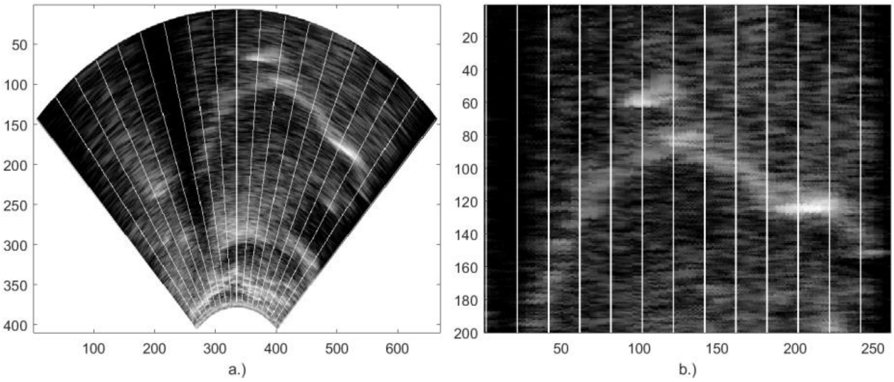


Figure 3
Some radial sections (drawn by white lines) of US frames in the original (a) and in the matrix-structured transformed plane (b)

The speaker for the MRI records is a native American English male speaker (John Esling), who vocalizes series of sounds of type VCV with vowel V and consonant C. In the US records, series of sounds arising from native Hungarian and Chinese female speakers are recorded, which are of CVCV, CVC, and VCV structures. All of these speakers are young adults, and the Chinese participant speaks in a Shaanxi Xi'an dialect. According to the presented frames of Figure 2, the obtained curves follow the line of the edge of the tongue authentically.

2.2 Geometric Transformations

Due to the screening effect of the hyoid bone and the mandible, US images are able to visualize the movement of the tongue only partially, that leads to a more confined data set regarding the position of the tongue compared to MRI frames. Since the production of a more extended parameter set from a narrower one, is much more challenging than the reverse, we specified the contours of US records, as the base of transformations.

As mentioned above, in addition to the tongue contour, the curve fitted to the palate plays a key role in the examination of articulation. Hence, before implementation of the transformation, the palate contour is needed to be ascertained in US frames. It is not a trivial task, because it cannot be revealed immediately in the US records. The location of palate, however, can be given via estimation by presuming the boundary of the tongue and palate by selecting the points being in the highest positions, and touched by the surface of the tongue during articulation. This requires, of course, the investigation of such consonants during the utterance of that the tongue surely touches the hard or soft palate. This condition is fulfilled automatically in the case of the available US package containing various audio items, since, during the articulation of consonants being present in the recorded sentences, the tongue comes into contact with the palate at different places. We implemented the drawing of the contour of the palate essentially by the solution of an extremum search problem, the result of that is presented by the red curve of Figure 4, and the tongue contour belonging to the frame is demonstrated by the green curve.

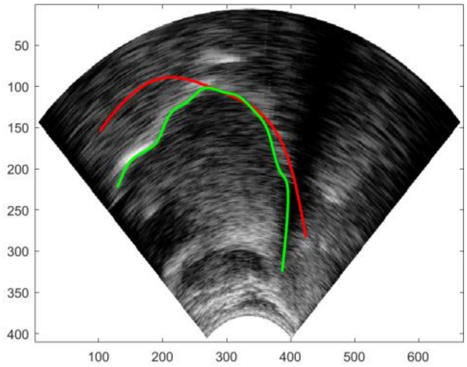


Figure 4

Tongue contour (green) fitted to the surface of the tongue and palate contour (red) arising from an extremum search problem in the case of US frames

For the transformation of the curves of Figure 4, we searched for such a reference point that can be identified with convincing certainty in the US and MRI frames, as well. We defined this point at the peak of the epiglottis, the position of that is marked by the red circle drawn in the images of Figure 5.

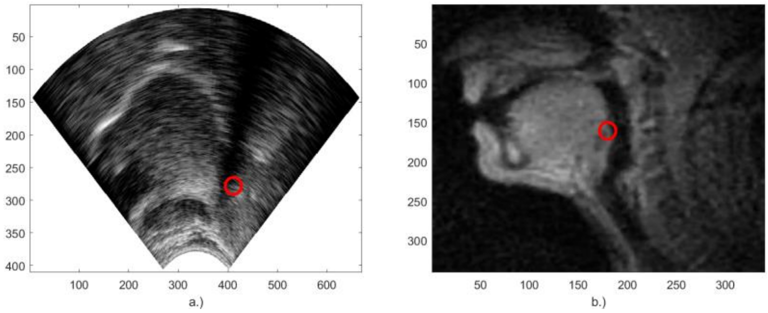


Figure 5

The peak of the epiglottis localized by red circles in the US (a.) and MRI (b.) frames

In the course of the visual study of dynamic US records we concluded that, during articulation of certain sounds, the tongue is pulled back insomuch that it touches the epiglottis. Hence, we designated the peak of the epiglottis as the starting point of the palate contour, and we determined the angular range covered by the tongue contour belonging to the selected sound *k*, which is limited by the values -39.6° and 19.4°. We performed the transformation of the curves of the tongue and palate contour in the polar coordinate system by scaling the radial and angular range given by the pairs of values ($\mathbf{r}$, $\boldsymbol{\phi}$) describing the points of the curves, and by shifting the initial angle $\boldsymbol{\phi_O}$ of the angular range according to the formulas

$$r' = Rr$$

$$\varphi' = FI\,\varphi$$

$$\varphi_0' = \varphi_0 + FIKORR \qquad (1)$$

The scale factors R and FI of relationships (1) enable the normalization of the radial and angular range, and the term FIKORR is responsible for the rotation of the angular range. By fixing the values R=0.31, FI=1, $FIKORR$=12.6°, it is allowed to transplant the tongue and palate contours to the MRI frame. According to Figure 6, the tongue and palate contour fit to the MRI frame in an acceptable way, where the angular range of the tongue contour extends between the values -27° and 32°. Ultimately, the radial geometry of US frames is embedded into the rectangular geometry of MRI frames by the transformations (1) executed in the system of polar coordinates.

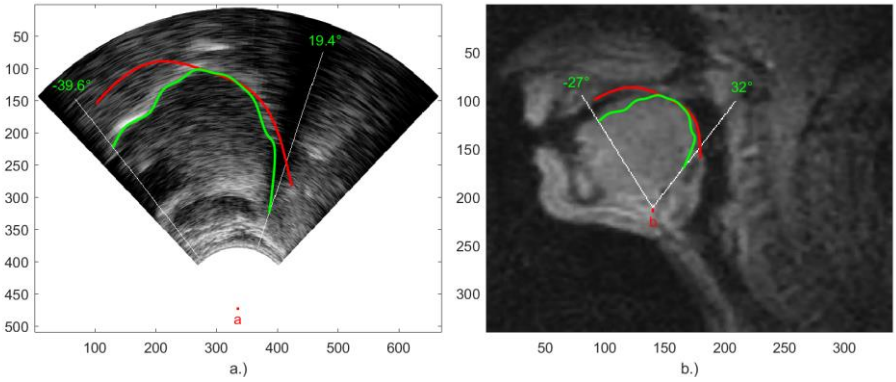


Figure 6

The angular range of the fitted (a) and transformed (b) US tongue contours drawn in the US (a) and MRI (b) frames, where the US palate contour is presented by red curves

By means of the transformation, the biunique correspondence of the points of tongue contours fitted to the US and MRI frames becomes possible that can be traced by dint of Figure 7. Due to the factor FI=1, the transformation is isogonal, therefore, the four contour points specified by the four inner radial sections selected in the US frame can be mapped along the same four radial sections drawn in the MRI frame to the MRI tongue contour illustrated by the blue curve. Thus, passing along a given section, two points can be found on the green and blue curves that can be assigned in pairs unambiguously.

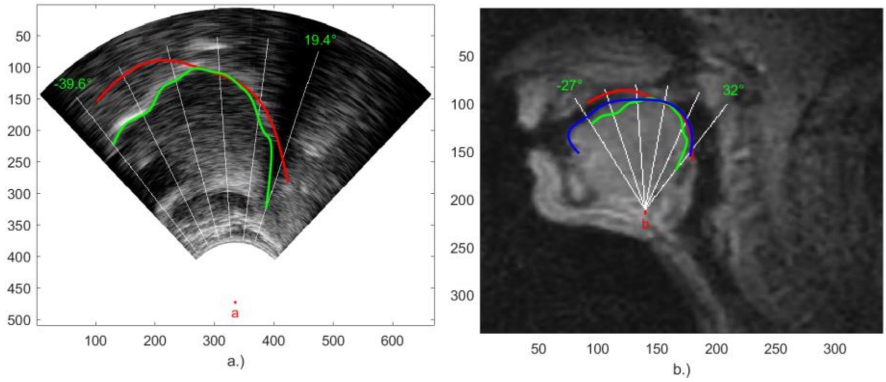


Figure 7

The biunique correspondence of the points of the transformed US tongue contour (green) and the fitted MRI tongue contour (blue) along the designated radial sections drawn by white lines in the US (a) and MRI (b) frames

The transformation can be realized in the reverse direction, as well, which means that a contour of an MRI frame can be projected to a US frame. For this purpose, the inverse transformations of (1) should be applied in the form of

$$r = r'/R$$

$$\varphi = \varphi'/FI$$

$$\varphi = \varphi_0' - FIKORR \qquad (2)$$

Using the transformations of (2), the tongue and palate contour of an MRI frame can be transferred to the appropriate US frame, as it is exemplified by Figure 8, where the tongue contour of Figure 7b is projected. The transformed curves accurately demonstrate those sections of the tongue and hard palate which do not appear in the US record because of the screening effect of the mandible.

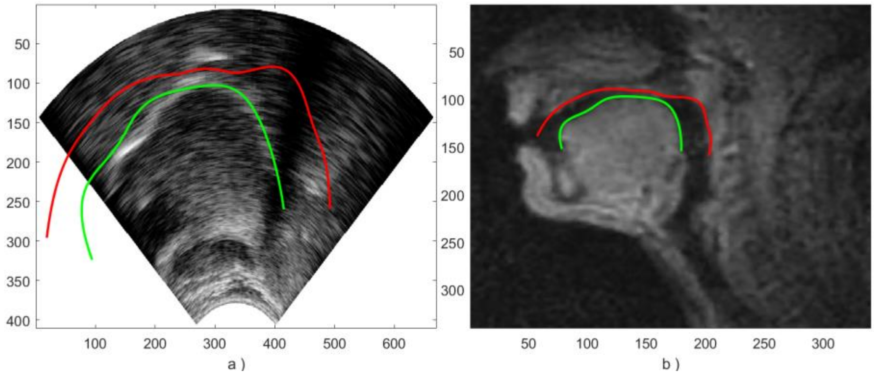


Figure 8

The fitted (b.) and transformed (a.) MRI tongue contours drawn in the MRI (b.) and US (a.) frames, where the MRI palate contour is presented by red curves

It is important to emphasize that the transformation of (1) and (2) containing exact steps cannot be applied uniformly in the case of all speech sounds, since the parameter set describing the transformation can change sound by sound. This circumstance makes the compact combination of US and MRI records more difficult, but this problem can be resolved by the optimization of the parameters of the transformation, extending to several speech sounds, or by involving machine learning algorithms belonging to the forefront of statistical methods. In the next chapter, the application possibilities of machine learning are presented.

3 Machine Learning

We created our programs in the MATLAB environment, and we implemented machine learning by such an algorithm that determines the weight factors of the neural network, by the scaled conjugate gradient method [21]. Knowing the input parameters, this optimization procedure solves the system of equations assigned to the problem by an iterative method, while the output parameters calculated by the procedure converge to the prescribed values. The advantage of the method is the fast convergence that can be ensured by minimizing the number of steps of the iterative algorithm, thus, machine learning training can be carried out in a relatively short time. The iterative steps are realized along such a direction that enables faster convergence than the most negative gradient corresponding to the steepest descent, while it preserves the error minimization obtained in the previous steps. Training stops when the maximum number of epochs is reached, or the maximum amount of time is exceeded, or performance is minimized to the goal, or the performance gradient falls below the minimum performance gradient, or validation performance has increased more than maximum validation failures times since the last time it decreased.

We placed two hidden layers in the neural network, which individually contained 30 neurons. We designated the input parameters needed for learning by dint of four chosen points of the dynamically changing tongue contour, to that we assigned the discrete cosine transform of the tongue contour in the output side of the system. The four feature points coincide with those four points that are determined by the four inner radial sections of the angular range, as shown in Figure 7b. As illustrated by Figure 9, the feature points of the US and MRI tongue contours are stamped by magenta and yellow markers, respectively, together with the ordinal numbers of the given points along the green and blue curves. In this manner, the feature points of the US and MRI tongue contours correspond to each other pairwise, unambiguously, along a given radial section. It can be seen that the magenta markers follow each other in reverse order compared to the yellow markers. This effect is caused by the vertical reflection of the US tongue contour when embedding it into the MRI frame. The relative positions of the four feature points are identical in each frame, in the sense that the four points can be found at about 20%, 40%, 60%, 80% of the angular range [-27°, 32°], in the case of all tongue contours. So the feature points are fixed automatically in all frames.

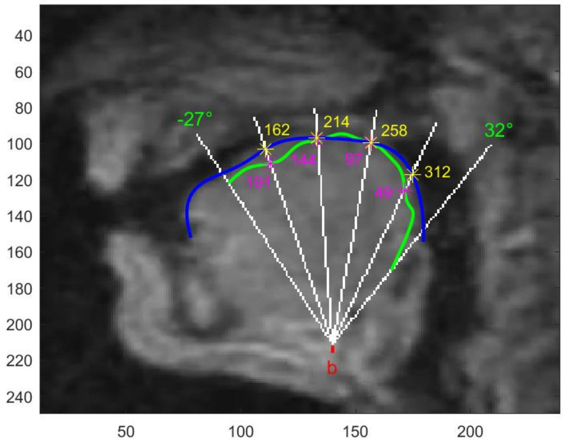


Figure 9

The feature points of the US (green) and MRI (blue) tongue contours stamped by magenta and yellow markers along the selected radial sections indicated by white lines

We executed the learning, first by fixing the input and output parameters arising from the MRI source, and then we tested the results in the same MRI frames. Based on a similar principle, we repeated the procedure for the US frames. Finally, combining the input parameters gained from the US source with the output parameters originating from the MRI source, we ran the algorithm again, then we tested the results in the MRI frames. The following sub-sections discuss the three different approaches.

3.1 MRI-MRI Learning

The subsection summarizes the results of machine learning accomplished in the case of the MRI records. The base of learning is formed by the phonemic configurations belonging to the speech sounds *ɔ, aː, ʦ, ʧ, d, ʥ, ʤ, ɛ, eː, g, ɟ, i, j, k, l, n, ɲ, o, ø, r, ʃ, s, t, c, u, y, z, ʒ*. The input parameters are given by the **y** coordinates of the four selected points of the tongue contour, measured in the plane of the image, while the set of output parameters is determined by the first twenty coefficients of the discrete cosine transform of the tongue contour. After running the learning algorithm, the trained tongue contour can be reconstructed by inverse discrete cosine transform. It practically means that the production of the complete curve occurs by using just four points. Our results are demonstrated through the example of sounds *j* and *t*.

Figures 10a and 10c present tongue contours fitted to the tongue positions corresponding to sounds *j* and *t*. Figures 10b and 10d display trained tongue contours belonging to the same sounds *j* and *t*. When comparing the fitted and trained tongue contours, no significant visual distinction shows up, the difference is minimal between the two curves, which can be determined also quantitatively for example by the values of the Mean Absolute Difference (MAD), Root Mean Squared Distance (RMSD), Mean Sum of Distances (MSD), or Nearest Neighbor Distance (NND).

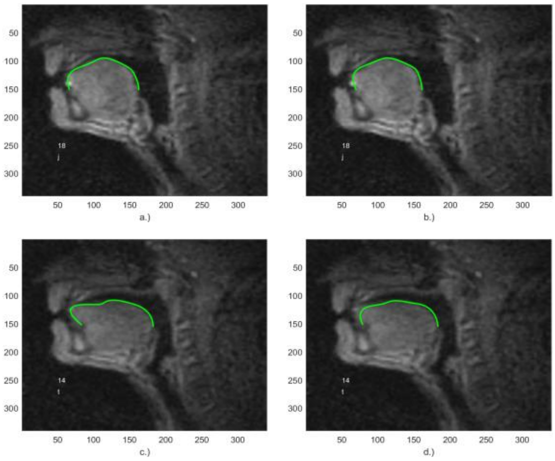


Figure 10

Fitted (a, c) and trained (b, d) MRI tongue contours in the case of sounds *j* (a, b) and *t* (c, d)

The results illustrated in Figure 10 reflect that the learning algorithm works effectively, confirmed by as well by the graphs of Figure 11 and showing the mean squared error of training, testing and validation. It can be seen that, besides rapid decrease, the errors of learning and testing are essentially identical.

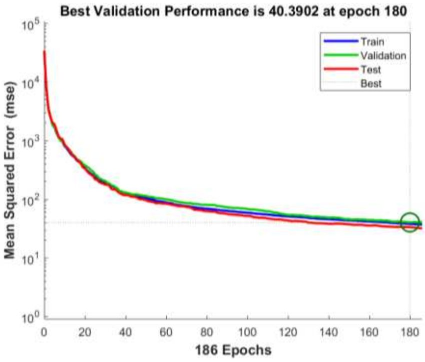


Figure 11

The mean squared error of training, testing, and validation in the case of MRI-MRI learning

3.2 US-US Learning

The subsection summarizes the results of machine learning performed in the case of the US records. In this case, learning is built upon utterances of CVCV type. The interpretation of the input and output parameters is the same as in the previous subsection, and at this time, the steps are led through the example of sounds *g* and *ʃ*.

Figures 12a and 12c demonstrate tongue contours fitted to the tongue positions corresponding to sounds *g* and *ʃ*. Figures 12b and 12d depict trained tongue contours belonging to the same sounds *g* and *ʃ*. Comparing the fitted and trained tongue contours, no considerable distinction can be observed between the two curves.

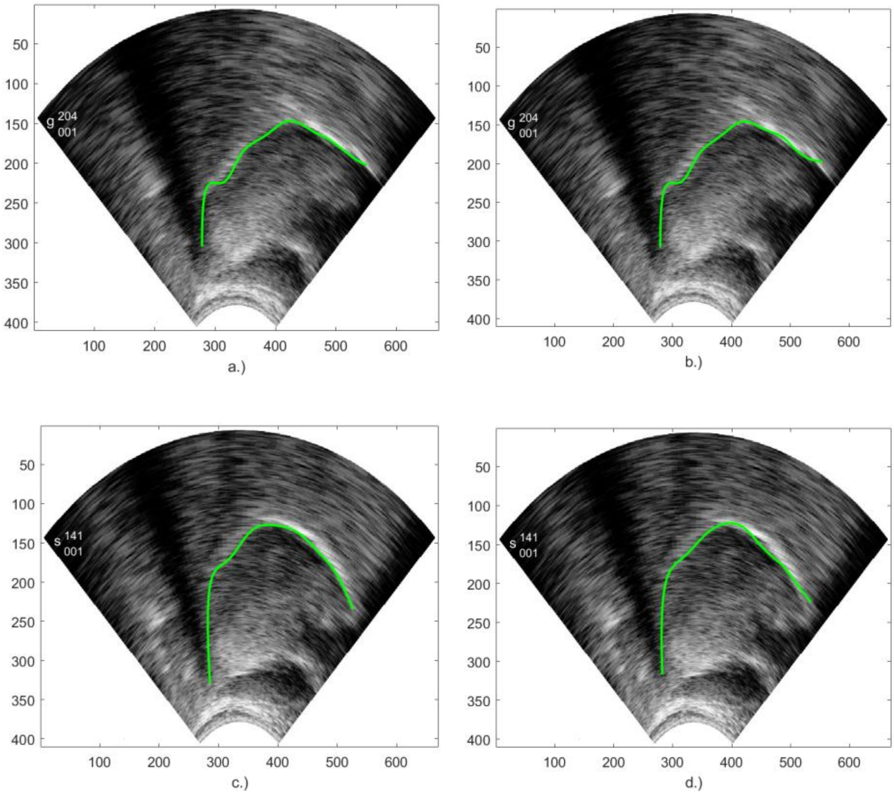


Figure 12

Fitted (a, c) and trained (b, d) US tongue contours in the case of sounds *g* (a, b) and *ʃ* (c, d)

Figure 13 illustrates the formation of the mean squared error of training, testing, and validation, the tendency of that is similar to the curves obtained during learning implemented by the MRI records.

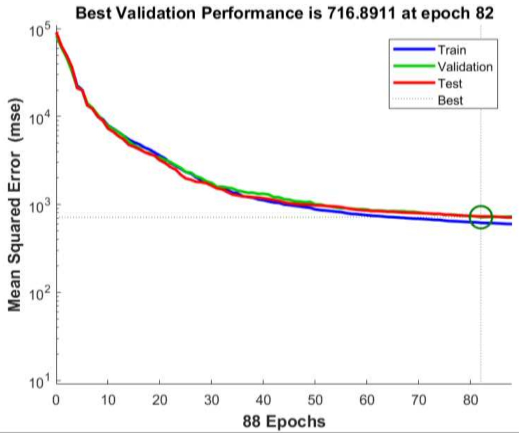


Figure 13
The mean squared error of training, testing, and validation in the case of US-US learning

3.3 US-MRI Learning

In the previous two subsections, the input and output parameters of machine learning originated from the same source, since MRI tongue contour was trained by MRI data, and US tongue contour was trained by US data. It is also worth examining how successful the parameters of the two different sources can be connected. It is quite challenging because the tongue contours of a female and male speaker of different anatomies need to be harmonized. Based on our expectations, however, even the parameters of the transformation carry quantitative information about the differences of the anatomies of the two speakers. For this purpose, we constructed the neural network such a way that its input parameters are created by the four selected points of the US tongue contour, and its output parameters are generated by the discrete cosine transform of the MRI tongue contour. Thereby, such a learning mechanism can be established in which MRI tongue contour can be produced by the utilization of US data. We note that the size of the used database lags behind the cases discussed in the previous two subsections by orders of magnitude. The reason for this is that the MRI and US records hold, not the same utterances, in all cases, furthermore, the number of frames assigned to the individual speech sounds does not match and that makes the harmonization of the parameters for the learning algorithm more difficult. Synchronization of the utterances and number of samples, however, is also currently in progress.

Figure 14a exemplifies the tongue contour fitted to the tongue position corresponding to sound *k*. Figure 14b presents the trained tongue contour belonging to the same sound *k*. The result can be interesting even from several viewpoints, since beyond the fact that the input and output parameters connected by the neural network arise from records of utterers of various native language and different gender made by different imaging techniques, neither the condition can be neglected that learning produces a wider data set starting from a narrower one. Namely, as mentioned earlier, US records are not able to display the rear part of the tongue and the region of the tongue tip that is visible in MRI records without any obstacles. This predicts that, using the partial data originating from US records and involving learning algorithms, the contour of the complete edge of the tongue can be estimated effectively.

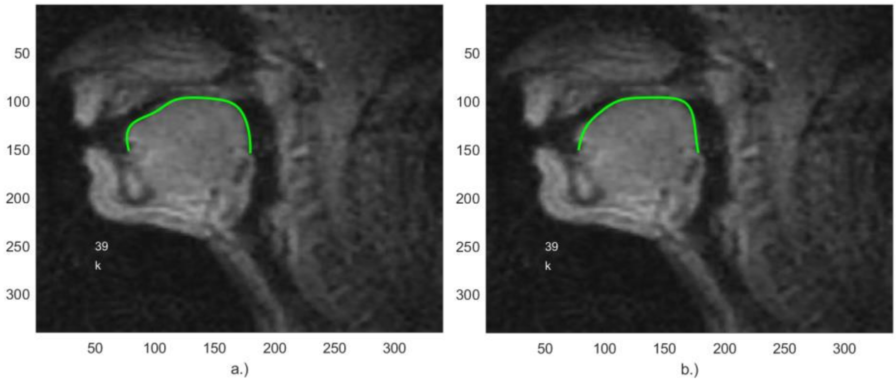


Figure 14

Fitted (a) and trained (b) MRI tongue contours in the case of sound *k*

Conclusions

The main goal of this work was the development and refinement of methods that can be applied towards articulatory speech synthesis. The tools of investigation are constituted by dynamic MRI and US records. The examinations basically run along two threads that approach the problem of harmonization of the relevant anatomic contours of MRI and US frames from different viewpoints. At the starting level, geometric transformations are performed, which interconnect the tongue and palate contours of the MRI and US frames in a bi-unique way. Although this procedure is based on exact mathematical considerations – according to the present stage of research work – it cannot be applied for all speech sounds, in a uniform manner, because the parameter set of the transformation does not contain the same values for each speech sound. So this solution seems to be quite tedious. In pursuance of our future plans, it will be resolved by the optimization of the parameters of the transformation, to produce satisfactory matching of MRI and US contours. Therefore, by way of statistical methods, machine learning is involved in the study, the application is associated with our automatic tongue contour tracking algorithms. Machine learning is implemented in respect of MRI-MRI, US-US, and US-MRI sources by the appropriate combining of the input and output parameters of the neural network. Currently, only a limited number of training and testing configurations are available, but the source data are being gradually expanded. The actual results exhibit only a narrow slice of the ongoing research work, since the fields of articulatory speech synthesis and machine learning, raise, in themselves, a large number of problems, that can be regarded as temporarily partially solved. Accordingly, the future trends of research can be determined by the perfection of the models of speech synthesis created by statistical or rule-based algorithms and built on visual information. It has a potential fundamental importance, for example, in speech therapy for clinical purposes, in the shaping of non-native language learning trainings or in the construction and development of the synthesizers needed for vocalizing silent speech.

Acknowledgement

We would like to thank the MTA-ELTE Lendület Lingual Articulation Research Group, for providing the recordings with the Micro system.

References

[1] Czap, L., Pintér, J. M.: Intensity feature for speech stress detection. Proceedings of the 16th International Carpathian Control Conference Miskolc, Hungary: IEEE IAS/IES/PELS, 2015, 91-94

[2] Olaszy, G.: Making Speech Database for Machine Speech Production. (Beszédadatbázisok készítése gépi beszédelőállításhoz) Beszédkutatás99, 1999, 68-89

[3] Olaszy, G., Németh, G., Olaszi, P., Kiss, G.: Profivox: the Most Modern Native Speech Synthesiser (Profivox: a legkorszerűbb hazai beszédszintetizátor) Beszédkutatás 2000, 2000, 167-179

[4] Németh, G., Olaszy, G., Fék, M.: Development and Experimental Results of a Novel Corpus-based Machine Text-to-Speech System (Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei) Beszédkutatás, 2006, 183-196

[5] Sproat, R. W.: Multilingual text-to-speech synthesis, KLUWER Academic Publishers, 1997

[6] Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. Int. J. Speech Tech., 6, 2003, 365-377

[7] Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Comm., 56, 2014, 85-100

[8] Zappi, V., Vasuvedan, A., Allen, A., Raghuvanshi, N., Fels, S.: Towards real-time two-dimensional wave propagation for articulatory speech synthesis. Proceedings of Meetings on Acoustics 171ASA, 26, 2016, 045005

[9] Czap, L., Pintér, J. M., Baksa-Varga, E.: Features and Results of a Speech Improvement Experiment on Hard of Hearing Children. Speech Comm., 106, 2019, 7-20

[10] Wu, Z., Valentini-Botinhao, C., Watts, O., King, S.: Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2015, 4460-4464

[11] Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Andrew, N., Raiman, J., Sengupta, S., Mohammad, S.: Deep voice: Real-time neural text-to-speech. Proceedings of the 34th International Conference on Machine Learning, 70, 2017, 195-204

[12] Roehling, S., MacDonald, B., Watson, C.: Proceedings of the Australasian International Conference on Speech Science and Technology, 2006, 130-135

[13] Li, X., MacDonald B., Watson, C. I.: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, 5009-5014

[14] Czap, L., Mátyás, J.: Virtual speaker. Proceedings of 6th International Carpathian Control Conference ICCC 2005 Miskolc, Hungary: University of Miskolc, 2005, 351-358

[15] Lulich, S. M., Berkson, K. H., de Jong, K.: Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion. Journal of phonetics, 71, 2018, 410-424

[16] Li, M., Kambhamettu, C., Stone, M.: Automatic contour tracking in ultrasound images. Clinical linguistics and phonetics, 19, 2005, 545-554

[17] Csapó, T. G., Deme, A., Gráczi, T. E., Markó, A., Varjasi, G.: Synchronised Speech and Tongue Ultrasound Records by the Sono-Speech System. 13th Conference on Hungarian Computational Linguistics (Szinkronizált beszéd- és nyelvultrahang-felvételek a Sono-Speech rendszerrel) University of Szeged, Institute of Informatics, Szeged, 2017, 339-346

[18] Zhao, L., Czap, L.: Automatic tracking of tongue contours in ultrasound records (A nyelvkontúr automatikus követése ultrahangos felvételeken) Beszédkutatás, 27, 2019, 331-343

[19] Csapó, T. G., Grósz, T., Gosztolya, G., Tóth, L., Markó, A.: DNN-based Ultrasound-to-Speech Conversion for a Silent Speech Interface, Interspeech 2017, Stockholm, Sweden, 2017, 3672-3676

[20] Czap, L., Image processing (Képfeldolgozás), Miskolc-Egyetemváros, Hungary: Miskolci Egyetem, 2007

[21] Moller, M. F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural networks, 6, 1993, 525-533

Proactive Maintenance Model Based on the Law on Change of Mechanical Vibration

Goran Otić[1], Goran Jovanov[2], Živoslav Adamović[3], Nemanja Jovanov[1], Stevo Jaćimovski[2]

[1]Faculty of Business and Law, "Union - Nikola Tesla" University, Knez Mihajlova 33, Belgrade 11040, Serbia; goran.otic@ppf.edu.rs, nemanja.jovanov@ppf.edu.rs

[2]University of Criminal Investigation and Polices, Cara Dušana 196, Belgrade 11080, Serbia; goran.jovanov@kpu.edu.rs, stevo.jacimovski@kpu.edu.rs

[3]University „Union-Nikola Tesla“, Faculty of Applied Sciences, Dušana Popovića 22a, Niš 18000, Serbia, zivoslav.adamovic@fpn.rs

Abstract: The basis for proactive maintenance in thermal power plants, is the analysis of the root cause of the failure, i.e. the determination of mechanism and cause of failure occurrence from the thermal power plant system. The root causes of system failures can be eliminated in this way, and the causes of failures can gradually be eliminated using an engineering approach from any assembly of device or machine. Successful proactive maintenance programs would gradually, over time, eliminate problems of the device by project-engineering solutions, which, as a consequence, would have a significantly longer device life cycle, reduced downtime and increase production capacity.

Keywords: proactive maintenance; life cycle management; life cycle costs; vibration control; reliability

1 Introduction

Maintenance management has been accepted as a serious issue, only after the dynamic industrial revolution during the World War II [16]. Rapid modernization and the increasing need for high yielding productivity have led to finer development and use of hi-tech and complex machines and equipment. Therefore, high cost capital is involved in the shop floor production and the occurrence of emerging and frequent failures may result in production downtime and huge losses for a company. Hansen I. H, in his paper, referred that maintenance cost can be the second largest component of a company budget, together with just energy costs [6]. So, controlled and appropriate maintenance activities are required in order to minimize the occurrence of such failures and increase the reliability of the company assets through effective plant maintenance practices. Today, 80% of the parameters measured are likely to be vibration based [16]. Hence, vibration monitoring and analysis are, often, the most widely used in condition based maintenance and greatly rely on instrumentation. Machine vibrations provide a lot of information concerning the condition of a machine. The measurement and analysis of the vibration response provide great deal of information relevant for defect conditions in different types of machines [13]. Vibration-based analysis techniques can be widely used for condition based maintenance because vibration spectrum can be collected for all machinery which consists of rotating or moving elements [11] [15]. Vibration analysis is one, among a number of techniques, in condition based maintenance implemented, in order to monitor and analyze certain machines, equipment and systems in a plant. Nevertheless, the primary concept behind the vibration analysis application is to monitor rotating machinery, to detect emerging problems and to eliminate the possibility of catastrophic failure. The maintenance is initiated when indicators show the sign of defects in the initial stages. In simple words, the main criterion is to maintain the right equipment at the right time. The practice of CBM (Condition Based Maintenance) is implemented by collecting and analyzing the real time data, so that maintenance activities and resources can be prioritized/ optimized accordingly [10]. Basic advantages of proactive maintenance lie, mainly, in reducing maintenance costs (direct and indirect) and increasing the efficiency (reliability and availability) of technical systems.

Therefore, the primary goal of proactive maintenance is to increase efficiency (certainty and reliability), i.e. reduction of downtime, which results in high utilization rates and thus also the productivity.

2 Model of Predictive Vibro-diagnostic Maintenance

Predictive Maintenance (PdM) is a proactive maintenance approach that emphasizes the forecast of how and when equipment will fail through data analytics, and performs maintenance precisely, before total failure occurs. This is achieved through the detection of possible failures by monitoring and analyzing various equipment operation variables, by using the assortment of diagnostic sensors and other monitoring instruments. For example, monitoring equipment changes regarding: vibration, temperature, pressure or voltage, to mention a few of these. The outcome then, is that maintenance will only be scheduled when a failure has been detected, rather than when equipment is perceived to require maintenance. Preventive Maintenance (PM), is the maintenance philosophy of performing maintenance tasks at predetermined intervals, using triggers. These triggers can be derived by a specific amount of calendar days or when a tool has elapsed a defined period of runtime [5]. Microsoft has published a predictive maintenance dataset in the past, which is designed to be used within their Azure platform as a learning device [9]. This dataset satisfies the need to be relevant and usable, for the development of and training for predictive maintenance model, as it has been created with that purpose in mind. However, caution should be taken when directly comparing the results of this project with other predictive maintenance solutions, as although this 'simulated' dataset provided by Microsoft is publicly available, there is no credibility that this dataset will represent realistic data values. Nonetheless, it is a valid dataset for the development of a predictive maintenance algorithm. Reliability-Centered Maintenance (RCM) is a hybrid maintenance philosophy, combining the use of PdM and Run-to-failure maintenance (RTF). RCM acknowledges that not all equipment has the same level of importance and takes a systematic approach to identify the right strategy, for the right need. Despite this, the predominant strategy is PdM, which is applied to the most critical systems of operation, while RTF is utilized for the least. [13].With supervised learning as the selected learning model, the attention must now be paid to the available supervised learning algorithms that will be used to map the predictive maintenance input and output. But first, supervised learning algorithms are prone to two specific modelling errors called 'Over fitting' or 'Under fitting' a predictive model. Furthermore, supervised learning algorithms also suffer a modelling problem called the 'Bias-variance trade-off'. These two specific circumstances need to be discussed, due to the fact that either of these modelling errors occur during the development of the predictive maintenance model [13]. Prior to the completion of the algorithmic procedure, it is necessary to check whether the solution variant meets the set criteria and limitations. If these are not satisfied, they must be rejected. This is particularly problematic at the failures belonging to the categories in which the costs represent project limitation, as it may be required to accept necessary solution that carries within itself higher risk for failure appearance. The completion of a procedure defined by general algorithm of preventive installation represents the end of analysis of all predicted variant solutions. It is then when it should be transferred to the decision on specific variant solution. Defined goal function in our specific example is the selection of optimal vibro-diagnostic formats set which provides the maximum possibility of dynamic problem detection, selection and verification, together with economic justification and simplicity of use. However, it is possible to conclude that the existence of dynamic problems is a characteristic for certain types of rotating machines, which could be categorized in particular groups depending on nominal power, speed, type of foundation, etc. The proposed solution advocates for soft sensor based algorithms. The soft sensor algorithms provide information about the physical status of the components, as well as information about the performance of the systems. These algorithms take advantage of existing or available internal signals of the systems. The objective is to estimate inaccessible states and parameters of the systems using as few physical sensors as possible to acquire the necessary signals to work with.

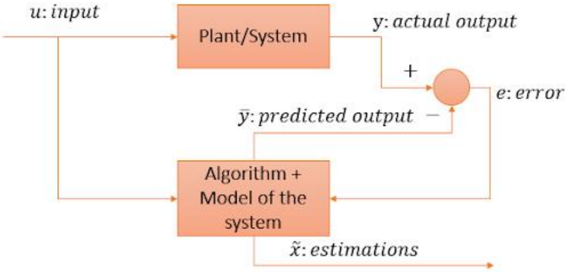


Figure 1
Soft sensor algorithm basic diagram

Many strategies for transitioning to condition-based maintenance leverage the Industrial Internet of Things (IIoT). However, organizations should first take advantage of existing operations' data sources, before investing in IIoT-enabled sensors or software. Capitalizing on the billions of data points already being generated by SCADA and automation systems augments maintenance records front-line workers use daily.

2.1 Stages in the Vibration Control Process

Condition Monitoring systems are the main technical reliability systems. These systems allow for the management of individual assets based on conditions. There is a wide array of available measurement technologies from portable systems to continuous on-line systems with permanently mounted transducers. Through effective monitoring, the technologies allow fault identification, diagnosis, and sometimes prognosis. Vibration signatures of the machine can offer an early warning to the operator for time based maintenance or to make a crucial decision before any serious problem or unscheduled downtime. The amplitude of the vibration signature gives an indication of the severity of the problem, whilst the frequency can indicate the source of the defect [12]. A system of vibro-diagnostics is basically a system which includes:

1) Establishing the laws for changing the parameters of the machine condition and its suitability for control.
2) Selection of vibration parameters and determination of characteristics of their changes and connections with machine condition parameters.
3) Determination of vibration parameter norms.
4) Determining the possibility of diagnosis setting.
5) The choice and technical-economic rationale of relevant method and measurement instrument.
6) Determining the optimal vibration control procedure or algorithm of vibration control (Fig. 2).

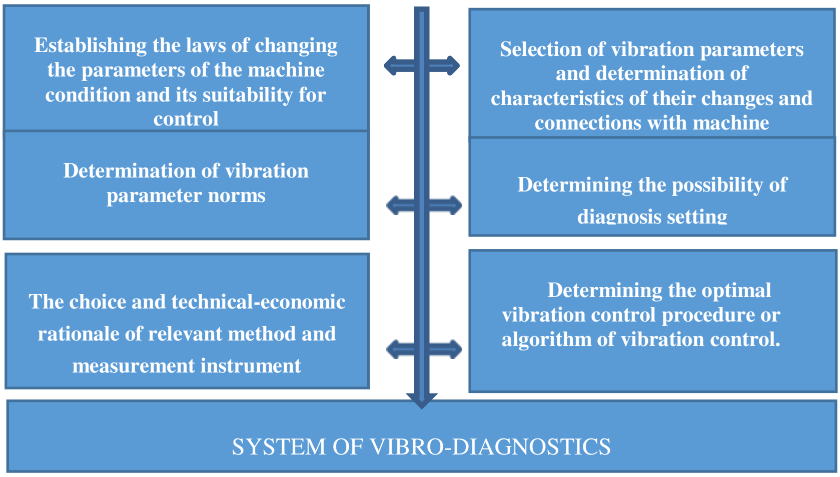

Establishing the laws of changing the parameters of the machine condition and its suitability for control
Determination of vibration parameter norms
The choice and technical-economic rationale of relevant method and measurement instrument
Selection of vibration parameters and determination of characteristics of their changes and connections with machine
Determining the possibility of diagnosis setting
Determining the optimal vibration control procedure or algorithm of vibration control.
SYSTEM OF VIBRO-DIAGNOSTICS

Figure 2
System of vibro-diagnostics

Except for the above mentioned, for the organization of vibration control process the following is required: the mode, technology, location and position of vibration control in the general maintenance system, bearing in mind the basic structure of diagnostics with control and levels of vibration.

Vibration Acceptance and Testing Criteria

It is highly recommended that equipment acceptance testing is performed on completely new and newly rebuilt hardware. The testing efforts should include the following:

1) Vibration measurements on each bearing in two radial directions and the axial direction (i.e. 12 measurements per four-bearing machine).
2) Vibration measurements at each hold-down bolt location.
3) Measurements of velocity and high frequency energy (spike energy or ESP) in rolling element bearing machines and velocity and displacement in sleeve bearing machines.
4) Measurements should be taken with the unit under a loaded condition (at least 70% load when possible). Uncoupled motor data is not acceptable for acceptance testing.

Vibration measurement is an effective, reliable and non- intrusive technique which monitors the condition of the machine during startups, shutdowns and normal operations [1]. Periodicity and scope of work for technical diagnostics are planned, while the foreboding character is provided through continuous monitoring of technical condition of the system with the aim to detect the pre-failure condition (ε_1) and the limit of deterioration ($\varepsilon_2 = \varepsilon_{max}$).

If the condition parameter reaches ε_1, this means that appropriate maintenance activities should be performed to avoid system failure (perform replacement or repair of the system component at the moment of diagnostic control at $\varepsilon \geq \varepsilon_1$). Thus, the measurement of advancing tolerance ($\Delta\varepsilon = \varepsilon_2 - \varepsilon_1$) – connected to the measurement of periodicity of diagnostic control ($\Delta \boldsymbol{T} = T_2 — T_1$).

Therefore, proactive maintenance strategy is a set of rules for determining the regime of diagnostics of system components in the real process of exploitation and making decisions on the necessity for their replacement or for necessary volume of maintenance, based on the information on real technical condition of system.

The best variant solution for specific technical system defines specific model of vibro-diagnostics maintenance (Fig. 3).

2.2 The Method of Data Collection and Processing

The controls and measurements of vibrations of turbine bearing housing, as well as of the bearing clearance of the turbo-generator's turbine have been performed in “Smederevo Stell Mill“. The mentioned measurements are performed by diagnostics devices where the measurement results are automatically processed by a program application and stored in the main maintenance system program. The measurements of bearing clearance have been performed with standard etalons - tickets.

The control and measurements are performed 4 times a year.

For the mentioned research, the diagnostic parameters data of the vibration value of turbo-generator’s turbine bearing housing have been used, at which increased values of bearing clearance appeared within a period of two years, i.e. the data from 8 controls have been used.

Mathematical expectation $m_\varepsilon(t)$ and standard deviation $\varepsilon(t)$ are approximated by line functions. When determining the correlation coefficient, correlation theory method has been used.

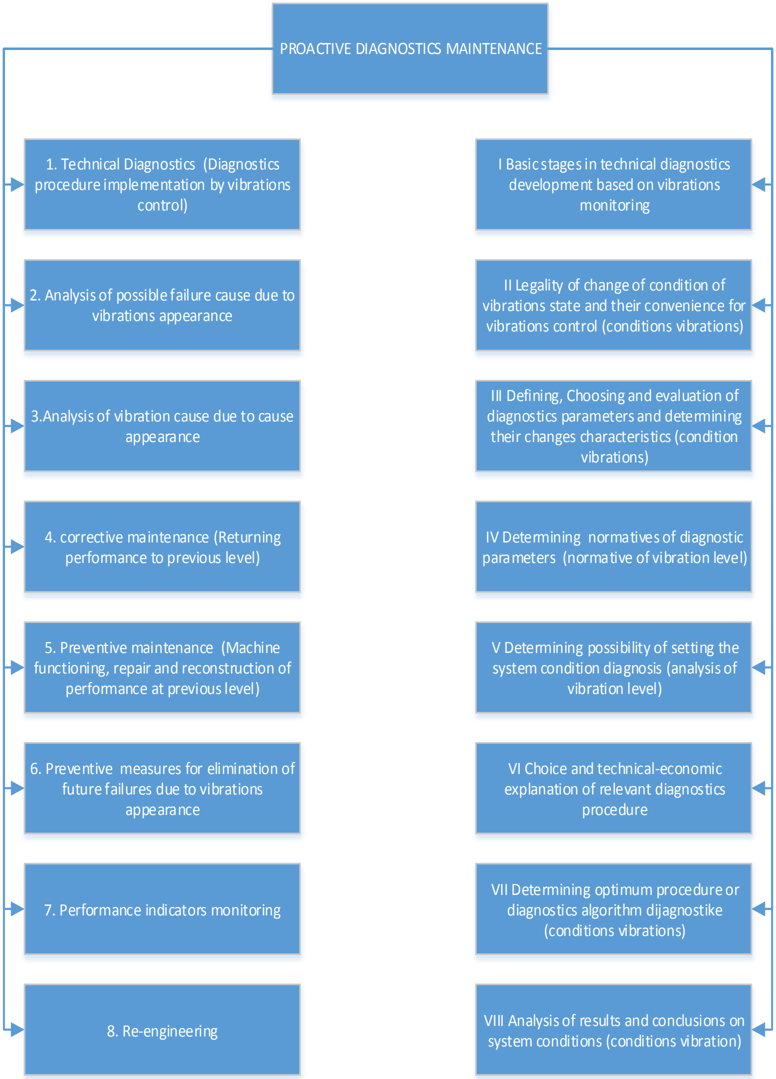

PROACTIVE DIAGNOSTICS MAINTENANCE
1. Technical Diagnostics (Diagnostics procedure implementation by vibrations control)
2. Analysis of possible failure cause due to vibrations appearance
3.Analysis of vibration cause due to cause appearance
4. corrective maintenance (Returning performance to previous level)
5. Preventive maintenance (Machine functioning, repair and reconstruction of performance at previous level)
6. Preventive measures for elimination of future failures due to vibrations appearance
7. Performance indicators monitoring
8. Re-engineering
I Basic stages in technical diagnostics development based on vibrations monitoring
II Legality of change of condition of vibrations state and their convenience for vibrations control (conditions vibrations)
III Defining, Choosing and evaluation of diagnostics parameters and determining their changes characteristics (condition vibrations)
IV Determining normatives of diagnostic parameters (normative of vibration level)
V Determining possibility of setting the system condition diagnosis (analysis of vibration level)
VI Choice and technical-economic explanation of relevant diagnostics procedure
VII Determining optimum procedure or diagnostics algorithm dijagnostike (conditions vibrations)
VIII Analysis of results and conclusions on system conditions (conditions vibration)

Figure 3

Model of proactive diagnostic maintenance

3 Methodology

Change of structural parameters and thus, the change of technical condition is random process that is performed under the influence of a wide range of exploitation factors. This process can fully be described by the density of the distribution of condition parameters f (ε, t) at any moments of time. It is adopted here that f (ε, t) obeys the normal law of distribution.

At same-type components and/or systems, the parameters of technical condition reach the limit value (deterioration limit ε_2) at different moments of time. In this way, the density of distribution of condition in failure $\varphi(t,\ \varepsilon_2)$ is formed. Here too, it is adopted that $\varphi(t,\ \varepsilon_2)$ obeys the normal law of distribution.

To determine the equation linking the distribution density functions $f(\varepsilon,\ t)$ and $\varphi(t,\ \varepsilon_2)$, we use Figure 4, where we adopt a linear change of system condition.

$$P\{t_x \leq T_i\}_{\varepsilon_1} = P\{\varepsilon > \varepsilon_2\}_{T_2} \tag{1}$$

i.e. using known laws from probability theory, the following equations can be written:

$$\int_{\varepsilon_1}^{\varepsilon_2} f(\varepsilon, T_2)\, dx = \int_{T_1}^{T_2} \varphi(t, \varepsilon_1)\, dt \tag{2}$$

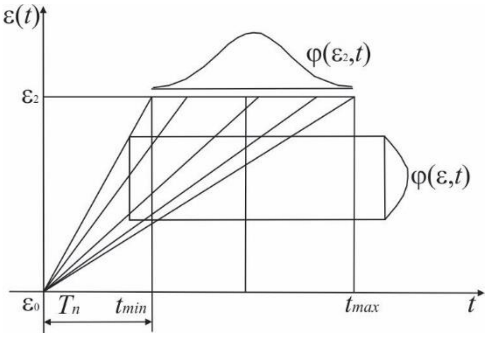


Figure 4
Correlation of parameters of technical condition and technical condition of system and after sorting mathematical model can be obtained

Therefore, it can be concluded that monotonic process ε(t) with given T1 and ε_2, following the moment of the diagnostic control T_2 and the pre-critical level ε_1, satisfies the last equation for general conditions of technical exploitation of the system.

During the real exploitation process, all values of $\varepsilon(t_x)$ will be grouped around the mean value of $\varepsilon(t_x)$ and will have diffusion around it expressed by standard deviation (standard deviation appears due to frequent starting and stopping of the system, due to different modes of exploitation process, etc.).

Now, the equation can be provided, which fully describes the model of change of parameters with the use of confidence intervals

$$\varepsilon(t) = u \cdot t + \varepsilon_0 \neq \sigma_{tot} \tag{3}$$

where:

u – condition change speed $\left(u = \frac{d\varepsilon}{dt}\right)$

ε_0 – minimum value of condition parameter

Having in mind that given confidence level is $p_z = R_z$, and the allowed probability of failure occurrence $p_0 = 1 - p_z$ can be written for some point of time t_z

$$p_0(t_x) = \int_0^{t_x} \varphi(t_x, \varepsilon_2)dt = \frac{\Phi(\varepsilon(t_x)\text{-}\varepsilon_2)}{\sigma_{tot}} \tag{4}$$

where:

Φ – Gaussian function, and

$\varepsilon(t_x)$ – mean value of condition change

Therefore, the following equation results:

$$\int_0^{T_1} \varphi(t, \varepsilon_1)dt = \int_{T_1}^{T_2} \varphi(t, \varepsilon_2)dt \tag{5}$$

Which can help in explaining the physical meaning of mentioned theorem, depending on whether it is a continuous or periodic diagnostic control.

For normal law of change of conditions parameters, the mathematical expectation mc (t) and mean square deviation $\sigma_\varepsilon(t)$ are approximated by linear dependences:

$$m_\varepsilon(t) = m_a + m_b t \tag{6}$$

$$\sigma_\varepsilon(t) = \sigma_a + \sigma_b t \tag{7}$$

Where m_a and σ_a, and parameters of technical condition at the moment $t = 0$, and represent the deviation of condition parameters from its initial value ε_0, which may be constructively allowed deviation (e.g. the initial gap in sliding pitch). Such approximation will be of great use at determining the ratings of diagnostic controls.

Now the density of distribution $f(\varepsilon, t_2)$ can be determined according to [1]:

$$f(\varepsilon, t_2) = \frac{1}{\sqrt{2\pi}(\sigma_a + \sigma_b \cdot t)} \exp\left(\text{-}\frac{\varepsilon\text{-}m_a\text{-}m_a \cdot m_b}{2 \cdot (\sigma_a + \sigma_b \cdot t)}\right) \tag{8}$$

It should be also stated herein that for each controlled parameter (vibration and noise level, quantity of products deteriorated in oil, etc.) it is necessary to determine the failure limits (ε_2).

When selecting diagnostic parameters of the system, it is necessary to determine the character of their relationship to the parameters of condition. In doing so, one or more diagnostic parameters may define only one condition parameter.

The choice of diagnostic parameters (ρ) can be derived based on several basic criteria, using the following characteristics: informative, relative relationship, consent, variation and relation.

The informative nature of diagnostic parameter (or "diagnostic weight") can be estimated over the number of information on technical condition of system that contains that parameter.

Mean value of the information can be used not only to select diagnostic parameters (ρ).

Max relative parameter ratio can be defined as follows:

$$\mathrm{M_d} = \frac{\rho_{\mathrm{max}}\text{-}\rho_{\mathrm{n}}}{\varepsilon_2\text{-}\varepsilon_{\mathrm{n}}} = \frac{\Delta\rho}{\Delta\varepsilon} \tag{9}$$

Where M_d should have the highest values possible.

The studies show that the most favorable ratio is M_d>2.30

Between the diagnostic parameters and the parameters of the technical condition, the required agreement must exist, i.e. to monotonic increase or decrease should correspond relevant change of ***p***, but inversely a proportional change may occur anyways.

The variation represents the deviation of parameters from mean static value.

The correlation coefficient r (in this case r is the degree of connection between ε and ρ) can be taken as a measure of the relationship between the diagnostic parameters and the relevant condition parameters, when solving specific tasks.

3.1 Anticipation of the Turbo Machine Condition

Anticipation is the prediction (forecast) of behavior of system condition parameters in the future, after performed diagnostics, with the aim to ensure the required efficiency of the exploitation process. Determining the "usability reserve" of a system is necessary in terms of its optimal reduction, as well as determining the timing of the following diagnostic condition controls or determining the moment of performing the necessary maintenance activities. Therefore, the anticipation results are the basis for decision making.

As the entry measurement in the anticipation of "usability reserve", necessary level of reliability occurs (p_z = R_z), which is expressed thru quantile of normal distribution ($u_{1\text{-}\rho 0}$), where numeric value is tabulated.

The moment of first diagnostic condition control: can be obtained from the condition that the system meets the required reliability (R_z).

Thus, the moment of the first diagnostic control of the condition can be obtained, after entering *f(ε, T₁,)* for normal distribution [4]:

$$\mathrm{T}_1 = \frac{1}{\mathrm{m_b}\text{-}\sigma_\mathrm{b}\mathrm{u}_{1\text{-}\rho_0}}\left(\varepsilon_2\text{-}\sigma_\mathrm{a}\mathrm{u}_{1\text{-}\rho_0}\text{-}\mathrm{m_a}\right) \tag{10}$$

where:

$u_{1-\rho_0}$ quantile of normal distribution (cumulative frequency), that corresponds to the probability of lawless P_z for the time *T*.

Generally, the limits of diagnostics measurements can be double – alternative and anticipation of "usability reserve" (period of leftover usage).

In both cases it is necessary to determine the limit values of condition parameters.

The measures ε_1 and $\Delta\varepsilon$ can be calculated (for normal law of parameters distribution) in accordance with the expression:

$$\varepsilon_1 = \frac{\sigma_\mathrm{a}(\varepsilon_2\text{-}\mathrm{m_b}\Delta\mathrm{T})+\sigma_\mathrm{b}(\varepsilon_2\mathrm{T}_1+\mathrm{m_a}\Delta\mathrm{T})}{\sigma_\mathrm{a}+\sigma_\mathrm{b}\mathrm{T}_1\Delta\mathrm{T}} \tag{11}$$

$$\Delta\varepsilon = \frac{\sigma_\mathrm{a}(\varepsilon_2\varepsilon_\mathrm{b}+\sigma_\mathrm{b}\sigma_\mathrm{a}\text{-}\mathrm{m_a}\sigma_\mathrm{b})\Delta\mathrm{T}}{\sigma_\mathrm{a}+\sigma_\mathrm{b}(\mathrm{T}_1+\Delta\mathrm{T})} \tag{12}$$

where:

$\Delta T=T_2 - T_1$ – period of diagnostic controls

T_2 – moment in which the next diagnostics control

T_1- moment of the first diagnostic control

If measured value of conditions parameters is below defined limit value, then the moment in which the next diagnostics control of condition is performed, should be looked for. This is how the expression for defining the moment of the second diagnostic control is got [3]:

$$\mathrm{T}_2 = \frac{1}{\mathrm{u}_1\text{-}\mathrm{p}_0}[\varepsilon_1 + \Delta\varepsilon(\mathrm{T}_1)] \tag{13}$$

In the moment T_2 the same procedure is repeated as in the moment T_1.

The calculation of moment of the next diagnostic control of condition T_{2+n} is performed according to the equation:

$$\mathrm{T}_{2+\mathrm{n}} = \frac{1}{\mathrm{u}}[\varepsilon_1 + \Delta\varepsilon(\mathrm{T}_{1+\mathrm{n}})] \tag{14}$$

The process is repeated in the moment when measured value of condition parameter becomes $\varepsilon(T_n)>\varepsilon_1$. It is then when relevant maintenance activities should be performed.

Maintenance activities, except mentioned examples, can be performed also in the case when it is required from the system to function longer than predicted by calculated moments of condition control. Such cases can be represented by the expression:

$$T_2 \geq T_{2+n} - T_{1+n} \tag{15}$$

In negative condition, the system continues to operate in the idle state until T_{2+n}, when the next diagnostic condition check is performed.

Based on previous considerations, a system state anticipation algorithm can be provided that allows the use of *N* diagnostic parameters, with always the lowest of *N* possible values for the (T_1, T_2, T_{2+n}) diagnostic controls selected. The calculation of condition parameters, for all diagnostic parameters, is done in moments T_{1min}, T_{2min}, T_{2+nmin}.

3.2 Application of the Mathematical Model of Vibro-diagnostics on the Case of Turbo-generator in Smederevo Steel Mill

The research of basic indicators of vibro-diagnostic maintenance model were performed on the example of turbo-generator in Smederevo Steel Mill.

As diagnostic parameter, in this case, the vibrations of the turbine bearing housing (p) were selected, resulting in increased gaps in bearings.

Based on the research performed within the period of 2 years the following were selected:

ε_0=115 [μm]

ε_2= 155 [μm]

R_Z= 0.98

u=0.118 [μm/h]

u_{1-p0}=3.8

mathematical expectation $m_i(t)$ and standard deviation $\sigma_i(t)$ were approximated by line functions.

The coefficients m_a and m_b were calculated in accordance with:

$$m_a = \frac{t_{i+1} m_\varepsilon(t_i) - t_i m_\varepsilon(t_{i+1})}{t_{i+1} - t_i} \tag{16}$$

$$m_b = \frac{m_\varepsilon(t_{i+1}) - m_\varepsilon(t_i)}{t_{i+1} - t_i} \tag{17}$$

The coefficients σ_a and σ_{ib} were calculated by analogue functions as m_a and $m_{b.}$ the estimate of mean value m of whole population (total quantity of replaced component part of the system), based on the research of limited sample, in case of normal law, was performed with the help of standardized normal distribution and with help of so called ***t***-distribution. This is how the following values were calculated:

$m_a = 50.42$

$m_b = 0.290$

$\sigma_a = 17.40$

$\sigma_b = 0.056$

In order to determine the character of dependence of parameters of technical condition from diagnostic parameter (ε to ρ), the theory of correlation was used.

As this method requires many numerical operations, a computer was used. Based on algorithms and program, from wanted correlation (in this case linear dependence) were calculated the correlation coefficients. Correlation coefficient and coefficient of regressive functions direction are as follows:

$r=0.55$

$a_1=1.10$

$b_1=0.070$

Comparing correlation functions with real values, it was determined that these were represented with enough punctuality.

The control of hypothesis on normal distribution $f(\varepsilon_1,t)$ in accordance with the criterion of Pearson and Kolmogorov, has shown its correspondence with examined data.

For shown parameters in accordance with the relations, the moments of the first diagnostics condition control (T_1) have been calculated, as well as pre-critical level (ε_1)

$T_1= 640$[h] $\varepsilon_1=124$[μm]

Having in mind that the calculated values of technical condition parameters ($\varepsilon(T_1)$) are less than pre-critical condition ε_1, the turbine continued the process of exploitation without performing maintenance activities.

The first moment of diagnostics control for bearing as a whole is determined from the conditions T_1 = min (T_1). The adopted value is T_1 = 790 hours. The moment of the next diagnostic control T2 is calculated according to the formula and is T_2 = 3436 hours. For the measured values of diagnostic parameter at the moment T_2, the technical state parameter ($\varepsilon(T_2)$) was calculated.

$p(T_2)=134$ [μm]

$\varepsilon(T_2)=141$ [μm]

For the determination of the value of required interval T_2 = 2360 hours, T_2 = T_2 - T_1 = 2620 has been obtained, which means that it was necessary to perform the planned maintenance activities at the moment T_2 = 3428 hours (the lowest T_2 is adopted). After this maintenance model has been developed in accordance with the condition for turbine bearing, the maintenance workforce hiring cost has been reduced and the availability of the system has been increased. The developed program is universal and can be used on N different diagnostic parameters for different technical systems.

3.3 Optimization of Cost of Vibration Diagnostics and Failures of Technical Systems

The essence of this submodel is that it is possible to calculate the optimal interval for the vibration diagnostics of technical systems with high malfunction costs (n_{opt}), which is often the case in thermal power plants and hydropower plants. Total malfunction costs of failure of technical T_T systems decrease at the beginning, and later they increase (Fig. 5), while the costs of diagnostics T_D increase with the increase of vibration interval between two vibration diagnostics.

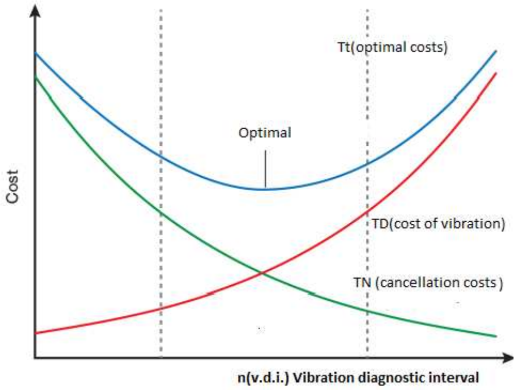


Figure 5

Total costs of vibration diagnostics and malfunctions of technical systems

From this graph, the optimal point in terms of costs and number of failures can be identified within the center of the intelligent maintenance sector; intelligent maintenance can be realized with an online condition monitoring solution [9].

Total costs of T_T malfunctioning of technical systems occur as the sum of average costs: T_C vibration diagnostics, T_D malfunction, T_R maintenance and T_L outage due to the vibration diagnostics process.

The sum of all average costs is:

$$\mathrm{T_{T_T}} = \mathrm{T_C} + \mathrm{T_D} + \mathrm{T_R} + \mathrm{T_L} = \frac{\mathrm{T_C}}{\mathrm{n}} + \frac{\mathrm{n}+1}{2}\frac{\mathrm{T_D}}{\mathrm{n_R}} + \frac{\mathrm{T_R}}{\mathrm{n_R}} + \frac{\mathrm{n_L T_L}}{\mathrm{n_R}}\,[\text{euro/piece}] \tag{18}$$

where: n – vibration diagnostic interval

n_R – estimated number of technical systems between immediate vibration diagnostics

n_L - number of missed technical systems in process interruption due to vibration diagnostics

It is evident that T_{TT} has a continuous flow and all derivatives, so its minimum value can be calculated with optimal vibration diagnostic interval, usingrequired condition for calculating extremes with the first derivative of the function.

$$\frac{dT}{dn}\left(\frac{T_C}{n}+\frac{n+1}{2}\frac{T_D}{n_R}+\frac{T_R}{n_R}+\frac{n_L T_L}{n_R}\right) = -\frac{T_C}{n^2}+\frac{T_D}{2n_R} \tag{19}$$

by equalizing the first derivative of the function with zero and with sufficient condition that the second function derivate is greater than zero, the optimal vibration diagnostics interval is obtained.

$$n_{opt} = \sqrt{\frac{2n_R T_C}{T_D}} \quad [\text{piece}] \tag{20}$$

optimal total malfunction costs are now obtained [9].

$$T_T = \sqrt{\frac{2T_C T_D}{n_R}} + \frac{\frac{T_D}{2}+T_R+n_L T_D}{n_R} \tag{21}$$

The last expression is missing the measurement for repairs T_R, so the optimal interval will not be correctly determined.

At successive diagnostics of vibrations, the probability of occurrence of the *1st*, *2nd*, *3rd* and *n-th* malfunctioning technical systems are mutually equal (Fig. 6)

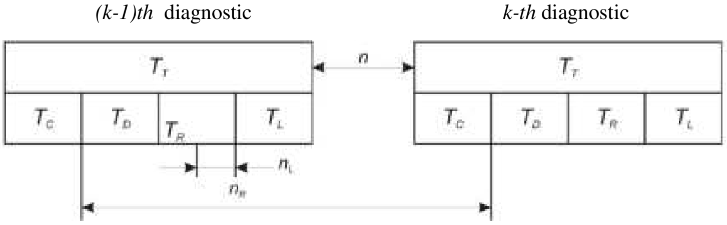


Figure 6

Cost ratio and number of technical systems for successive vibration diagnostics

When malfunctioning technical system is found to be in k-th vibration diagnostics, it causes the occurrence of malfunctioning technical system between *(k-1)th* and *k-th* diagnostic vibration, and therefore the average number of malfunctioning technical systems is *n/2*. Now the actual number of technical systems is between two consecutive settings (n_R+ *n/2*) instead of n_R.

Based on that:

$$\mathrm{T_T} \cong \frac{\mathrm{T_C}}{\mathrm{n}} + \left(1\text{-}\frac{\mathrm{n}}{2\mathrm{n_R}}\right)\left(\frac{\mathrm{n}+1}{2}\mathrm{T_D} + \mathrm{T_R} + \mathrm{n_L T_D}\right) \tag{22}$$

After the second derivative (greater than zero) *dT$_T$/dn*, the optimal interval is obtained for vibration diagnostics of technical systems with high malfunctioning costs:

$$\mathrm{n_{opt}} = \sqrt{\frac{2\mathrm{T_C}(\mathrm{n_R}+\mathrm{n_L})}{\mathrm{T_D}\text{-}\frac{\mathrm{T_R}}{\mathrm{n_R}}}} \tag{23}$$

For the period of *2* years:

T$_N$= 470000 euro

T$_D$=490000 euro

Conclusions

Predictive maintenance requires predictions that are reliable. Examination of vibration phenomenon provides us with the data concerning the volume of working parameters changes and the intensity of vibrations.

On the basis of the obtained results, we evaluate the safety. Often cited, in most cases, it is necessary to determine the cause of non-stationary occurrences that should be either removed or amortized. Working ranges that should be avoided are also determined in many cases.

The primary sources of vibrations for rotating pumps are mechanical, hydraulic and electric processes, caused by the design, manufacturing technology, working regime and exploitation.

The primary goal of proactive maintenance is to find degradation mechanisms which lead to the failure of some elements.

Compared to predictive maintenance, i.e. the maintenance by condition, proactive maintenance does not imply determination of the stage in which an element will fail (what its remaining useful life cycle is), but it tries to discover the mechanism that leads to failure, to mitigate it or to completely eliminate it, in order for the life cycle of an element to be optimally prolonged.

Therefore, the goal is to delay at maximum, or even to completely eliminate, the failure occurrence. Quality issues can hardly be solved without knowledge and an operating knowledge management system. In order to fulfill strategic goals, it is worthwhile to only take quality efforts into consideration [2].

References

[1] Adamović, Ž., Josimović, LJ., Vulović, S., Ilić, B., Spasić, D.: Vibro-diagnostic Maintenance of Technical Systems, Serbian Society for

Technical Diagnostics Adam institute, Smederevo, 2016, ISBN-978-86-83701-39-1

[2] Bencsik, A., Horváth-Csikos G.: The Role of Knowledge Management in Developing Quality Culture: Acta Polytechnica Hungarica, Vol. 15, No. 8, 2018

[3] Diewald, W., Nordmann, R.: Parameter Optimization for the Dynamics of Rotating Machinery, Proceedings of the Third IFToMM International Conference on Rotordynamics, Lyon, France,1990

[4] Gunter, E.: Understanding amplitude and phase in rotating machinery, Vibration instittute 33Annual Meeting, Harrisburg, PA., June 23-27, 2009

[5] Geissbauer, R., Vedso, J., Schrauf, S.: '2016 Global industry 4.0 survey', industry 4.0: Building the digital enterprise, 2006, https://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016.pdf

[6] Hansen, I. H.: Performance Measurement of the Maintenance Function Within Ecomold Ltd Master thesis in Industrial Economy and Information Management, Agder University College, Grimstad, 2006, https://uia.brage.unit.no/uia-xmlui/bitstream/handle/11250/138332/ Hansen.pdf?sequence=1&isAllowed=y

[7] Lifson, A., Simmons, H. Smalley, A.: Vibration Limits for Rotating machinery, Mechanical Engineering, pp. 60-65, 1987, ISSN: 0025-6501

[8] Ličen, H., Zuber, N.: Predictive maintenance of rotating equipment based on vibration measurements and analysis, Technical Diagnostics journal, VI/No. 1, Belgrade, 2007

[9] Mauntz, M., Peuser, J.: Identification of Critical Operation Conditions of Industrial Gearboxes by 24/7 Monitoring of Oil Quality, Oil Aging, and Additive Consumption, 6th International Conference on Fracture Fatigue and Wear, IOP Conf. Series: Journal of Physics: Conf. Series 843, 2017

[10] Morales, D., K.: CBM Policy Memorandum. Washington DC: Deputy under Secretary of Defense for Logistics and Material Readiness, 2002

[11] Muszynska, A.: Vibrational Diagnostics of Rotating Machinery Malfunctions, International Journal of Rotating Machinery, Vol. 1, No. 3-7; Amsterdam, 1995

[12] Peng, Z, Kessissoglou, N.: An integrated approach to fault diagnosis of machinery using wear debris and vibration analysis, Wear 255(7):pp. 1221-1232, 2003

[13] Curran, K., King, R.: Predictive Maintenance for Vibration-Related Failures in the Semi-Conductor Industry, Comput. Eng. Inf.Technol, 8:1 Journal of Computer Engineering & Information Technology, 2019

[14] Scheffer C, Girdhar P: Practical machinery vibration analysis and predictive maintenance. Elsevier, Amsterdam, 2004, ISBN-9780750662758

[15] Vulović, S.: Integrated maintenance model based on the change establishment of principles of Mechanical vibrations change and its impact on prognosticis of condition of rotor engines (PhD dissertation), University in Novi Sad, 2018, www.cris.uns.ac.rs

[16] Vyas, R. K., & Pophaley, M.: Plant maintenance management practices in automobile industries: A retrospective literature review, Journal of Industrial Engineering and Management, pp. 512-541, 2010, ISSN: 2013-8423

Accurate Calculation of Stratified Ground Low-Frequency Return Impedance

Karolina Kaszás-Lažetić, Dragan Kljajić, Nikola Djurić and Miroslav Prša

Faculty of Technical Sciences, University of Novi Sad
Trg D. Obradovića 6, 21000 Novi Sad, Serbia
e-mail: kkasas@uns.ac.rs, dkljajic@uns.ac.rs, ndjuric@uns.ac.rs, prsa@uns.ac.rs

Abstract: This paper presents the analytical and numerical calculations of the current distribution and impedance per unit length in real, multi-layer and two-layer homogeneous grounds. Realistic situations consisting of different thicknesses of ground layers were considered in these calculations. Combinations of parameters were considered at frequencies between 50 Hz and 2500 Hz (the 50th harmonic, of the basic frequency). Calculations were conducted by applying a software tool developed previously for a homogeneous ground, based on a combination of analytical and numerical mathematical procedures and modified, using boundary conditions. The problem was also solved numerically by applying the COMSOL Multiphysics software package, based on FEM. The results are validated by comparing them with the results obtained from empirical formulae applied in practical cases.

Keywords: Stratified ground; Two-layer soil; Current distribution; Ground return impedance; Poynting vector flux

1 Introduction

The probability of current, in the ground, is quite high, in all systems for electrical energy transmission and distribution. In power electrical theory and practice, current can be an essential part of the system, as in the case of a single-wire-earth-return, or it can occur by accidental ground faults, lightning strikes or utility overvoltage. In most cases only low frequencies appear in the ground, given the basic industrial frequency and several higher harmonics, and it is of great importance to be well acquainted with both the current distribution in the ground and with the ground impedance.

On the other hand, depending on the electromagnetic sources, that produce current in the ground, time variations of the current could be rapid, so electromagnetic transients and fast transients should be studied as well. In this case, the electromagnetic field cannot be treated as quasi-static; instead, electromagnetic wave equations have to be applied to determine Transverse Electromagnetic (TEM) waves or quasi TEM waves. Those cases deal with frequencies up to several tenths or hundreds of a MHz.

In order to find an accurate solution for practical problems dealing with the optimization of power transmission and distribution grounding systems, in this paper, only the first case (low frequency, quasi-static electromagnetic fields), is investigated.

For the case of lumped parameters (grounding resistors, capacitors and inductors) applications, which are frequently defined and applied in the power engineering practice, the results could be less accurate, especially in the presence of higher harmonics.

Some of practical approaches have used the concept of the simplified soil model, in which the empirical formulae and diagrams are usually applied. These methods could be efficient in most usual problems, but could also cause significant inaccuracies in applied calculations.

Our first attempt to determine return parameters of low-frequency transmission lines deals with homogeneous soil as the current conducting media. The problem has been identified and calculated in a number of ways that can be classified into three main groups. All the methods, together with their classifications and results are described in detail in [5].

Stratified ground is also the subject of several papers, but mostly at higher frequencies [2] [10] [11] [12]. This is expected given that, in nature, the soil layer thickness is usually much smaller compared to the penetration depth of the low frequency electromagnetic fields. Nevertheless, most of the mentioned papers describe the behavior of the multi-layer soil at a higher frequency range, from 50 Hz to 1 MHz, and only a few contributions at extra low frequencies for homogeneous soil were found in [8].

This paper's significant contribution is the new approach, based on the Poynting theorem, partially presented in [5], improved and successfully adjusted for a multi-layer ground analysis. The developed method takes into account all existing parameters, including soil layer thickness, soil layer resistivity values, conductor height and operational frequency. The other novelty in this paper is the introduction of the third coordinate system to define and apply boundary conditions necessary for an analytical calculation of the electromagnetic field in the case of stratified ground. The proposed procedure has not been used elsewhere in the explored references. Some initial results in the current distribution calculation inside a two-layer soil are given in [3] [4], while some of the other authors' results dealing with multi-layer soils are presented in [7] [13] [16].

The paper consists of three main parts. Section 1 provides background information about the problem, and points out the importance of the new method. Section 2 provides insight into the model geometry and explains the application of the proposed method to multilayer soil. Section 3 presents the most significant results, while the conclusion emphasizes the most important achievements.

2 Theoretical Approach

For the stratified soil return impedance determination, we assume the same system adopted in [5] and depicted in Figure 1; the system contains an overhead conductor that is parallel to the ground surface, with ground serving as a return conductor. The problem could be treated as two-dimensional, in a plane that is perpendicular to the overhead conductor (*x*-*y* plane, *A*–*A* crossing).

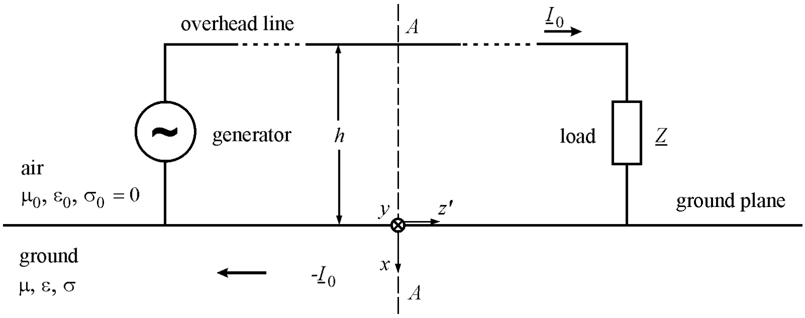


Figure 1
The principle of the ground as a return conductor

In the investigated case of stratified ground, the earth is assumed to have an infinite cross-section, with *n* layers of generally different thicknesses and generally different electromagnetic parameters, cf. Figure 2, for the *A*–*A* crossing.

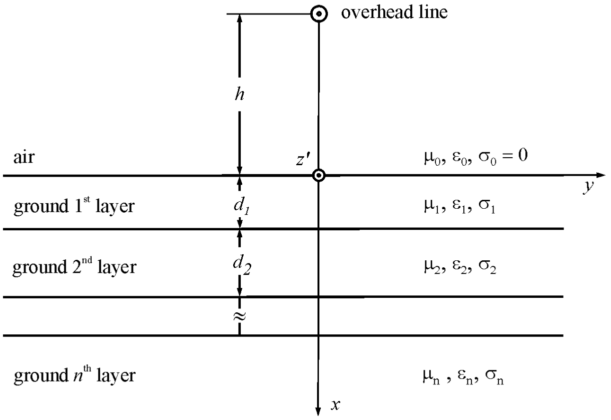

overhead line
h
air
z'
μ_0, ε_0, $\sigma_0 = 0$
y
ground 1st layer
d_1
μ_1, ε_1, σ_1
ground 2nd layer
d_2
μ_2, ε_2, σ_2
≈
ground nth layer
μ_n , ε_n, σ_n
x

Figure 2
Overhead line over a multi-layer ground

The multi-layer ground conducts the current $-\underline{I}_0$ from the load impedance $\underline{Z}$ towards the generator. Due to the skin effect and the proximity effect with the overhead conductor, and due to the presence of layers with different conductivity values, the current density vector is not uniformly distributed across the infinite cross-section. In homogeneous soil, the current density's maximal magnitude appears under the overhead conductor $(x, y) = (0, 0)$ and decreases with the increasing distance. This maximal value depends on the height of the overhead conductor h, the current frequency f, permeability μ and soil conductivity σ.

In the multi-layer soil case, depending on the layer conductivity values, the maximal current density magnitude is expected to appear at the same point or at the upper boundary between the two neighboring layers.

Adopting the standard analytical approach to the ground return impedance determination, the complex Poynting vector, $\underline{\boldsymbol{P}} = \underline{\boldsymbol{E}} \times \underline{\boldsymbol{H}}^*$, where $\underline{\boldsymbol{E}}$ is the complex electric field strength vector and $\underline{\boldsymbol{H}}^*$ is the conjugate complex magnetic field strength vector, needs to be calculated. From the determined complex Poynting vector, the ground return impedance per unit length, $\underline{Z}'_{\mathrm{G}}$ can then be calculated as (cf. [14]),

$$\underline{Z}'_{\mathrm{G}} = \frac{1}{\left|\underline{I}_0\right|^2} \cdot \int_S \left(\underline{\boldsymbol{E}} \times \underline{\boldsymbol{H}}^*\right) \cdot \mathrm{d}\boldsymbol{S} . \tag{1}$$

The analytical integration of the above integral can be difficult to perform, but the sufficiently accurate solution can be derived by numerical integration.

The determination of the current distribution within the ground, taking into account soil stratification, is the starting point for the calculation of the magnetic field strength vector. For this reason, the first step should be the determination of the electric field strength vector.

2.1 Calculation of the Electric Field Strength Vector within any Homogeneous Media

In order to solve the problem within a stratified soil, the partial differential equation in homogenous soil should be solved first. For this reason, a geometric representation of the problem defined by cross section $A - A$ in Figure 1 is presented in Figure 3 also adopted in [5].

First of all, an adequate coordinate system has to be chosen. According to the problem geometry, the cylindrical coordinate system (r, φ, z) is adopted. In the chosen coordinate system, the z-axis runs along the overhead conductor. Both the electric field strength vector and current density vector have only z components, depending on the radius r.

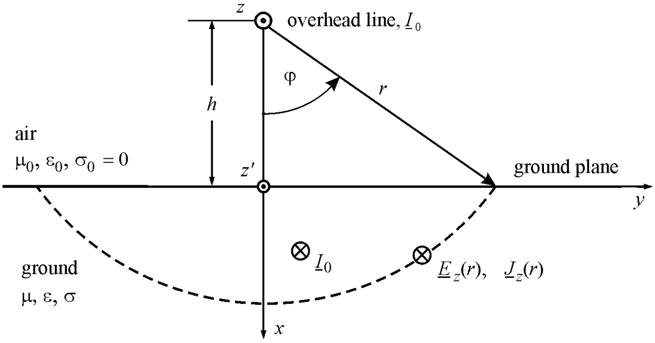

z
overhead line, $\underline{I}_0$
φ
r
h
air
μ_0, ε_0, $\sigma_0 = 0$
z'
ground plane
y
$\underline{I}_0$
$\underline{E}_z(r)$, $\underline{J}_z(r)$
ground
μ, ε, σ
x

Figure 3

Definition of the chosen coordinate systems for the electric field strength vector calculation

The entire calculation process for a complex electric field strength vector calculation in homogeneous soil is described in detail in [5]. Starting from the first two Maxwell's equations in a complex domain for a quasi-stationary electromagnetic field, it is shown, that the complex electric field strength vector in the ground is a solution to the partial differential equation in the cylindrical coordinate system [5] [14]:

$$\frac{\partial^2 \underline{E}_z}{\partial r^2} + \frac{1}{r}\frac{\partial \underline{E}_z}{\partial r} - \underline{K}^2 \underline{E}_z = 0 \tag{2}$$

In (2), the complex constant $\underline{K}^2$ is defined as:

$$\underline{K}^2 = \frac{\mathrm{j}\omega\mu\sigma}{\pi}\arccos\frac{h}{r} \tag{3}$$

For the constant value of the coefficient $\underline{K}^2$, (2) is Bessel's equation with the general solution:

$$\underline{E}_z(r) = \underline{A}I_0(\underline{K}r) + \underline{C}K_0(\underline{K}r) \tag{4}$$

According to [1], the complex electric field strength vector is defined as:

$$\underline{E}_z(r) = \underline{C}K_0(\underline{K}r) \tag{5}$$

The function $K_0(\underline{K}r)$ is divided into the real and imaginary part, and the complex electric field strength vector takes the form of:

$$\underline{E}_z(r) = \underline{C}\left[\ker(ar) + \mathrm{j}\,\mathrm{kei}(ar)\right] \tag{6}$$

In the above equation the parameter a is:

$$a = \sqrt{\frac{\omega\mu\sigma}{\pi}\arccos\frac{h}{r}} = \sqrt{\frac{\omega\mu}{\rho\pi}\arccos\frac{h}{r}} \tag{7}$$

In order to determine the complex constant $\underline{C}$, the complex current density vector should be integrated over the ground cross-section, S_{Gcs}, (x-y plane):

$$\underline{I}_0 = \int\limits_{S_{\mathrm{Gcs}}} \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S}. \tag{8}$$

The complex constant $\underline{C}$ is a function of frequency f, conductor's height above ground h, and soil resistivity value, ρ.

2.2 Complex Current Distribution inside a Multi - Layer Ground

Although many studies are based on the assumption that ground is homogeneous with a constant conductivity value, in reality the soil is not homogeneous; its conductivity changes from point to point. In practice, ground could be thought of as being composed of several layers with different electromagnetic properties.

To have an adequate soil model, the influence of ground stratification on ground impedance calculation should be taken into account.

The soil characteristics can be represented by an idealized multi-layer model, shown in Figure 2.

The top layer is assumed to have a finite depth d_1 from the surface, conductivity σ_1 and permittivity ε_1. The next layer is also assumed to be finite, of thickness d_2, and is characterized by conductivity σ_2 and permittivity ε_2. The last observed layer is assumed to be infinite, with the corresponding electromagnetic characteristics. Assuming that the investigated ground is not made of a ferromagnetic material, permeability values of all layers are set to be the same as the permeability of air, $\mu_1 = \mu_2 \ldots = \mu_n = \mu_0$. Should a soil layer contain any ferromagnetic material, the corresponding value of permeability must be defined and included in the calculations.

In the case of a multi-layer soil, the complex current density vector calculations are slightly different. The complex electric field strength vector is considered to be parallel to the surface between any two soil layers. The boundary conditions that have to be satisfied are that the complex tangential components of the electric field strength vector must be the same on both sides of the surface that separates the i^{th} and the j^{th} layer:

$$\boldsymbol{n} \times (\boldsymbol{E}_1 - \boldsymbol{E}_2) = 0 \quad \rightarrow \quad \underline{\boldsymbol{E}}_{\mathrm{it}} = \underline{\boldsymbol{E}}_{\mathrm{jt}} = \boldsymbol{i}_{\mathrm{z}} E_{\mathrm{z}}. \tag{9}$$

For this reason, the complex electric field strength vector is the same as in the previous model for homogeneous ground. The complex current density vector also has only the z component and will be calculated from the complex electric field strength vector in each soil layer:

$$\underline{J}_1(r)=\sigma_1\underline{E}_z(r) \quad \underline{J}_2(r)=\sigma_2\underline{E}_z(r) \quad ... \quad \underline{J}_n(r)=\sigma_n\underline{E}_z(r). \tag{10}$$

In this case, the current density vector depends on coordinates x and y, and on the conductor height h.

2.3 Complex Magnetic Field Strength Vector

The determination of the complex magnetic field strength vector $\underline{\boldsymbol{H}}$ is difficult because of the different current density vector values in each ground layer and the calculation in three coordinate systems. In addition, the resultant magnetic field strength vector is composed of two contributions.

The first contribution is produced by the complex current $\underline{I}_0$ in the overhead conductor. This part of the magnetic field strength vector at an arbitrarily chosen point on the ground surface, T (0, y), denoted by $\underline{\boldsymbol{H}}_{\mathrm{C}}$, according to Figure 4, is:

$$\underline{\boldsymbol{H}}_{\mathrm{C}}=\frac{\underline{I}_0}{2\pi r}\boldsymbol{i}_{\varphi}. \tag{11}$$

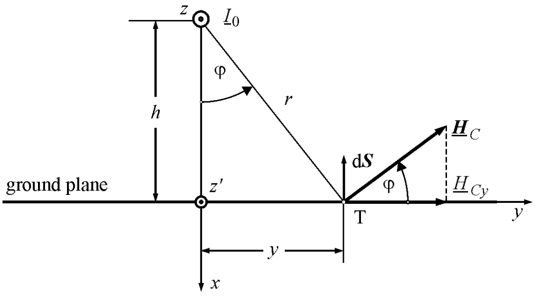


Figure 4

Magnetic field due to the overhead conductor

From (1), it follows that only the y component of magnetic field strength vector is relevant, which is:

$$\underline{H}_{\mathrm{Cy}}=\frac{\underline{I}_0}{2\pi r}\cos\varphi=\frac{\underline{I}_0}{2\pi r}\frac{h}{h^2+y^2}, \tag{11}$$

where

$$\cos\varphi=\frac{h}{r} \quad \text{and} \quad r^2=h^2+y^2. \tag{12}$$

The current density $\underline{\boldsymbol{J}}_{\mathrm{z}}(r)$ within the ground produces the additional contribution to the magnetic field strength vector $\underline{\boldsymbol{H}}_{\mathrm{G}}$. The resultant y component of vector $\underline{\boldsymbol{H}}$ will be the sum of the two parts,

$$\underline{H}_{\mathrm{y}}=\underline{H}_{\mathrm{Cy}}+\underline{H}_{\mathrm{Gy}}. \tag{13}$$

The magnetic field strength vector produced by the current inside the ground is much more difficult to calculate, due to the fact that the layers are parallel to the surface. For this reason, we decided to determine the complex magnetic field strength vector by applying a mathematical procedure in a Cartesian coordinate system, defined by the axes x, y, z' (see Figure 5).

The complex magnetic field strength vector in an arbitrary point $\mathrm{T}(0, y)$, on ground surface, produced by the complex current in the ground, can be determined by applying Biot-Savart's law, where:

$$r=|\boldsymbol{r}|,\quad r^2=d^2+x'^2=\left(y-y'\right)^2+x'^2,\quad \boldsymbol{r}_0=\frac{\boldsymbol{r}}{r}=\frac{\boldsymbol{r}}{|\boldsymbol{r}|}. \tag{14}$$

Complex magnetic field strength vector in point $\mathrm{T}(0, y)$ is then:

$$\underline{\boldsymbol{H}}_{\mathrm{G}}(0,y)=\frac{1}{4\pi}\int_{\mathrm{S}}\frac{\underline{\boldsymbol{J}}(x',y')\times\boldsymbol{r}}{|\boldsymbol{r}|^3}\mathrm{d}S=\frac{1}{4\pi}\int_{\mathrm{S}}\frac{\underline{\boldsymbol{J}}(x',y')\times\boldsymbol{r}_0}{|\boldsymbol{r}|^2}\mathrm{d}S. \tag{15}$$

All quantities in (14) are presented in Figure 5.

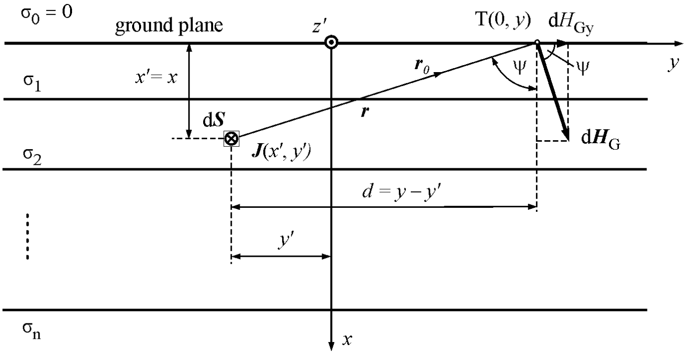


Figure 5
Magnetic field due to the current inside the ground

In order to determine the Poynting vector on the ground surface, the only relevant component of the complex magnetic field strength vector on ground surface is the y component:

$$\underline{H}_{\mathrm{Gy}}(0,y)=\underline{H}_{\mathrm{G}}\cos\psi=\frac{1}{4\pi}\int_{\mathrm{S}}\frac{\underline{J}(x',y')}{r^2}\cos\psi\,\mathrm{d}S=\frac{1}{4\pi}\int_{\mathrm{S}}\frac{\underline{J}(x',y')}{r^2}\frac{x'}{r}\mathrm{d}S. \tag{16}$$

In (16), according to Figure 5, $\cos\psi$ is:

$$\cos\psi=\frac{x'}{r}=\frac{x'}{\sqrt{x'^2+\left(y-y'\right)^2}}. \tag{17}$$

Hence, in order to calculate the *y* component of the complex magnetic strength field vector, the following surface integral must be solved:

$$\begin{aligned}\underline{H}_{\mathrm{Gy}}(0,y)&=\frac{1}{4\pi}\int_{\mathrm{S}}\frac{\underline{J}(x',y')}{x'^2+\left(y-y'\right)^2}\frac{x'}{\sqrt{x'^2+\left(y-y'\right)^2}}\mathrm{dS}=\\&=\frac{1}{4\pi}\int_{0}^{\infty}\int_{-\infty}^{\infty}\frac{\underline{J}(x',y')}{\sqrt{x'^2+\left(y-y'\right)^2}}\frac{x'}{\sqrt{x'^2+\left(y-y'\right)^2}}\mathrm{d}x'\mathrm{d}y'\end{aligned}\tag{18}$$

Knowing that the complex current density vector is a combination of the modified Bessel functions, the integral could be difficult to solve analytically. Nevertheless, this integral could be successfully solved by applying any numerical integration procedure. The integration can be performed on an arbitrarily fine mesh and the obtained results can be very accurate.

Knowing both the complex electric field strength vector $\underline{\boldsymbol{E}}$ and the complex magnetic field strength vector $\underline{\boldsymbol{H}}$, we can calculate the complex Poynting vector and its flux over the ground surface:

$$\underline{\boldsymbol{P}}=\int_{\mathrm{S}}\left(\underline{\boldsymbol{E}}\times\underline{\boldsymbol{H}}^{*}\right)\cdot\mathrm{d}\boldsymbol{S}=\underline{Z}_{\mathrm{G}}\left|\underline{I}_0\right|^2.\tag{19}$$

Applying (1), the ground return impedance can be calculated.

3 Calculation Results and Discussion

The calculation of current distribution was carried out over several two-layer soils, for four different soil layer resistivity values: ρ = l/σ = 50 Ωm, 250 Ωm, 1000 Ωm and 2500 Ωm, together with five values of the overhead conductor height: $h = 10$ m, 15 m, 20 m, 25 m and 30 m. In the case of the two-layer soil, several typical combinations of soil resistivity values are calculated. In this paper two of these combinations are presented and discussed.

As most soil types are non-magnetic, the relative permeability of the ground was assumed to be unity, with the relative permittivity also considered equal to one. The Bessel function values necessary for solving (2) were found in [1]. All calculations in frequencies ranging from 50 Hz to 2500 Hz were examined, assuming a sinusoidal current in the overhead conductor, presented in the complex domain as:

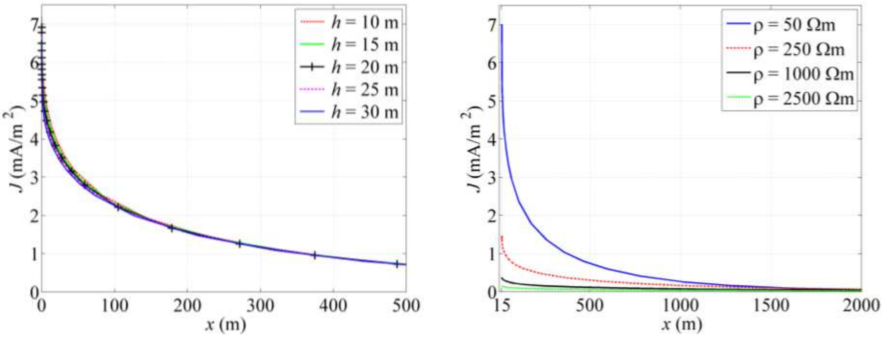
$$\underline{I}_0=\left(1+\mathrm{j}0\right)\ \mathrm{kA}.\tag{20}$$

3.1 Current Distribution in a Homogeneous Ground Model

In the case of a homogeneous ground, the entire calculation procedure described with (2) through (9) can be applied to directly determine the current density vector distribution. This procedure, together with the appropriate results, is described in detail in [5]. In this paper some of these results will be repeated for comparison purposes between the homogeneous ground and the multi-layer or the two-layer soil.

The calculated current distribution inside the homogeneous soil, at $\rho = 50\ \Omega$m and $f = 50$ Hz, for five different conductor heights, is shown in panel (a) of Figure 6, while panel 6 (b) presents the current distribution for a single conductor height of $h = 15$ m, for four different homogeneous soil resistivity values. The diagrams are adopted from [5].

Figure 6 demonstrates that the influence of conductor height on the current distribution within the ground is negligible for the same value of soil resistivity, while the soil resistivity values have a significant impact on the current distribution.



(a) (b)

Figure 6

Current density vector magnitude as a function of overhead conductor height (a) and of soil resistivity values (b)

It is apparent from Figure 6 (b) that the skin effect is the strongest in the case of the smallest soil resistivity value, $\rho = 50\ \Omega$m, and weakest for the largest soil resistivity value, $\rho = 2500\ \Omega$m e.g. its influence decreases with increasing soil resistivity. For all four soil resistivity values the current density vector magnitude is the largest at the surface and decreases rapidly with increasing distance from the above conductor.

In the three other cases, i.e. for higher soil resistivity values, the skin effect is less distinct and the penetration depth is much higher.

3.2 Current Distribution in a Multi-Layer Ground Model

As an example of a multi-layer soil, we investigated two-layer soil typical for our region. Typical thicknesses of the layers in our region are 1 m, 2 m and 5 m, with two-layer resistivity combinations presented in Table 1, obtained from local measurements [9].

Table 1

Two-Layer Soil Resistivity Combinations

Layer	ρ [Ωm]						
1	50	50	50	100	100	100	500
2	100	500	1000	1000	3000	50	50

The distribution of current in all combinations of the two-layer soil presented in Table 1 was calculated. In this paper two typical combinations of current distribution are presented and discussed as an example. The corresponding parameters for these combinations are as follows:

- $\rho_1/\rho_2 = 50\ \Omega\text{m}/1000\ \Omega\text{m}$ $d_1 = 1$ m, $d_2 = \infty$
- $\rho_1/\rho_2 = 500\ \Omega\text{m}/50\ \Omega\text{m}$ $d_1 = 1$ m, $d_2 = \infty$

The current distribution for the first soil resistivity combination, at 50 Hz and at 450 Hz, is shown in Figure 7.

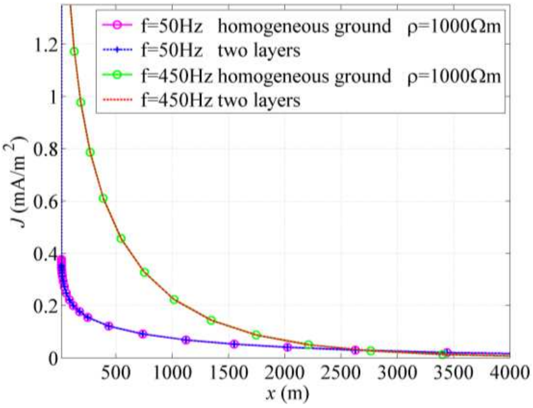


Figure 7

Current distributions inside a two-layer soil, 50 Ωm/1000 Ωm, at 50 Hz and 450 Hz, in a wide range

Due to the negligible thickness of the first layer compared to the overall ground depth, the influence of the soil resistivity values is not visible in Figure 7. Figure 8 depicts a zoomed-in version to the depth of 5 m.

The drop in current density vector magnitude on the boundary that separates the two layers is evident at both frequencies. Due to the less emphasized skin effect at a lower frequency, the difference between the current density vector magnitudes in a two-layer soil is smaller at 50 Hz. Due to the more significant skin effect, the current density vector magnitude drop is much higher at 450 Hz.

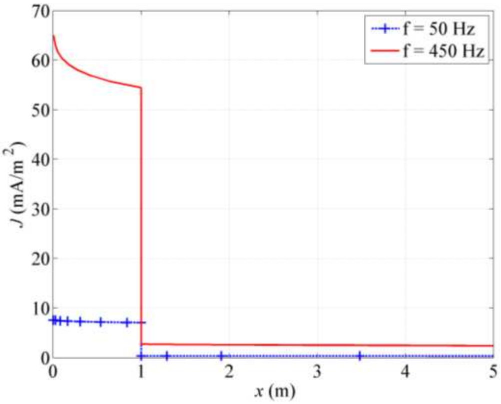


Figure 8

Current distribution inside a two-layer soil, 50 Ωm/1000 Ωm, at 50 Hz and 450 Hz, in a limited range

The same current distribution can also be presented by contour lines and an appropriate color-map. As in Figure 7, the case of current density vector magnitude presentation in a wide range beneath the ground surface, the influence of two different resistivity values cannot be noticed.

For this reason, only the current density vector magnitude distribution to the depth of 10 m is presented in Figure 9. The higher density of contour lines in the upper layer describes the higher current density vector values.

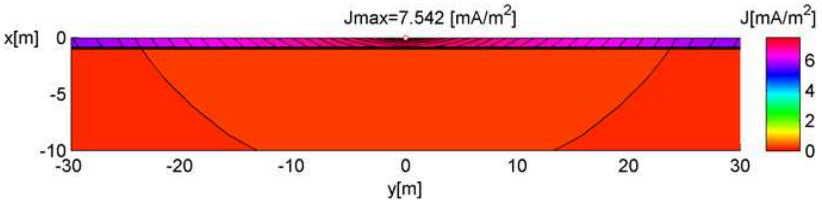


Figure 9

Current distribution inside a two-layer soil, 50 Ωm/1000 Ωm, at 50 Hz, up to the 10 m depth

A similar situation as in the first case, $\rho_1/\rho_2 = 50\ \Omega\text{m}/1000\ \Omega\text{m}$, appears in the second case, $\rho_1/\rho_2 = 500\ \Omega\text{m}/50\ \Omega\text{m}$ as well. When the calculated current distribution is presented in a wide range beneath the ground surface, the influence of two layers with different soil resistivity values is negligible, again due to the negligible thickness of the first layer, comparing to the observed ground depth. For this reason, the results are presented in two diagrams; in a wide ground range and in a limited range beneath the ground surface.

Current distribution for the second case of the two-layer ground is presented in Figure 10, while the same results on a zoomed-in region are presented in Figure 11.

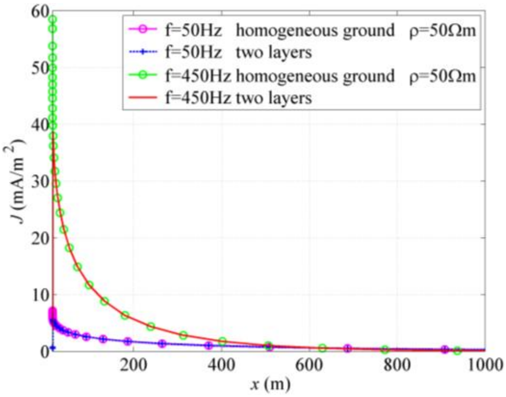


Figure 10

Current distribution inside a two-layer soil, 500 Ωm/50 Ωm, at 50 Hz and at 450 Hz, in a wide range

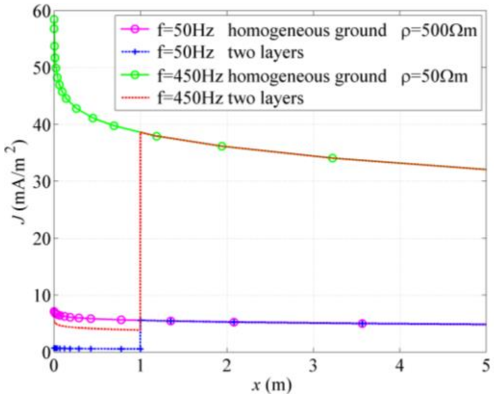


Figure 11

Current distribution inside a two-layer soil, 500 Ωm/50 Ωm, at 50 Hz and at 450 Hz, in a limited range

Distribution of current density vector magnitude, for the same soil resistivity combination, $\rho_1/\rho_2 = 500\ \Omega m/50\ \Omega m$, at 50 Hz, in a limited range, up to the depth of 10 m is presented in Figure 12.

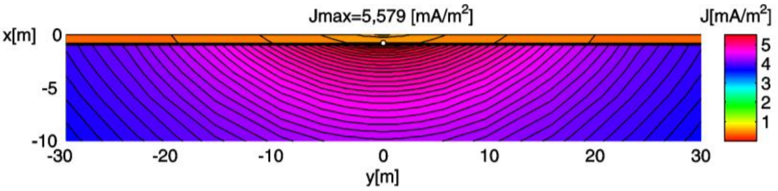


Figure 12

Current distributions inside a two-layer soil, 500 Ωm /50 Ωm, at 50 Hz, up to 10 m inside the ground

Figure 12 demonstrates that the highest value of the current density vector magnitude does not appear at the ground surface, but in the second layer, beyond the separating boundary.

3.3 Penetration Depth

Penetration depth is a measure of how deep any electromagnetic field can penetrate into a material. It is defined as the depth at which the intensity of the field inside the material falls to 1/*e* of its original value at the surface or more properly, just beneath the surface. The penetration depth is a good proxy for the skin effect and is given by the expression [14]:

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} = \sqrt{\frac{2\rho}{2\pi f \mu}} = \sqrt{\frac{\rho}{\pi f \mu}}. \tag{20}$$

The penetration depths in meters, for the five most common soil resistivity values in our region and 10 frequencies, are calculated applying (20) and given in Table 2.

Table 2

Penetration Depths in Meters, for Different Soil Resistivity Values and Different Frequencies

f [Hz]	ρ [Ωm]				
	50	250	500	1000	2500
50	503.29	1125.40	1591.55	2250.79	3558.81
100	355.88	795.77	1125.39	1591.55	2516.46
150	290.57	649.75	918.88	1299.49	2054.68
250	225.08	503.29	711.76	1006.58	1591.55
350	190.23	425.36	601.55	850.72	1345.11
500	159.16	355.88	503.29	711.76	1125.40
750	129.95	290.58	410.94	581.15	918.88
1000	112.54	251.65	355.88	503.29	795.77
1500	91.89	205.47	290.58	410.94	649.75
2500	71.18	159.16	225.08	318.31	503.29

Comparing the layer thickness values to the penetration depth values, it can be concluded that the penetration depth is in all cases much higher that the layer thickness.

As also concluded in [15], due to the very small first-layer thickness compared to the penetration depth, the first layer practically has no influence on the current distribution and, consequently, on the ground return impedance. Both cases could be treated as homogeneous cases with the second layer resistivity values. At all observed frequencies, a significant first layer impact can be expected only if the layer thickness varies between 10 m and 50 m [15].

This conclusion is very important, because it enables all following calculations to be performed for the homogeneous soil model, with the resistivity equal to the resistivity of the second layer. It also eliminates the necessity for the complex magnetic field strength vector calculation, applying (18), or numerical integration.

Nevertheless, the calculations of magnetic field strength vector and ground return impedance were performed for both cases; two-layer soil and homogeneous ground by applying numerical integration.

3.4 Ground Return Impedance Calculation

In order to verify the developed method and calculated impedances, the results obtained by this method were also compared to the results obtained by the numerical procedure based on FEM and the results of the simplified Carson's (Carson-Clem) formula for ground return impedance [17].

Numerical procedure based on FEM was performed by applying a computer package COMSOL MULTIPHYSICS', AC/DC Module. For the entire calculation, the mode "2D Quasi-static, Magnetic/Perpendicular Induction Currents, Vector Potential" was chosen, together with the "Time Harmonic" simulation. The calculations were carried out for the two-layer ground.

The simplified Carson's (Carson-Clem) formula for ground current impedance is presented in [17], as:

$$\underline{Z}' = R'_{\mathrm{C}} + 9{,}8696 f \cdot 10^{-4} + \mathrm{j}2{,}8937 f \cdot 10^{-3} \log \frac{658{,}86875\sqrt{\dfrac{\rho}{f}}}{GMR} \ (\Omega / \mathrm{km}) \qquad (21)$$

Where:

- R'_{C} is the resistance per unit length of the overhead conductor in Ω/m
- ρ is soil resistivity in Ωm
- f is frequency in Hz
- *GMR* is effective radius of the overhead conductor in meters

An improved version of (21) is presented in [6].

Expression (21) cannot be applied in the case of a multi-layer or two-layer ground, hence in (21) the single soil resistivity, the resistivity of the lower ground layer, was taken into account.

Ground resistance per unit length, calculated for the soil resistivity combination $\rho_1/\rho_2 = 500$ Ωm/50 Ωm and $d_1 = 1$ m is shown in Figure 13. The results of the numerical procedure (FEM) for the same ground parameters are also depicted in Figure 13. The results obtained via (21), denoted as "Carson", are calculated for the homogeneous ground with resistivity value of ρ = 50 Ωm. All three calculations were carried out for the conductor height $h = 15$ m.

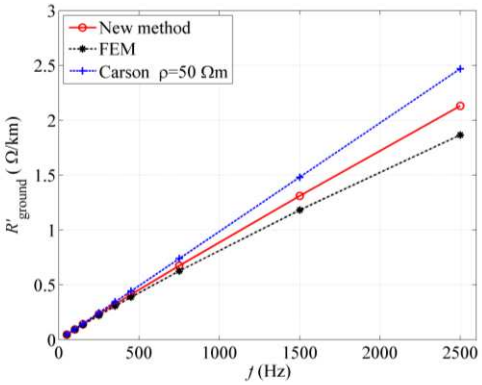


Figure 13

Ground resistance per unit length calculated by applying three different methods

Good agreement of all resistance calculation results is obvious in Figure 13 and this is an excellent verification of the developed method. At the same time, this shows that the simplified formula (21) can be applied for an accurate resistance per unit length determination for two-layer soil as well. This confirms that the simplified expression (21) is still useful, not only for the calculations inside a homogeneous soil, but also for the calculations in any stratified ground with dominant thickness of the lower layer.

Figure 14 displays the relationship between the frequency and the ground reactance per unit length, for a conductor height of h = 15 m, soil resistivity values combination $\rho_1/\rho_2 = 500$ Ωm/50 Ωm and the results obtained by (21) applied in homogeneous ground, ρ = 50 Ωm. The radius of the overhead conductor is set to be $r_c = 0.0144$ m.

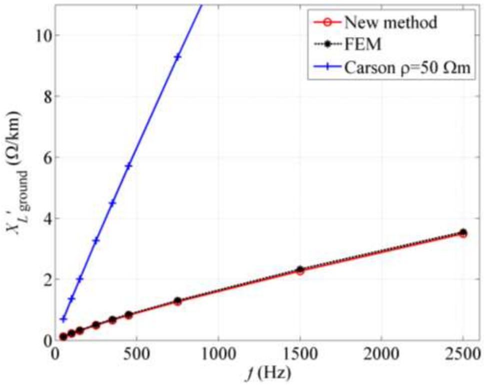


Figure 14

Ground reactance per unit length for h = 15 m and different soil resistivity values, $\rho_1/\rho_2 = 500$ Ωm/50 Ωm

Comparing the calculated results presented in Figure 14, at the chosen soil resistivity combination $\rho_1/\rho_2 = 500\ \Omega\text{m}/50\ \Omega\text{m}$, good agreement is seen between the new, analytical, method and the numerical FEM procedure, but not with the simplified "Carson" formula, for the homogeneous ground with resistivity $\rho = 50$ Ωm.

Analysis of Figure 14 reveals that the ground reactance per unit length does not increase linearly, but rather slowly with an increasing frequency, predicting that the ground inductance per unit length, will be decreasing with increasing frequency, as expected.

Knowing that the ground return impedance is an inductive load, the ground return inductance can be determined as well. Ground inductance per unit length could be very useful, especially in the case of electromagnetic disturbances on neighboring electronic or telecommunication systems.

Figure 15 depicts ground inductance per unit length as a function of frequency, for a constant conductor height of h = 15 m and soil resistivity values combination $\rho_1/\rho_2 = 500\ \Omega\text{m}/50\ \Omega\text{m}$ and d_1 = 1 m.

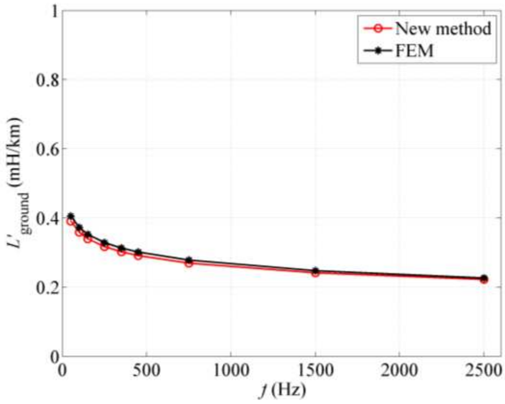


Figure 15

Ground inductance per unit length, for h = 15 m and $\rho_1/\rho_2 = 500\ \Omega\text{m}/50\ \Omega\text{m}$

The skin effect at lowest frequencies is not significant and ground inductance per unit length has the highest value, decreasing slowly with the increasing frequency. In contrast, at the higher frequencies the skin effect is more pronounced and ground inductance per unit length is at its lowest and decreases even more slowly with the increasing frequency.

In Figure 15, an excellent agreement between the results obtained by the new method and the one obtained by FEM can once again be observed.

The inductances per unit length calculated via the simplified Carson's formula (21) were deemed unreliable and these values are not presented in Figure 15.

Conclusions

In this paper, a novel procedure developed previously for the determination of ground impedance in homogeneous soil, described in detail in [5], was improved, adjusted and applied, for the calculation of the ground return impedance, in a two-layer ground. The inhomogeneous ground was simplified by assuming that ground consists of homogeneous layers, of different thickness, parallel to the surface. The mathematical procedure presented in this paper is also based on a strict electromagnetic approach. The presented method is convenient for an accurate multi-layer ground impedance calculation. The method enables an exact treatment of the skin effect within the multi-layer ground.

The results lead to the most important conclusion: due to the small layer thickness compared to the penetration depth in all investigated cases, all effects are the same as in the case of homogeneous soil, with the resistivity of dipper layers. This conclusion is even more valid in practice, when the thickness of the first layer usually does not exceed a few meters. The distribution of current differs only close to the ground surface, while the differences decrease rapidly in the ground. Hence, the application of the proposed method enables correct calculation of the electric and magnetic fields both in the multi-layer ground and in the space between the conductor and the ground surface.

When the layer thickness is close to the penetration depth (between 10 and 50 m, for low soil resistivity values, at industrial frequency), the suggested magnetic field calculation could be applied, providing an accurate calculation of the ground return impedance. The developed method thus, represents an efficient tool for the multi-layer ground impedance calculations, in electrical power transmission and distribution systems, that include ground return, in which, ground currents have a particular significance. The most common, single line-to-ground fault case can also be successfully treated inside the multi-layer ground.

An excellent agreement between the results of analytical process presented in this paper and the numerical procedure (FEM) was achieved and discussed in this work.

References

[1] Abramowitz M. and Stegun I. A., Eds., "Bessel Functions of Integer Order" in *Handbook of Mathematical Functions*, 9th Ed. New York, NY: Dover Publications, 1970, pp. 355-433

[2] Arnautovski-Toševa V. and Grcev L., "High frequency current distribution in horizontal grounding systems in two-layer soil" in *Proc. 2003 International Symposium on Electromagnetic Compatibility*, pp. 205-208

[3] Kasas-Lazetic K., Prsa M. and Mucalica N., "Current Distribution and Resistance per Unit Length of a Two-Layer Ground" (in Serbian) in *Proc. 2010 International Conf. INFOTEH,* pp. 390-394

[4] Kasas-Lazetic K., Prsa M. and Mucalica N., “Resistance per Unit Length of Homogenous and Two-Layer Ground” (in Serbian) in *Proc. 2010 Conf. on Electrical Distribution Systems of Serbia, CIRED*, pp. 71-75

[5] Kaszás-Lažetić K., Herceg D., Djurić N. and Prša M., “Determining Low-Frequency Earth Return Impedance: A Consistent Electromagnetic Approach” *Acta Polytechnica Hungarica*, 2015, Vol. 12, No. 5, pp. 225-244

[6] Krolo I., Vujević S. and Modrić T., “Highly accurate computation of Carson formulas based on exponential approxiamtion” *Electric Power System Research*, 2018, Vol. 162, pp. 134-141

[7] Ma J., “Fast and high precision calculation of earth return mutual impedance between conductors with a multylayered soil” *COMPEL*, 2018, Vol. 37, No. 3, pp. 1214-1227

[8] Micu D. D., Czumbil L., Prsa M. and Kasas-Lazetic K., “Interfstud electromagnetic interference software - An accurate evaluation of current distribution in soil and in underground pipelines” in *Proc. 2012 International Symposium on Electromagnetic Compatibility*, pp. 1-5

[9] Mucalica D., “Calculation and measurement of characteristic parameters of MV overhead and polyethylene cable systems” M. Sc. thesis, Dept. Elect. Eng., Belgrade Univ., Belgrade, Serbia, 2000 (in Serbian)

[10] Nakagawa M. and Iwamoto K., “Earth-Return Impedance for the Multi-Layer Case” *IEEE Trans. on Power App. Syst.*, 1976, Vol. PAS 95, No. 2, pp. 671-676

[11] Olsen R. G. and Willis M. C., “A comparison of exact and quasi-static methods for evaluating grounding systems at high frequencies” *IEEE Trans. Power Del.*, 1996, Vol. 11, No. 2, pp. 1071-1080

[12] Papadopoulos T. A., Papagiannis G. K. and Labridis D. A., “Wave propagation characteristics of overhead conductors above imperfect stratified earth for a wide frequency range” *IEEE Trans. Power Del.*, 2009, Vol. 45, No. 3, pp. 1064-1067

[13] Papagiannis G. K., Tsiamitros A., Labridis D. P. and Dokopoulos P. S., “A Systematic approach to the evaluation of the influence of multilayered earth on overhead power transmission lines” *IEEE Trans. Power Del.*, 2005, Vol. 20, No. 4, pp. 2594-2601

[14] Popovic B. D., “Some basic theorems of electromagnetic field” in *Electromagnetics*, 2nd Ed. Belgrade: Gradjevinska knjiga, 1986, p. 52 (in Serbian)

[15] Satsios K. J., Labridis D. P. and Dokopoulos P. S., “The Influence of nonhomogeneous earth on the inductive interference caused to

telecommunication cables by nearby AC electric traction lines", *IEEE Trans. Power Del.,* 2000, Vol. 15, No. 3, pp. 1016-1021

[16] Tsiamitros D. A., Papagiannis G. K. and Dokopoulos P. S., "Equivalent Resistivity Approximation of Two-Layer Earth Structures for Earth return impedances Calculations" in *Proc 2005 Power Tech. IEEE Russia*, pp. 1-7

[17] Write Sh. H. and Hall C. F. (Central Station Engineers of the Westinghouse Electric Corporation), "Characteristics of overhead conductors" in *Electrical Transmission and Distribution - Reference Book*. Pennsylvania East Pittsburgh, 1950, pp. 33-64

Practical Application Possibilities for 3D Models Using Low-resolution Thermal Images

András Molnár, István Lovas*, Zsolt Domozi*

John von Neumann Faculty of Informatics,
Óbuda University, H-1034 Budapest, Bécsi út 96/B, Hungary
E-mail: molnar.andras@uni-obuda.hu

*Doctoral School of Applied Informatics and Applied Mathematics,
Óbuda University, H-1034 Budapest, Bécsi út 96/B, Hungary
E-mail: lovas.istvan@uni-obuda.hu, domozi.zsolt@phd.uni-obuda.hu

Abstract: Everyday used cheap thermal cameras can only take low-resolution images. Low-resolution images can be used as the input data of photogrammetry procedures with difficulty or cannot be used at all, as little information is stored of the actual object. Based on the little amount of information, conventional procedures are not capable of identifying a correlation between individual images. Intensity also differs between individual pixels compared to conventional RGB images, thus gradient-based solutions fail to be successful. A method has been developed, which can be used to create thermal image orthophotos from thermal images combined with RGB images. The procedure can result in several output images, from which the most important is the false color thermal image, which is within the visible light spectrum, as on it the original object and the amount of thermal radiation are both visible. Another advantage of the procedure is that not only the information of the visible light spectrum can be visualized but also the data which is invisible to the naked eye.

Keywords: thermal photogrammetry; picture fusion; data visualization; 3D objects; low-resolution

1 Introduction

Since high resolution thermal cameras are very costly devices, we examined the scenario how photogrammetric procedures can be performed using low resolution thermal camera images. In the course of the experiments several problems and errors had been found, but these were resolved. A procedure was created which resulted in that real thermal image orthophotos and 3D models can be created routinely. The dynamic range of thermal images is significantly lower compared to images taken within visible range. In Figure 1 (a), a stream and its environment can be seen. The photo was taken within a (thermal) range of 7.5-13.5 µm. It is clearly visible that the dynamic range of the image is very small. The intensity distribution function seen in

Figure 1 (a) testifies to a small dynamic range. In Figure 1 (b), we can see the RGB image of the same area at the same time and in the same position. Based on intensity distribution functions it is clearly visible that the dynamic range of individual color channels is much wider compared to the IR image. For the human eye it has more contrast, thus we gain a more satisfactory image.

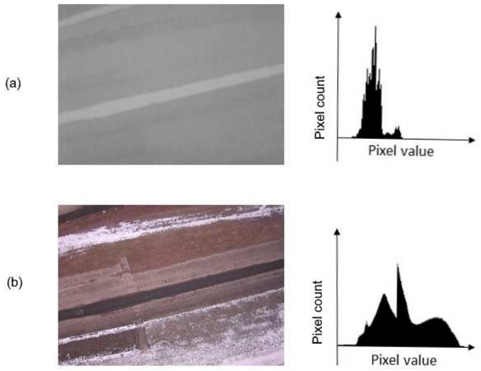


Figure 1
Grey-scale thermo photo with histogram

Low dynamic range images (Figure 1 (a)) not only play a role in the subjective assessment of the image but considering its information content it is of lesser value compared to color images (Figure 1 (b)). In images containing less information it is more difficult to find clearly distinct pixels, thus the efficiency of determining point pairs between the images significantly declines.

The gray image (Figure 1 (a)) shows the thermal image of a given area with a bit depth 14 (resolution of 1 pixel intensity). In principle, this would mean that the detail of the mapped area (the distinctiveness of the individual pixels) is sufficiently high, i.e. we get a good contrast image. Next to the photo is a histogram of the image that shows how much of the available 14-bit resolution has been used by the camera. Ideally, the entire 14-bit representation range would be covered, but the histogram shows that only less than half of the available range contains data, i.e., a real pixel intensity value. In other words, the image is strongly ”underexposed.” In the subjective formulation, we feel the same as light and obscure (weak contrast).

In the color image (Figure 1 (b)) shows the RGB image of the area shown in Figure 1 (a). The two recordings were made from the same position at the same time. The grayscale conversion histogram of a color image is a good illustration of the wider range of distribution of each pixel value. Using the above terminology, this means a more detailed, contrasting picture.

Beyond subjective judgment, it can be seen that when searching for point pairs between images, algorithms can use more information for color images or 8-bit grayscale images formed from them than for a higher bit depth but still information-deficient thermal image.

Processability fundamentally does not depend on the subjective assessment of the images, but on its information content. The pixels of the RGB images applied during the experiments were of 3x8 bit resolution. In case of thermal images this resolution was 1x14 bit. As the resolution of the data describing a sole pixel of the RGB image used is substantially higher (24 bit), than the resolution of a sole pixel of the thermal image (14 bit), it makes sense why photogrammetric processing based solely on thermal images was less efficient.

The images seen in Figure 1 and Figure 2 were taken at the same time and in the same position. The grayscale conversion of the image of Figure 1, taken using an RGB camera, can be seen. The images represent a bridge spanning across a stream. In Figure 2, we can see an image taken using an IR camera which represents the object identical to Figure 2. The figures represent two images which were photographed in different positions. [1]

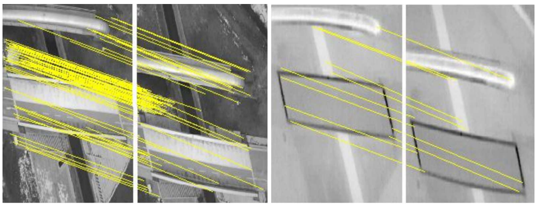


Figure 2
Point pair detection applied on grayscale image pairs using the SIFT algorithm. [2] [3]
(left: grayscale from RGB, right: grayscale from IR image)

Figure 2 represents the point pairs detected by the SIFT [4] algorithm. In both cases the parameters of the algorithm were identical. It is apparent that in the grayscale image converted from the RGB image the algorithm found many point pairs, whereas in IR images it found substantially fewer. During the photogrammetric process, these point pairs will supply spatial points. In case of a sufficiently numerous number of point pairs the spatial point cloud will be dense, thus the orthophoto created from it will be rich in detail. In case of few point pairs the point cloud will be very sparse. This, practically, is not sufficient for further processing.

In case of input images containing less information, the photogrammetric process either does not provide a result or provides one with many errors (Figure 3). Photogrammetric processing was done using Agisoft Photoscan. It is visible from the result that during the creation of the point cloud the software made three individual objects which are in fact three sections of a singular object.

Based on the facts introduced above, the photogrammetric procedure not only demands high resolution overlapping images, but it also demands the adequate dynamic of the images. If the pixels possess the necessary amount of information (rich in detail, high contrast, high dynamic range images), and resulting from their resolution many point pairs can be localized in the overlapping areas, successful photogrammetric processing can be expected. With respect to this expectation, we planned the structure of the connection between the color image and the thermal image taken at the same time and in the same location. Based on the aboves this

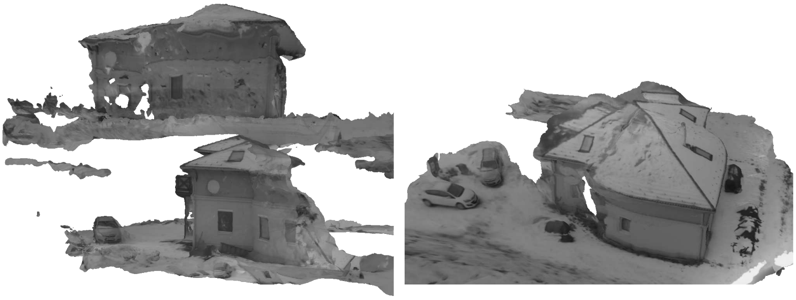


Figure 3
The result of faulty photogrammetric processing

method can be used efficiently on any object which is differ from buildings [5] but it has heat emission like car-parts etc.

The method discussed in this paper is a type of structure from motion and multi-view stereo (SfM-MVS) technique. Essentially, the object under examination is either surrounded by several cameras or they go around it with a single device. The device can be fitted with several tightly or loosely fitting image capture devices to meet special needs. With tight-fitting, the cameras record information received from one other. With loose-fitting, they work independently, and subsequently, during the processing, it is possible to combine and further examine the individual images. The integration of an ordinary camera and a low-resolution thermal imaging camera into one device makes it possible to take high-resolution thermal images of objects using the SfM technique. Our device also produces an RGB and IR image of the area with one shot. These images are not identical, calibration is required.

2 Materials and Methods

With conventional photography, distortions due to the structure of the lens are usually not considered. However, with 3D reconstruction, the determination of the parameters of the cameras is essential. The lack of parameters would result in inaccuracy during processing. Two parameters are distinguished; intrinsic and extrinsic parameters. Intrinsic parameters describe the physical characteristics of the camera. These values are independent of the spatial position, always taking a fixed value. It is therefore sufficient to calculate them once during calibration. It is usually given in matrix form:

$$K = \begin{bmatrix} f_x & \alpha & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where f_x, f_y are focal lengths and α describes the angle between x and y axes.

Another problem can be the radial and tangential distortion caused by the glass lenses in the camera lens. It depends on these parameters, for example, whether the edges of an object, in reality consisting of straight lines, become curved (radial distortion). Extrinsic parameters describe the relationship between the camera system and the world coordinate system, by a rotation matrix R and a translation vector T.

With an RGB camera, the calibration procedure can be performed using a printed chessboard [6]. The parameters described above can be determined by detecting the corner points of the squares with regular, known dimensions on the chessboard, however, in the case of a thermal imaging camera, this method cannot be used with a paper chessboard. The simple reason for this is that the thermal imaging camera cannot distinguish black and white colors.

A special chessboard was made to calibrate the thermal imaging camera. The black squares of the chessboard are made up of aluminium leaves with good thermal conductivity, which were glued to a white surface with poor thermal conductivity. Before taking the calibration images, the chessboard was heated using a heat source for a few minutes. The result can be seen in the Figure 4.

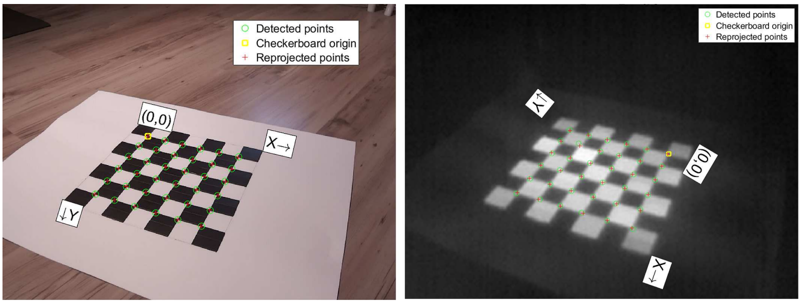


Figure 4
The calibration itself was performed using the Matlab Single Camera Calibrator App

During calibration, the goal is to determine the lens parameters of two different cameras. The image on the left shows the calibration of a color camera, while the image on the right shows the calibration of a thermal camera.

The basis of the procedure created is to take advantage of the connection between the RGB and the IR image. The photogrammetric procedure is performed on the RGB image but the transformations calculated during the procedure will also be performed on the IR image. This is only possible if the two images overlap, in other words, they contain the same area [7]. In this phase, where IR and RGB images reperesenting the same areas, images can be considered as two different data source [8].

As the overlap of RGB and IR images is not ensured, the first phase of our procedure is to create this [9], [10]. Considering the fact that the IR image contains less area than the RGB image, in other words, the content of the IR image is the subset of the RGB image, overlap transformations are performed based on the parameters of the

IR image. In this procedure, the relative fixed position of the two cameras is taken advantage of. This means that we perform magnification on the IR image based not on image content but on values determined during previous measurements, then, similarly we perform predetermined cropping on the RGB image. As the result of the procedures, RGB-IR image pairs are created which are identical in pixel size and content too. Further conversion is performed on the IR image. The original IR image is a 14 bit one, which with conversion is converted to an 8 bit one. This conversion results in data loss. If the aim is to display relative temperature conditions on the orthophoto or on the 3D model, then this data loss does not cause a substantial error.

Of course, radiometric data vanish as well which means that temperature measurement will not be possible on the processed end result. If data loss is to be avoided, it is possible to extend the 14 bit thermal image to a 24 bit one (2x8 bit): the thermal image is 2x7 bit, which is abrased on 2x8 bit after the conversation. The 2x8 bit thermal image information in this case is stored on the green (G) and the blue (B) color channels of the merged image.

With the second step, the RGB image is converted into a grayscale, but color image. This means that all three color channels of the color image contain identical pixel information. The data of the blue (B) color channel of the grey image created this way is replaced with the data of the IR image. As a result, a merged image is created whose one color channel (R) contains the grey image, the other color channel (B) contains the thermal image. The green color channel (G) is not used. [11]

The specially created merged image is suitable for photogrammetric processing. During our experiments, the Agisoft Photoscan software was used. The 3D models or the orthophotos created were subjected to post processing.

During post processing, using the software created, in case of the 3D model we converted the texture images and in case of the orthophoto, the orthophoto itself. Basically, the full color of the image was replaced based on the blue color channel (as this contains the temperature information of the given pixel),. Thus, an artificially colored thermal image is created where the color of the pixel is proportionate to the radiation temperature of the original surface. The recoloring program provides an opportunity for the modification of the histogram of the thermal image for the sake of providing a better visual experience. By modifying the histogram a better contrast can be ensured, also small-scale temperature differences can be made visible for the human eye. It is true however, that this phase of post processing does not provide additional information, the visual experience of the result significantly improves.

During the transformations of visible light (RGB) and thermal camera images, rotation was not used. The reason for this is that we assume the exposition of the two cameras occurring at an identical time as well as their relative fixed position. In case of a slowly moving camera system this condition is usually fulfilled. At quick moves (mainly when rotating the camera) however the sync error of the thermal camera and the visible light range camera results in serious deviations regarding the content of the fixed image pairs. Figure 5 represents the image pair “lagging” in time and its merged result. The bottom left image represents the thermal image

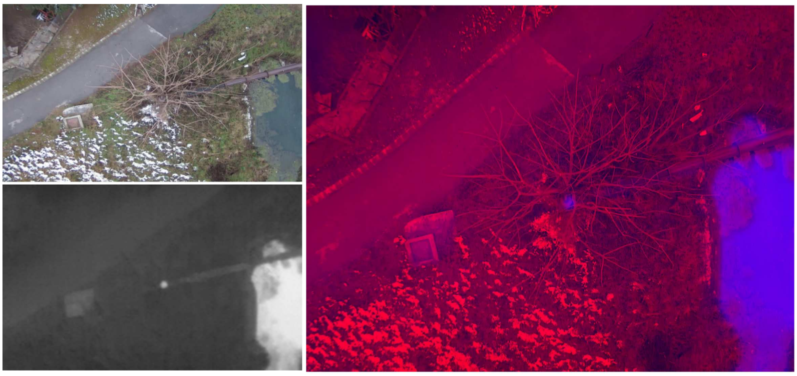


Figure 5
Exposition asynchronicity error

displayed in grayscale, the top left image illustrates the color image of the same area. The large image of Figure 5 is the merged image of the above illustrated image pair. It is visible in the original image pair too that image contents do not match. The result can be observed in the merge image too. As the blue channel of the merged image contains the information of the thermal image the “ghost image” of the source and the spring appears. The image was taken just at the time of the camera’s rotation; in this way not only translation but also rotation deviation can be observed too. Such images are to be excluded from post processing. Although it is obvious that mismatched images are not suitable for further processing when shooting at high speed moving, there may be a difference in rotation between the two images.

For the sake of improving visual experience the already prepared temperature colored models (3D model or orthophoto) were improved using edge enhancements. The edges were displayed in the models with white color. This is clearly distinct from the colors containing temperature information and with defined contours it provides a more easily interpretable visual experience. The detection and the enhancement of the edges were done using gradient-based Mamdani fuzzy logic. [12]

As a first step, as it is conventional for edge detecting algorithms, the gradient vector (2) of the grayscale image was determined:

$$\nabla f = \left[\frac{\partial f}{\partial x}\,\frac{\partial f}{\partial y}\right]^T \tag{2}$$

In case of digital images, partial derivatives can be approximated with finite difference, with the following general differential equation:

$$\frac{\partial f}{\partial x} = \lim_{\varepsilon\to 0}\left(\frac{f(x+\varepsilon,y)}{\varepsilon} - \frac{f(x,y)}{\varepsilon}\right) \approx \frac{f(x_{n+1},y)-f(x_n,y)}{\triangle x} \tag{3}$$

The derivatives can be defined using a convolution mask in X and Y directions. There are a number of possibilities for a convolution mask.

The establishment of a rule base took place in accordance with the following two GMP (Generalized Modus Ponens) conclusions, which contain two antecedents and one consequence:

- ”If Ix is zero and Iy is zero then $Iout$ is white”;
- ”If Ix is not zero or Iy is not zero then $Iout$ is black”;

where Ix is the value of the gradient belonging to axis X, while Iy is the value of the gradient belonging to axis Y.

Membership functions are continuous, trapezoid-shaped functions as seen on Figure 6. For the fuzzy set defuzzification of the conclusion gained as a result, a method of center of gravity location was used.

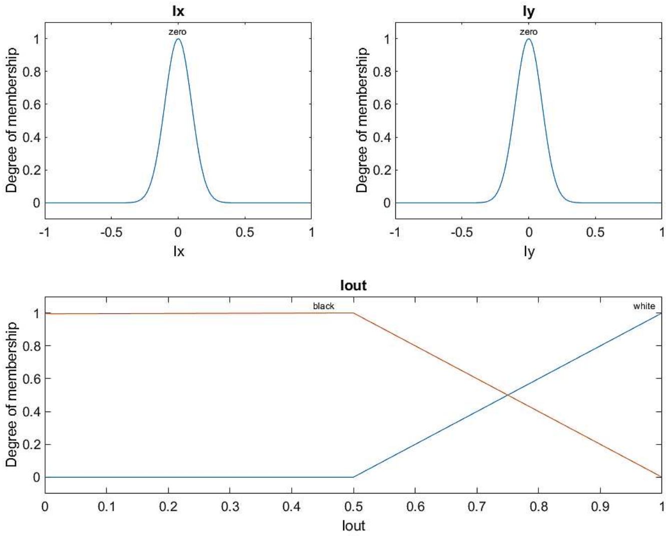


Figure 6
The membership functions of the inputs and the outputs.

The result of merging the image gained after edge enhancement and the palletized thermal image can be seen in Figure 7.

Many different resolution enhancement methods are known, such as RCNN, video stream analysis, or high-resolution image built from time-series images. The problem is that each requires the creation of multiple images, which in many cases is not possible or makes it unnecessarily complicated to produce a high-resolution image.

The presented solution is able to create a high-resolution thermo image from a single image. The procedures presented by others are not capable of this case.

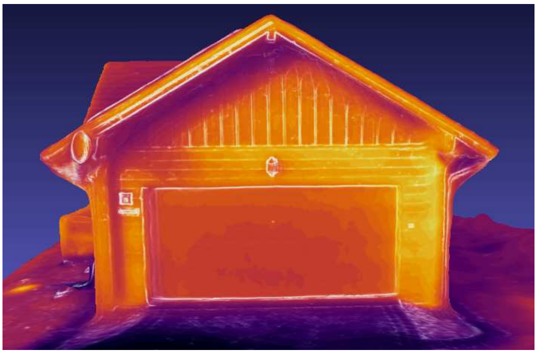


Figure 7
Contoured thermal image produced using fuzzy-based edge detection.

3 Results

By utilizing the connection between the color and thermal image taken of the object it became possible to connect the photos in good quality using the photogrammetric method. During the procedure, we created cuts with content identical to the thermal images, which were converted to grayscale images in the next phase. The respective thermal image was connected to the grey scale image using one of the color channels. Thus, the photogrammetric method was performed on specially-made, merged images. At the end of the process, texturing was conducted on the basis of thermal images. [13] In Figure 8, the 3D model of an apartment block can be seen, which was produced using photogrammetry. Grayscale images were used for texturing. 1920x1080 pixel resolution images serve as the basis of the procedure. Texturing was done based on the data of the 3D model, supplied by a thermal camera, which was produced using the procedure described above. As a result, a high quality 3D thermal model was produced. The resolution of the images of the thermal camera is altogether 160x120 pixel. In these images it is not possible to detect a sufficient number of high quality feature points, thus they cannot be connected using the photogrammetric method.

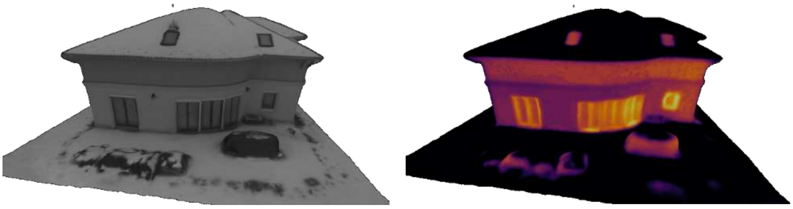


Figure 8
Grayscale and thermal 3D reconstruction of a residential home.

At the photo shoot of the building, the temperature of the environment was -5 °C. As during the time of the photo shoot the weather was sunny in the few hours before that the camera detected not only the radiation resulting from the temperature loss of the house, but also the reflection resulting from the sun's radiation, as well as the back-radiation of its absorbed energy in the form of thermal radiation on the detected thermal image [14].

This is clearly visible on the black objects in front of the building (Figure 8). In the thermal image (Figure 8), these are warm, though these are not heated objects, only garden furniture covered with black foil protecting them from the weather.

At the border of Dunaalmás, a natural thermal hot spring can be found. Water of 32 °C comes rushing from the depths of the earth which flows into the Danube not far from the spring. 1200 image pairs (RGB and IR) have been created of the area of the spring. The IR orthophoto seen in Figure 9 with 1.5 cm/pixel field resolution was made using these images. The lightest area of the image can be clearly seen, where the 32 °C water of the spring rushes to the surface. From this point, the warm spring water flows to the right in the picture. The track of the warm stream looks a little colder in the orthophoto, but in reality, the water in the photographed area does not cool down. Plants are growing on the surface of the stream water and these "shade" surface thermal radiation. During the time of the photo shoot, air temperature was 1 °C.

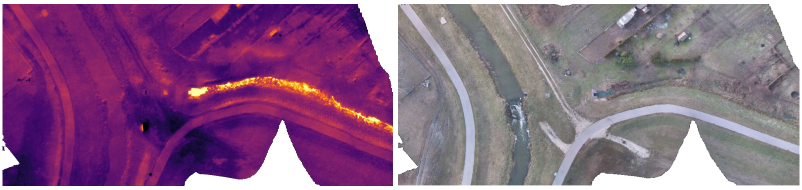


Figure 9
The thermal orthophoto and the color ortophoto of the 32 °C thermal spring (Dunaalmás)

The color (RGB) orthophoto of the area of the spring can be seen in Figure 9. The spring and the original stream separated from it by a dam can be clearly seen. In the color photo which is natural to the human eye there is not any difference between the two water surfaces, but by comparing the RGB and the IR othophotos, the two water surfaces immediately become clearly distinct.

The colored model of a wooden house can be seen in Figure 10. The model creation procedure was performed in the manner similar to the stone house introduced in Figure 8. It is noticeable in the image that the contours are blurred; regarding the visual experience the image is rather disturbing for the eye. As the observed wooden house was not heated, its thermal image results from the thermal radiation arising from the absorption of the external radiation, as well as from the reflection of the immediate thermal radiation. Hence, the walls are warmer than the windows of the house.

In Figure 10, the 3D model of the wooden house was supplied with contours during post processing. The edges highlighted with white color carry little additional information. In this case, the solar panels engineered on the house become visible during the application of contours.

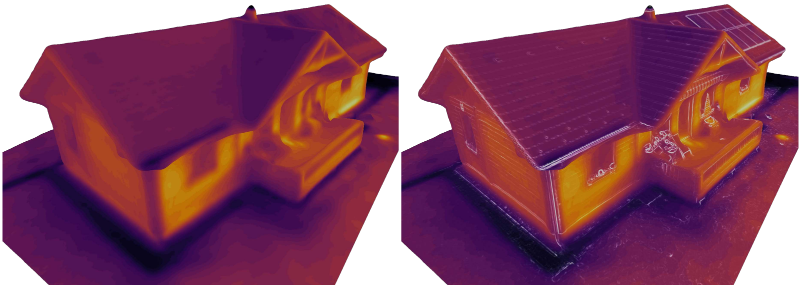


Figure 10
Thermal 3D reconstruction and contoured thermal 3D reconstruction of a wooden house

The model supplied with contours however looks more contrasted; it offers a more pleasant view to the eye. The application of edge enhancement therefore provides a more easily interpretable result in 3D thermal images.

We used Fuzzy edge enhancement, because using Matlab to enhance the visual experience on our data, this was the best procedure.

4 Discussion

A 3D thermal image is capable of the thorough examination or inspection of vehicles. Based on the thermal image, the thermal raditation of a passenger car with a closed passenger compartment can be examined. The photos help in the optimization of the planning of the heating system, particularly in the planning of cooling/heating installation. In case of providing sufficient external conditions, satisfactory or just the insufficient heating of the given glass surfaces can be examined (windscreen, side windows). The operation of further heating devices can be examined similarly, by using a thermal image, like for example the heating of side mirrors (Figure 11) or that of the windscreen.

In Figure 11, the cold air inlet of the engine becomes obvious. Based on the image, it can be seen that the cold air inlet is independent in its total cross-section, in other words an obstruction negatively influencing the cooling of the engine is not visible.

Conclusions

On the basis of the connection between the thermal and RGB images created of the object, joining images was made possible using photogrammetry. The procedure has created cuts from RGB images equaling the size of the thermal images, and then these images have been converted to grayscale. The procedure used one of the channels of the grayscale image to visualize the thermal image, in this way the details of individual areas of the object and its thermal information became visible. In such images, parts suitable for identification can be found now, based on which the joining has taken place, even by using the photogrammetric method.

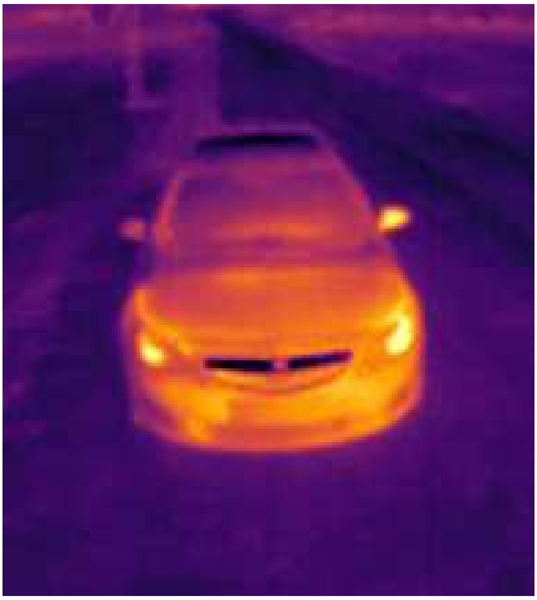


Figure 11
The thermal image of the passenger car in which the heated side mirror on the driver's side can be seen

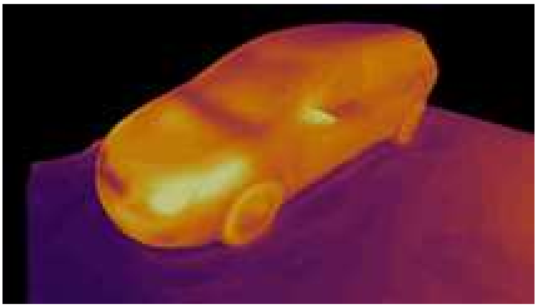


Figure 12
The 3D thermal image of the passenger car

The efficiency of the procedure has been examined on the 3D model of a building and a personal car too. As the basis for testing efficiency, 1920x1080px resolution images were used. Using the above-detailed procedure, texturing took place with the produced images, thus a high-resolution 3D model was created, whose thermal information was gained from low-resolution images, as the resolution of the thermal camera is 160x120px. All elements of the procedure developed have been accomplished using the program Matlab. The Fuzzy edge enhancement implemented also in program Matlab.

References

[1] G Kertész, S Szénási, Z Vámossy, "Multi-directional image projections with fixed resolution for object matching", Acta Polytechnica Hungarica 15 (2), 211-229, 2018.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60, 2 pp. 91-110. 2004.

[3] D. G. Lowe, https://www.cs.ubc.ca/ lowe/keypoints/, 2018.

[4] D. G. Lowe, "Object recognition from local scale-invariant features", Proceedings of the Seventh IEEE International Conference on Computer Vision, 2/8, p. 1150–1157, 1999.

[5] A. Molnar, I. Lovas, Z. Domozi, "Photogrammetry on low resolution thermal pictures", in P. Iványi, B.H.V. Topping, (Editors), "Proceedings of the Sixth International Conference on Parallel, Distributed, GPU and Cloud Computing for Engineering", Civil-Comp Press, Stirlingshire, UK, Paper 30, 2019. doi:10.4203/ccp.112.30

[6] Zhang, Zhengyou. "A Flexible New Technique for Camera Calibration." IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000): 1330-1334.

[7] S. Vidas, P. Moghadam, M. Bosse, "3D thermal mapping of building interiors using an RGB-D and thermal camera", 2013 IEEE International Conference on Robotics and Automation, 2013.

[8] Balla, D., Zichar, M., Kozics, A., Mester, T., Mikita, T., Incze, J., Novák, T. J. (2019). A GIS Tool to Express Soil Naturalness Grades and Geovisualization of Results on Tokaj Nagy-Hill. Acta Polytechnica Hungarica, 16(6).

[9] X. ZHAO, J. HE, Y. LUO, N. HUANG and Y. NI, "Analysis of the Thermal Environment in Pedestrian Space Using 3D Thermography Generated With Unmanned Aerial Vehicles and Infrared Cameras," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 4328-4331, doi: 10.1109/IGARSS.2019.8898994.

[10] H. Jung, J. Lyou, "Matching of thermal and color images with application to power distribution line fault detection", Control Automation and Systems (ICCAS) 2015 15th International Conference on, pp. 1389-1392, 2015.

[11] Dorit Borrmann, Florian Leutert, Klaus Schilling, Andreas Nüchter, "Spatial projection of thermal data for visual inspection", Control Automation Robotics and Vision (ICARCV) 2016 14th International Conference on, pp. 1-6, 2016.

[12] M. Takacs, "Mamdani-type Implication Inference with Degree of Coincidence", (2003) 1st Slovakian - Hungarian Symposium on Applied Machine Intelligence (SAMI 2003), Herlany, Slovakia, February 12-14, ISBN. 9647154140, pp.67-73., 2003.

[13] M. Scaioni, E. Rosina, L. Barazzetti, M. Previtali, V. Redaelli, ”High-resolution texturing of building facades with thermal images”, Proc. SPIE 8354, Thermosense: Thermal Infrared Applications XXXIV, 83540I (18 May 2012); doi: 10.1117/12.920613;

[14] M.C. Harvey, J.V. Rowland, K.M. Luketina, ”Drone with thermal infrared camera provides high resolution georeferenced imagery of the Waikite geothermal area, New Zealand”, Journal of Volcanology and Geothermal Research, Volume 325, Pages 61-69, 2016.

Three-Dimensional Modeling and Analysis of Mechanized Excavation for Tunnel Boring Machines

Danial Mohammadzadeh S. [1,2,3,4]**, Nader Karballaeezadeh** [5]**, Amirhossein Sanaei Zahed** [4,6]**, Amir Mosavi** [7*]**, Felde Imre** [7]

[1] Department of Civil Engineering, Ferdowsi University of Mashhad, Mashhad, University street 1, P.O. BOX 9177948974, Iran
danial.mohammadzadehshadmehri@mail.um.ac.ir

[2] Department of Civil Engineering, Mashhad Branch, Islamic Azad University, Mashhad, University street 1, P.O. BOX 9187147578, Iran

[3] Department of Civil Engineering, Faculty of Montazeri, Khorasan Razavi Branch, Technical and Vocational University (TVU), Mashhad, University street 1, P.O. BOX 9176994594, Iran

[4] Department of Elite Relations with Industries, Khorasan Construction Engineering Organization, Mashhad, University street 1, P.O. BOX 9185816744, Iran

[5] Faculty of Civil Engineering, Shahrood University of Technology, Shahrood, University blvd. 1, P.O. BOX 3619995161, Iran
N.karballaeezadeh@shahroodut.ac.ir

[6] Toos Instite of Higher Education, Khorasan Razavi, Mashhad, University street 1, P.O. BOX 9188911111, Iran, Ah.sanaei@toos.ac.ir

[7] John von Neumann Faculty of Informatics, Óbuda University, Bécsi út 96/b, H-1034 Budapest, Hungary, felde@uni-obuda.hu, amir.mosavi@nik.uni-obuda.hu

Abstract: Urban train infrastructures are very important for reliable urban mobility. This paper proposes a three-dimensional modeling of mechanized drilling corridors. Drilling in urban areas is always a risky and complex project. One of the most important issues during the construction of subway tunnels is the investigation of the impact of drilling steps on the ground subsidence and impact on existing structures. For this purpose, different types of mechanized drilling methods are often used, resulting in a considerable reduction in the displacements caused by tunnel drilling. In this study, part of the route of an urban train tunnel, that passes under a traffic interchange, is examined. The shear strength capacity of the slab pile was calculated, using the relevant equations, and then, the modeling of the soil mass was performed, using the PLAXIS 3D finite element program. The proposed depth of the tunnel construction, by the consulting company, is 18 meters. Due to drilling problems, *a depth of 14 meters has been suggested as an alternative. Analysis of both the depths of 14 and 18 meters, showed that the displacements at both depths, were approximately the same. However, the impact of the tunnel, on the capacity of the piles' tip, at a depth of 18 meters, is greater than at the depth of 14 meters. Thus, the suggested optimum depth is 14 meters, which is more suitable, than the initial suggested depth of 18 meters.*

Keywords: Tunnel; mechanized drilling; optimization; urban train lines; computational mechanics; smart cities;PLAXIS 3D; numerical simulation; finite element simulation

1 Introduction

Urbanization and urban development have necessitated the need for effective public transportation systems [1-5]. The urban train network is one of the most important transportation systems in a city [6-9]. The construction of the train network above ground is less costly, but, because of land restriction and increasing surface congestion, the underground train network is more preferable [10]. For the construction of underground tunnels in urban areas, engineers often perform excavation operations near underground services, cultural heritage monuments and residential/commercial buildings. The prediction of the tunneling-induced settlement and the related impact on existing structures help engineers to estimate potential damage [11]. Due to the low depth of these underground tunnels, train stations are usually built on soft soils. The construction of tunnels in soft soils results in soil movement. This issue could lead to the instability of the integrity and damage to existing structures [12]. Thus, the optimal implementation of these underground spaces and ensuring their security during the long-term construction process, is a factor that has been taken into account, by designers of underground structures. To decrease these movements, engineers utilize Tunnel Boring Machines (TBM) for the creation of tunnels in urban areas. Because of temporary supports and face pressure, the TBM diminishes soil disturbance, due to tunneling, providing protection to existing structures [13-15].

Evaluation of tunnel construction using TBM and its impact on the soil movement requires 3D soil-structure interaction modeling. Due to limitations of the analytical approaches and also, the development of computer coding, the use, by engineers, is increasing. Muniz de Farias et al. and Negro and Queiroz summarized the finite element models used for tunneling studies before 2000 [16, 17]. They showed that the most popular approach is the finite element method (FEM). The numerical mechanized tunneling modeling aims to take into consideration of processes that take place during tunnel excavation. In order to take into account all considerations, a three-dimensional numerical model should be used. Nowadays, software packages such as PLAXIS 3D [18-20], Abaqus [21, 22], and FLAC 3D [23-28] are normally used for 3D analysis. As a main aspect of the tunnel excavation, the behavior of the ground must be taken into account.

Therefore, a realistic model of the ground is essential in determining the displacement and stresses of the ground. In this study, 3D modeling of TBM, crossing under Mianrood bridge, Shiraz metro line 2, in Iran, is performed using the PLAXIS software. The main goal is to determine the optimum depth for tunneling. Plaxis is a powerful software for tunnel analysis. The results are presented as displacement and stress contours, together with the curves of bending moment, shear force and axial force. This study is very important because drilling, in particular in urban areas, is sensitive and requires great precision. The effect of drilling operations on ground surface settling is one of the most critical concerns, during the construction of metro tunnels. This paper is organized into four sections: Section 1 introduces the work, Section 2 describes the case study and methodology, Section 3 shows the modeling results and discussion and finally, Section 4 provides the relevant conclusions.

2 Materials and Methods

2.1 Materials

Similar to many metropolitan cities in Iran, Shiraz is known as a tourist hub. Therefore, it needs a subway network to reduce urban congestion. Metro line 2 of Shiraz has a length of approximately 14 kilometers, comprising of 13 stations. Figure 1 shows this metro line. According to geotechnical studies, tunnel excavation from Ghahramanan Station to Azadi Station was designed by earth pressure balance (EPB) and TBM machines, with a diameter of 6.88 m, in two twin tunnels. TBM drilling in the soil is always associated with sedimentation. Therefore, controlling the possible displacements and settlements for an underground structure is a critical parameter in project management, especially the project in which both TBMs must pass under existing vital structures. One of these structures is Mianrood Bridge. This bridge has frictional piles. Mianrood Bridge is located on the Ring Road of Shiraz. The bridge piles are frictional. The slab of the middle bases has dimensions of 6.8 x 16.4 meters. This slab is constructed on 8 piles with a diameter of 1.2 meters and a length of 25 meters. The side slabs are 6.8 x 17.43 meters and are located on 10 piles with a diameter of 1.2 meters and a length of 27 meters. Reducing displacements of ground level and pile slabs should be considered in the tunnel route design under this bridge. Reducing impacts on the bearing capacity of the piles is another important issue that should be considered. Based on the designed tunnel profile by the project consultant, the depth of the tunnel under the Mianrood Bridge is 18.1 m. The distance between the tunnel wall and the side piles of the bridge is approximately 5 meters. The tunnel wall is far from the middle piles of the bridge about 3 meters. According to Geotechnical studies, the soil of this area, up to the depth of 29 meters, is made of lean clay. At the depth of 29 to 30 meters, there is a middle layer of Silty Sand. Also, the depth of the groundwater level in this area, is about 4.75 meters. In this study, 3D modeling of the TBM crossing under the foundations of the bridge, was performed at two depths of 14 meters (authors' recommendation) and 18 meters (project consultant recommendation).

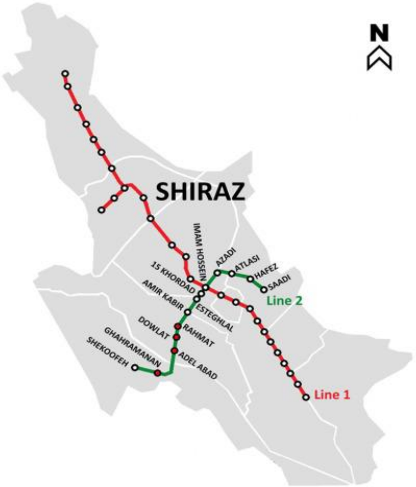


Figure 1

Shiraz metro line 2

2.2 Analysis of the Tunnel Settlement due to Drilling TBM and the Interaction Pile Base and Tunnel

The effective factors in the tunneling settlements are divided into three zones, as shown in Figure 2. which is adapted from [29].

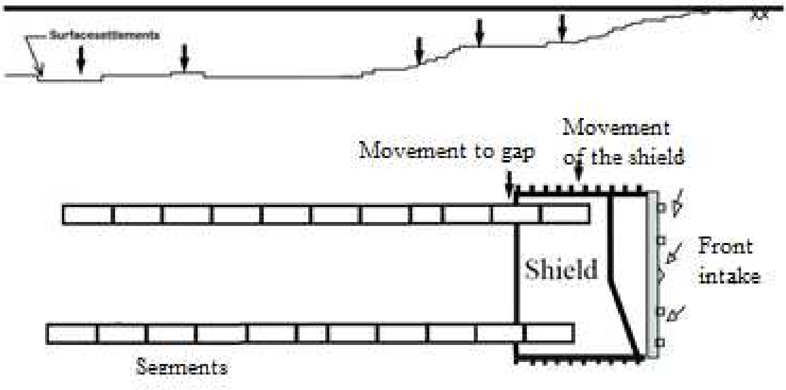


Figure 2

Various settlement Areas Created in Tunneling by the Shield Method

Zone 1

The settlement was caused by displacements created at the tunnel work front. If the pressure of the tunnel work front is wrong, or the operator does not control correctly, and the volume of soil exited from the work front exceeds the desired soil volume then, displacements at the work front may continue until ground level. Occasionally, the low pressure of the work front will overcome the water and soil pressure and the work front falls to the shield. Thus, after the shield passes through the area, the soil will be stressed and displaced and will settle again. Of course, high pressure at the work front, will lead to a severe depreciation of the cutting tools and related devices [30].

Zone 2

This area is within the shield length range (Figure 2). Usually, for easier movement of the shield, the drilling diameter is 1 or 3 cm larger than the outer diameter of the shield ends. They also form a spindle metal cylinder that reduces the front and rear friction of shields. For this reason, this area of several meters around the shield and the movement of the shield and its impact on ground displacements, can cause settlement in this area. Bentonite slurry injected around the shield during drilling, can be used to prevent this subsidence and to prevent soil shield friction [30].

Zone 3

Due to the difference between the outer diameter of the concrete rings (6.6 m in this project) and the drilling diameter (about 6.88 m), there is a gap between these rings and the soil. This gap is filled by the grout slurry (Figure 2). The subsidence in this region depends on the geological, resistive and grouting properties. By summing this subsidence and subsidence of zones 1 and 2, the whole subsidence of EPB Shield Tunneling is calculated [30].

2.3 Interaction Analysis of Tunnel and Pile

In general, the interaction between piles and tunnels has been studied in various studies. The basis of these studies is the depth of the tunnel, the depth of the piles, the horizontal distance of the tunnel to the pile and the effect of tunnel drilling interaction on pile displacement. Accordingly, there are three areas [31] as follows. A) Deep tunneling: where the pile tip is located above and close to the tunnel, B) Shallow tunneling: where the pile tip is located above and some distance from the side of the tunnel, and C) Shallow tunneling: where the pile tip is located below the zone of ground movement. In modes A and B, the tunnel is below the tip of the pile. In the case of A, the tunnel excavation settlement and the displacement affect the wall friction and the bearing capacity of the tip of the pile. Still, in some cases, this effect can cause pile failure. In the case of B, the impact radius of the tunnel displacement has less effect on the pile, and the pile is not located in the critical area. Thus, in these two cases, the greater the horizontal distance of the tunnel from this pile the less impact it has on the piles. In the case of C, the tunnel can affect the bearing capacity of the wall and tip of the pile. The closer the horizontal distance of the tunnel to the pile, the greater the impact, and the closer the tunnel to the pile tip, the greater the impact on the bearing capacity of the pile tip and the greater the bending moment on the pile. In most optimum case of C, the tunnel should have a more horizontal distance from the pile, and the depth of the tunnel should be chosen, so as, it has the least impact on the bearing of the capacity of the pile tip and also has the least settlement and displacement on the pile head and ground [31]. Figure 3 shows a schematic of the effects of the tunnel impact zone and the ratio of pile head displacement to ground displacement due to tunnel excavation and subsidence phenomena. In this figure, if the piles are in area A and above the tunnel impact surface, it is possible to move the pile head further than the ground surface displacement, above the tunnel (R>1). In area B, the displacement is equal (R=1). And in area C, pile head displacement is less than the ground level displacement above the tunnel (R<1) [32]. Further details available in Figure 3 which is adapted from [31].

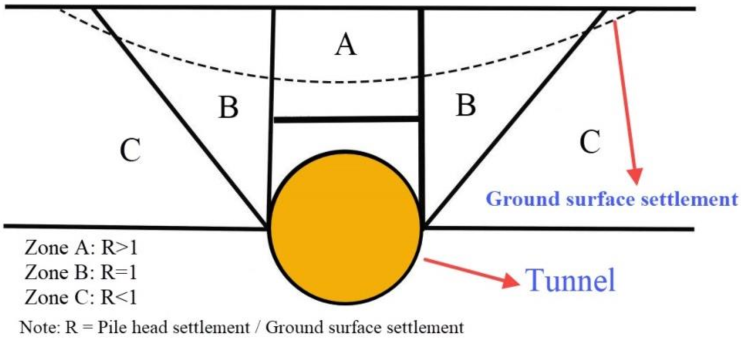

A
B
B
C
C
Ground surface settlement
Zone A: R>1
Zone B: R=1
Zone C: R<1
Tunnel
Note: R = Pile head settlement / Ground surface settlement

Figure 3
The Tunnel impact areas on the ground and piles settlement

2.4 Floor Slabs Load

For 3D modeling and tunnel front pressure determination, load on floor slabs is calculated by the reverse method. In the reverse method, considering the diameter and length of each pile and geotechnical characteristics of the soil, it can be to estimate the ultimate and the permissible bearing capacity of piles. This bearing capacity helps calculate the computational distributed load on each slab. The ultimate bearing capacity of frictional piles is obtained from the sum of the final bearing capacity in the wall and tip of the pile [33].

2.4.1 Ultimate Bearing Capacity of Pile Tip

Generally, the ultimate bearing capacity of the pile tip is calculated according to Equation (1) [33].

$$Q_{up} = A_p(N_c^* C_{ub} + 50N_q^* \tan \phi) \quad (1)$$

where Q_{up} is the ultimate bearing capacity of pile tip (Kg). A_p is the area of pile tip in squared meter (for Mianrood bridge, the diameter is 1.2 meters). C_{ub} is the undrained shear strength of soil in the pile tip which according to geotechnical studies of this area is 75 Kpa. N^*_q and N^*_c are bearing capacity Factors that depend on the internal friction angle of the soil. Considering Figure 4 adapted from [34] where the internal friction angle of the site which is approximately 26 degrees, N^*_q and N^*_c are 25 and 60, respectively. The bearing capacity of the pile tip, in this study, is equal to:

Q_{up}=1.13×(60×75 + 50×25 tan26) =5774 KN

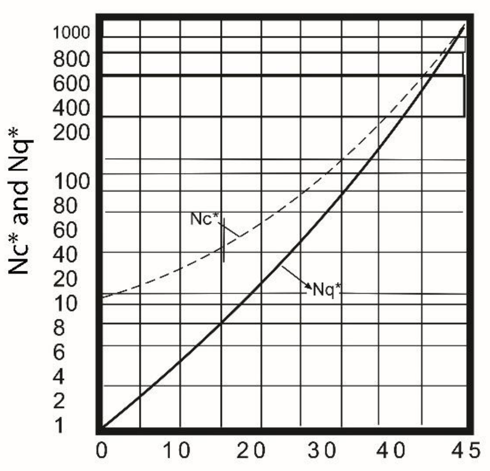


Figure 4
Variations of $N_c^*$and N_q^*values versus internal friction angle

2.4.2 Ultimate Bearing Capacity of Pile Wall

Calculating the ultimate bearing capacity of the pile wall, in an undrained condition, is performed by the alpha (α) method. This method assumes that the loading behavior of piles in low permeable clays, is similar to piles in undrained soils. The ultimate bearing capacity of the pile wall, in an undrained condition according to [33] is obtained from the Equation 2.

$$Q_{us} = A_s \alpha C_u \quad (2)$$

where, Q_{us} is the ultimate bearing capacity of the wall pile (Kg), is the pile wall area in the squared meter (πDL). C_{ub} is the undrained shear strength of the soil in the pile wall, which, according to geotechnical studies, is equal to 50 kPa. The α value is an experimental factor that decreases the adhesion of the shaft wall soil. This coefficient is less than one. Based on Figure 5, the value of this coefficient is 0.6 adapted from [34].

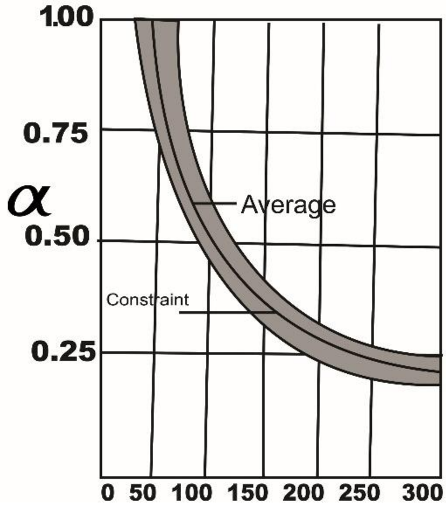


Figure 5
The graph of C_u (KN/m^2) versus α

Due to the geometry of the bridge and the designed piles, the piles of the middle slab are 25 meters long and the side slabs are 27 meters long.

$$Q_{us(25m)} = 25 \times 1.2 \times 3.14 \times 0.6 \times 75 = 4239KN$$

$$Q_{us(27m)} = 27 \times 1.2 \times 3.14 \times 0.6 \times 75 = 4578KN$$

By determining the ultimate bearing capacity of the pile, load allowed per pile can be calculated as:

$$Q_w = \frac{Q_{up} + Q_{us}}{F.S} \tag{3}$$

where Q_w is the permissible load on each pile (Kg). F.S is the safety factor for each pile (4 in this study). From Eq. 3, the permissible bearing capacity of the pile is:

$Q_w = (5774 + 4239)/4 = 2503KN$ The bearing capacity of piles in the middle slab

$Q_w = (5774 + 4578)/4 = 2588KN$ The bearing capacity of piles in the side slabs

The number of piles in the middle and side slabs is 8 and 10, respectively. Also, the middle slab area is 111 m^2, and the side slab area is 129 m^2. Considering the bearing capacity of each pile and the number of piles, the load on each slab is equal to:

$q = (8*2503)/111= 180\ KN/m^2$ Middle slab

$q = (10*2588)/129= 200\ KN/m^2$ Side slabs

2.5 Three-Dimensional Tunnel Modeling

For optimizing tunnel overburden (reducing TBM system depreciation and decreasing TBM drilling pressures), TBM crossing under the Mianrood bridge is modeled at two depths of 18 and 14 meters. This modeling is performed by Plaxis 3D software. Depths of 18 and 14 meters are suggested by the project consultant and the authors, respectively. The main reason for lowering the tunnels to a depth of 18 meters, by project consultant, is to lessen the impact of tunnel drilling, on the Mianrood Bridge. Figure 6 shows the geometry of the modeling of the tunnels and the Mianrood bridge in two overburden of 14 and 18 meters.

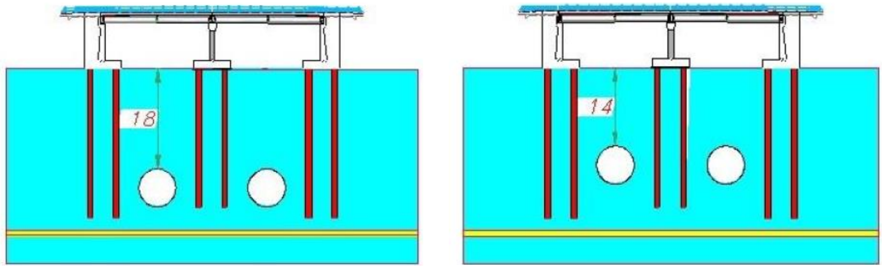


Figure 6
Two scenarios for modeling of tunnels under Mianrood bridge

Plaxis 3D is a powerful software that has many capabilities in simulating the drilling process, such as, applying a working pressure chest, considering the amount of shrinkage, applying injection pressure, and drilling and segmentation steps. Considering the tunneling steps in the EPB method and above capabilities, Plaxis 3D v1.2 software was used for calculating the settlement parameters. The program is based on the finite element equations used in the two-dimensional program. In addition to two-dimensional simulations, the program can simulate the three-dimensional behavior of underground structures in various soil environments. Materials construction was done by user-created areas in the model. Each member responds to forces or boundary constraints, according to stress-strain laws (linear or nonlinear). The speed of software computation depends on the number of model areas and the speed of the computer. The modeling steps are as follows:

Geometry: Due to the overburden and the distance of the tunnels at the cross-section, length of 60 meters and depth of 35 meters are considered for all models. Dimensions are chosen so as unrealistic boundary conditions do not affect the model. Two depths of 14 and 18 meters are provided for the tunnels.

Boundary Conditions: After defining two-dimensional and three-dimensional geometry, it is the turn of the boundary conditions. Hinged and roller conditions were considered for the bottom of the model and the sides, respectively. According to section 2.4.2, the slab load in the middle and side slabs are 180 KN/m^2 and 200 KN/m^2, respectively.

Geotechnical Properties of Aggregates: Tables (1) and (2) present the properties of material, shields, and segments. Triangular elements are used to mesh. The model is first trimmed in two-dimensional mode and then expanded to the third dimension. Figure 7 shows the whole model and its three-dimensional mesh.

Table 1
Geotechnical properties of soil layers

ID	Type	g_unsat	g_sat	Nu	E_ref	c_ref	φ
		KN/m^3	KN/m^3		KN/m^2	KN/m^2	degree
BH 15-1 (clay)	Drained	15	17	0.28	2.10e04	15	25
BH 15-2 (sand)	Drained	17	20	0.28	1.3ee04	1	31

Table 2
Shield and segment specifications

ID	Type	EA	EI	nu
		KN/m	KN.m^2/m	
Plate	Elastic	8200000	83800	0
Segment	Elastic	10570000	107900	0.15

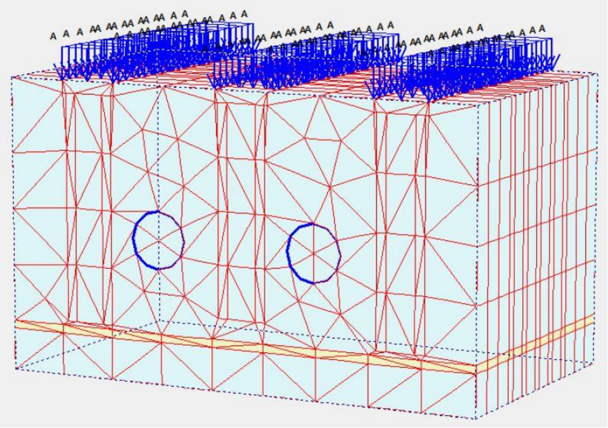
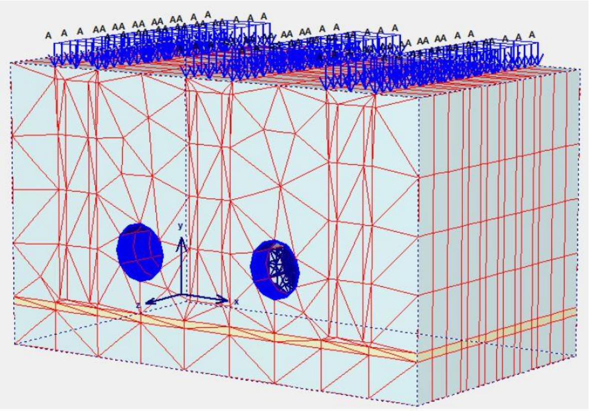


Figure 7
The three-dimensional model of tunnels crossing under the Mianrood bridge

Preliminary Conditions: After completing the previous steps, the initial conditions (effective stresses) are based on the soil lateral pressure coefficient and the water pressure. The water level is considered based on the geological maps in the model.

Computational Phases: At this stage, the authors attempted to simulate the actual drilling conditions. Overall, there are three steps in simulating drilling cycles in EPB modeling. First, it is the shield and chamber of the machine that performs the drilling and pressure on the chest. Since the shield length is 9 meters, shield drilling is done in three steps of 3 meters. In this section, shield parameters are considered for the lining. For modeling piles, the lengths of the advances are also proportional to the intervals of the piles. The injection operation behind the segments is modeled. In general, there should be a relative balance between the working pressure and injection pressure. The injection pressure is 50 kPa higher than the working pressure. It should be noted that at this stage there is no lining. After injection of slurry into the vacuum between the segment and the soil around the tunnel, the segments are considered for the lining. Figure 8 shows a complete drilling cycle with applied front work pressure and grout injection pressure, and tunnel lining system. According to the analytical equations, the working chest pressure values of 140 and 170 kPa in the 14 and 18 meter overburden, are considered in the tunnel crest.



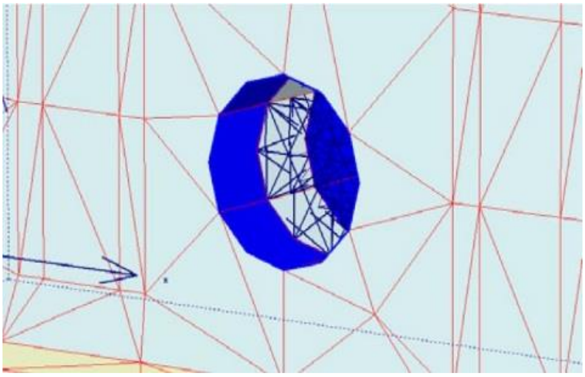


Figure 8
A complete drilling cycle with applied frontal and grout injection pressures and tunnel lining system

3 Results and Discussion

After modeling, the results show that the displacement of the ground and slabs at both 14 and 18 meters tunnel overburden, are close together. Figures 9 and 10 show the displacement and impact area at two depths of 18 and 14 meters. In order to study more precisely and compare the displacement of different points in two 14 and 18eters overburden, the authors defined 10 points in geometry and calculated all displacements at these points. The location of the points is given in Figure 11. Also, Table 3 presents the displacement of these points.

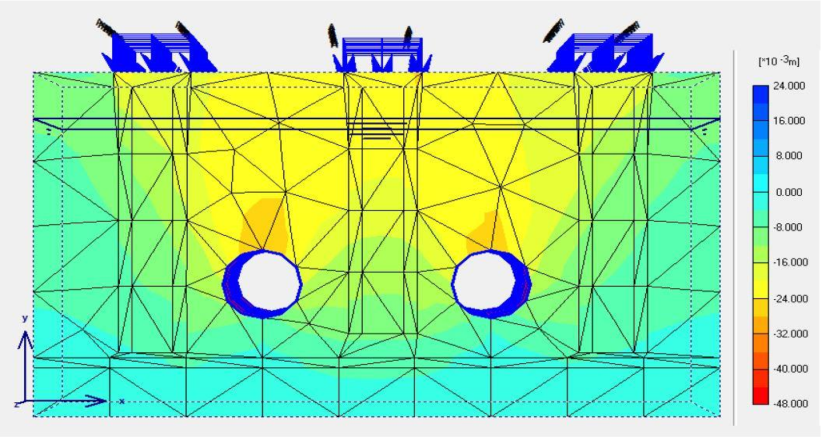


Figure 9
Displacement and impact area at 18 meters overburden

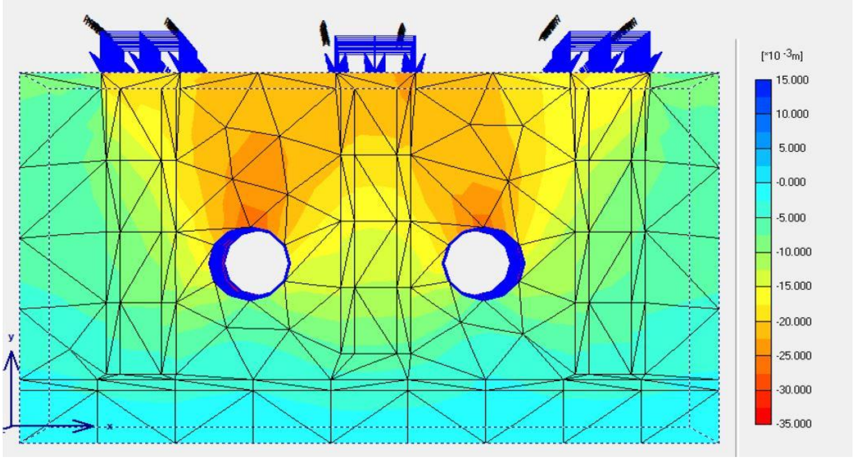


Figure 10
Displacement and impact area at 14 meters overburden

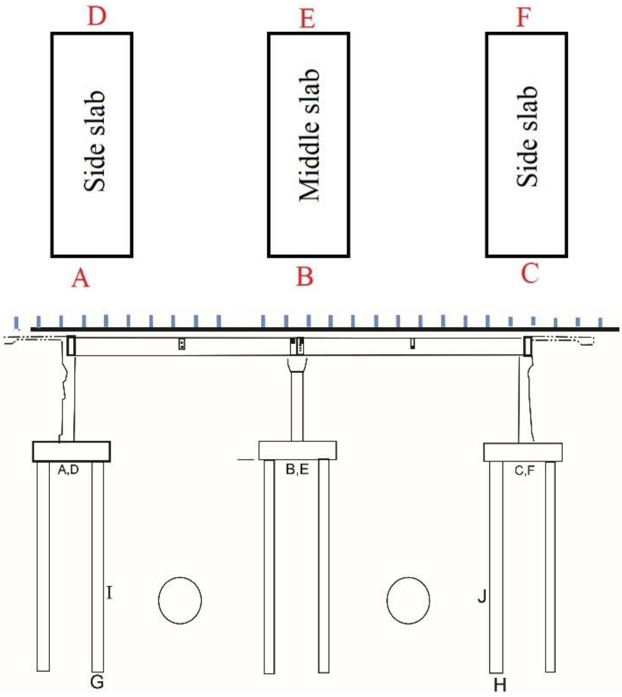


Figure 11
The location of specified points by authors for comparing tunnel displacement

Table 3
The displacement of specified points in the model at two overheads of 14 and 18 meters

Point	14 m Overburden	18 m Overburden
	Dis. (mm)	Dis. (mm)
A	17	17
B	20	22
C	14	15
D	13	13
E	15	16
F	11	12
G	11	12
H	11	12
I	15	15
J	13	14

Another important issue that should be considered is the distance from the bottom of the tunnel to the tip of the pile. This distance affects the capacity of the pile tip. Given that the depth of the piles in the middle slab of the bridge is 25 meters, if the tunnel is located at a depth of 18 meters, the floor of the tunnel is approximately 1.5 meters from the tip of the pile. When the tunnel is located at a depth of 14 meters, the distance between the tunnel floor and the pile tip is greater. Therefore, it is suggested that the designer changes the depth of the tunnel overburden from 18 meters to 14 or 15 meters. Figures 12 and 13 show the effect of the location of the tunnel on the pile tip.

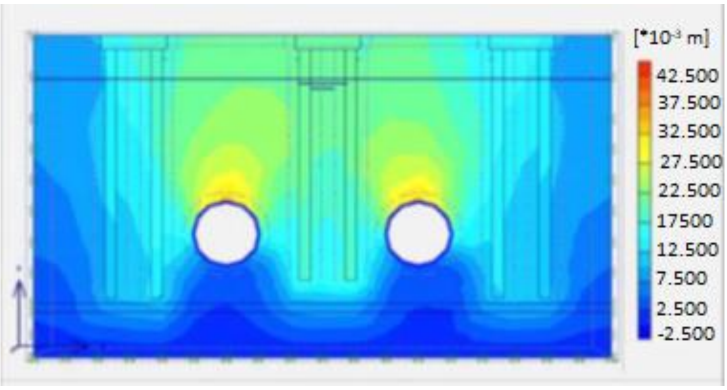


Figure 12
The effect of tunnel displacement on the capacity of pile tip (14 meters overburden)

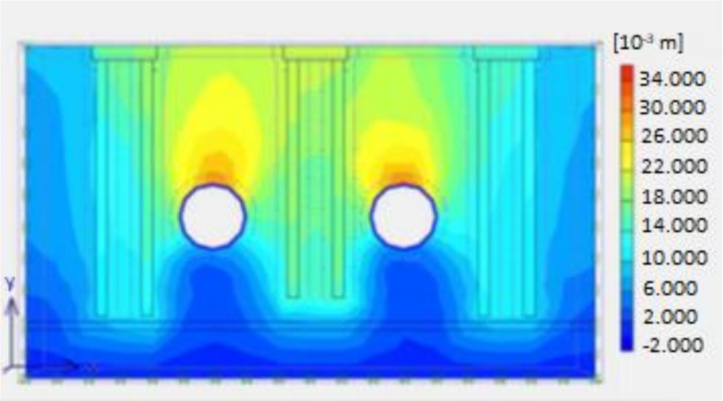


Figure 13
The effect of tunnel displacement on the capacity of pile tip (18 meters overburden)

In this study, the simulation of crossing Shiraz metro line 2 under the Mianrood bridge was done using the Plaxis-3D software. The suggested depth for the tunnel by the project consultant was 18 meters. Nevertheless, the authors attempted to optimize this depth. Therefore, they suggest that the depth of 14 meters is also examined. After simulating the crossing of the tunnel, it was found that both the depths of 14 and 18 meters, have a similar displacement. However, considering the effect of the tunnel on the capacity of the pile tip, the depth of 14 meters is more appropriate, because the distance of the tunnel bottom to the pile tip is less. Consequently, the authors' recommendation is to design the tunnel at a depth of 14 meters.

Conclusions

This study aimed to study, model and optimize a part of the route of the Shiraz metro line 2. The reason for choosing this part is that the tunnel in this part, passes under a traffic interchange (Mianrood bridge). In fact, the main goal was to find the optimum depth for the construction of the tunnel. The initial proposal for the tunnel depth was 18 meters and was suggested by the project consultant. Due to executive difficulties, as well as, the experiences of line 1, the authors have suggested a depth of 14 meters. The soil mass modeling around the tunnel was performed using the PLAXIS finite element program for the two depths of 14 and 18 meters. The modeling had two important consequences:

1) The displacements caused by tunneling at depths of 14 and 18 meters were similar.

2) In the middle slab of the bridge, the piles are 25 meters in depth. In the 18 meters overburden, the tunnel floor is less distant from the tip of the pile. Thus, the tunnel has a greater impact on the bearing capacity of the pile tip (compared to the 14 meters overburden).

As a result, the authors propose to change the tunnel depth, from 18 to 14 meters. For future studies, the authors intend to analyze other parts of the tunnels. Furthermore, modeling can be done using different software such as Abaqus and Flac, then comparing the results with Plaxis 3D.

Acknowledgment

The support of the Hungarian-Mexican bilateral project (2019-2.1.11-TÉT-2019-00007) is acknowledged.

References

[1] Xu, W., P. Zhao, and L. Ning, *Last train delay management in urban rail transit network: Bi-objective MIP model and genetic algorithm.* KSCE Journal of Civil Engineering, 2018, **22**(4): pp. 1436-1445

[2] Chen, H., B. Jia, and S. Lau, *Sustainable urban form for Chinese compact cities: Challenges of a rapid urbanized economy.* Habitat international, 2008, **32**(1): pp. 28-40

[3] Mohammadzadeh, D., et al. *Urban train soil-structure interaction modeling and analysis*. in *International Conference on Global Research and Education*. 2019, Springer

[4] Abdollahzadeh Nasiri, A. S., et al., *Evaluation of Safety in Horizontal Curves of Roads Using a Multi-Body Dynamic Simulation Process.* International Journal of Environmental Research and Public Health, 2020, **17**(16): p. 5975

[5] Karballaeezadeh, N., et al., *Estimation of flexible pavement structural capacity using machine learning techniques.* Frontiers of Structural and Civil Engineering, 2020, **14**(5): pp. 1083-1096

[6] Mayer, T. and C. Trevien, *The impact of urban public transportation evidence from the Paris region.* Journal of Urban Economics, 2017, **102**: pp. 1-21

[7] Ding, R., et al., *Heuristic urban transportation network design method, a multilayer coevolution approach.* Physica A: Statistical Mechanics and its Applications, 2017, **479**: pp. 71-83

[8] Li, L., et al., *Urban transit coordination using an artificial transportation system.* IEEE Transactions on Intelligent Transportation Systems, 2010, **12**(2): pp. 374-383

[9] Knox, P. L. and L. McCarthy, *Urbanization: an introduction to urban geography*. 1994: Prentice-Hall Englewood Cliffs, NJ

[10] Tayyaran, M. R. and A. M. Khan, *The effects of telecommuting and intelligent transportation systems on urban development.* Journal of Urban Technology, 2003, **10**(2): pp. 87-100

[11] Fargnoli, V., et al., *3D numerical modelling of soil–structure interaction during EPB tunnelling.* Géotechnique, 2015, **65**(1): pp. 23-37

[12] Mroueh, H. and I. Shahrour, *A simplified 3D model for tunnel construction using tunnel boring machines.* Tunnelling and Underground Space Technology, 2008. **23**(1): pp. 38-45

[13] Herrenknecht, M. *New developments in large-diameter tunnel design manufacture and utilisation for world-wide projects*. in *World Tunnel Congress*. 1998

[14] Kurihara, K. *Current mechanized shield tunneling methods in Japan (Invited lecture)*. in *Proc. of the World Tunnel Congress 98 on Tunnels and Metropolises*. 1998, Balkema

[15] Kuwahara, S. *Mechanized and automated tunnelling in Japan*. in *Proceedings of the International Symposium on Ground Challenges and Expectations in Tunnelling Projects, Cairo, Egypt*. 1999

[16] Farias, M. M., Á. H. Moraes Júnior, and d. A. A. Pacheco, *Displacement control in tunnels excavated by the NATM: 3-D numerical simulations.* Tunnelling and Underground Space Technology, 2004, **19**(3): pp. 283-293

[17] Negro, A. and B. Queiroz, *Prediction and performance of soft ground tunnels*, in *Geotechnical Aspects of Underground Construction in Soft Ground*. 1999: Balkema, Tokyo, Japan, pp. 409-418

[18] Afifipour, M., et al., *Interaction of twin tunnels and shallow foundation at Zand underpass, Shiraz metro, Iran.* Tunnelling and Underground Space Technology, 2011, **26**(2): pp. 356-363

[19] Broere, W. and R. Brinkgreve, *Phased simulation of a tunnel boring process in soft soil.* Numerical Methods in Geotechnical Engineering, Mestat (ed.), Presses de l'ENPC/LCPC, Paris, 2002: pp. 529-536

[20] Möller, S. and P. Vermeer, *On numerical simulation of tunnel installation.* Tunnelling and Underground Space Technology, 2008, **23**(4): pp. 461-475

[21] Migliazza, M., M. Chiorboli, and G. Giani, *Comparison of analytical method, 3D finite element model with experimental subsidence measurements resulting from the extension of the Milan underground.* Computers and Geotechnics, 2009, **36**(1-2): pp. 113-124

[22] Ng, C. W., K. M. Lee, and D. K. Tang, *Three-dimensional numerical investigations of new Austrian tunnelling method (NATM) twin tunnel interactions.* Canadian Geotechnical Journal, 2004, **41**(3): pp. 523-539

[23] Barla, G., et al. *Two and three dimensional modelling and monitoring of the Metro Torino*. in *11th International conference of Iacmag, Turin (Italy)* 2005

[24] Dias, D., R. Kastner, and S. Benmebarek. *Slurry shield tunnelling: comparison between in situ data and three dimensional numerical*

simulations. in *Proceedings of The International Conference On Soil Mechanics And Geotechnical Engineering*. 2002, AA Balkema Publishers

[25] Dias, D., R. Kastner, and M. Maghazi. *Three dimensional simulation of slurry shield tunnelling*. in *Geotechnical aspects of underground construction on soft ground*. 2000

[26] Lambrughi, A., L.M. Rodríguez, and R. Castellanza, *Development and validation of a 3D numerical model for TBM–EPB mechanised excavations.* Computers and Geotechnics, 2012, **40**: pp. 97-113

[27] Mollon, G., *Etude dèterministe et probabiliste du comportement des tunnels*. 2012, INSA de Lyon

[28] Mollon, G., D. Dias, and A.-H. Soubra, *Probabilistic analyses of tunneling-induced ground movements.* Acta Geotechnica, 2013, **8**(2): pp. 181-199

[29] Saadin, H. K. a. M., *Analysis and Prediction of Land Surface Settlement Due to Tunneling (Case Study: Tabriz Urban Train Line 2 Project).* Transportation Engineering, 2010, **4**

[30] Ng, C. W., H. Huang, and G. Liu, *Geotechnical Aspects of Underground Construction in Soft Ground: Proceedings of the 6th International Symposium (IS-Shanghai 2008)* 2008: CRC Press

[31] Jongpradist, P., et al., *Development of tunneling influence zones for adjacent pile foundations by numerical analyses.* Tunnelling and underground space technology, 2013, **34**: pp. 96-109

[32] Selemetas, D., J. Standing, and R. Mair. *The response of full-scale piles to tunnelling*. in *Geotechnical aspects of underground construction in soft ground. Proceedings of the 5th international conference of TC 28 of the ISSMGE, the Netherlands, 15-17 June 2005*, 2006

[33] Schroeder, F. C., *The influence of bored piles on existing tunnels.* 2003

[34] Das, B. M., *Principles of Foundation Engineering 6th Edition*. 2007: Thomson

Appendix

Term	Description
TBM	Tunnel Boring Machines
FEM	Finite Element Method
EPB	Earth Pressure Balance
Q_{up}	Ultimate Bearing Capacity of Pile Tip
Q_{us}	Ultimate Bearing Capacity of Pile Wall
F.S	Safety factor
Kg	Kilogram
KN	Kilonewton

A denotational semantics of a concatenative/ compositional programming language

Jurij Mihelič, William Steingartner, Valerie Novitzká

University of Ljubljana, Faculty of Computer Science and Informatics,
Večna pot 113, 1000 Ljubljana, Slovenia;
jurij.mihelic@fri.uni-lj.si

Technical University of Košice, Faculty of Electrical Engineering and Informatics,
Letná 9, 042 00 Košice, Slovakia;
{william.steingartner, valerie.novitzka}@tuke.sk

Abstract: A distinctive feature of concatenative languages is that a concatenation of their programs corresponds to a composition of meaning functions of these programs. At first programming in such languages may resemble assembly language programming. In spite of this, they also exhibit many similarities to high-level functional programming languages. We start our presentation with the definition of the language syntax. The main part of the paper consists of the definition of a meaning of programs in the language. To do this we employ a well-known method based on denotational semantics. We also informally introduce the language and its meaning as well as present its background and provide motivation for the work. Our exposition is accompanied by many examples and in the last part of the paper, we also discuss various language extensions and identify several proposals for further research.

Keywords: concatenative, compositional, denotational semantics, function, programming language, syntax, stack

1 Introduction

In this paper, we focus on defining the denotational semantics for a new concatenative/compositional language. We begin the paper with a review of the area and related work. First, in the next subsection, we review approaches to define a meaning of programs. Since the presented programming paradigm is not well known, we present its background with a related work. Finally, we will give an informal introduction to the discussed programming language.

1.1 Short overview of the area

Each programming language should have its own formal definition. This definition consists of formal syntax and formal semantics. The former can be concrete or abstract where a concrete syntax serves for syntax analysis and abstract syntax is suitable for defining the semantics of a program. The latter expresses the meaning of a program. For programs written in purely functional languages, a value of a term is a meaning of the program [8]. In contrary, the semantics of an imperative program is defined as a change of memory states (storage).

Our goal is to define a semantics for our language KKJ[1]. Since there are several well-known approaches to semantics that are reciprocally equivalent and they are used for different purposes, we briefly look over them. One of the most popular methods for defining the semantics of programming languages is structural operational semantics (also known as small-step semantics). This method models, in details, computations explicitly in particular steps of execution and describes the effects of program constructs on program states and each step is expressed as the transition relation [17]. Operational semantics specifies programming languages in terms of program execution on abstract machines which provide an intermediate language stage for compilation. They bridge the gap between the high level of a programming language and the low level of a real machine [4]. Structural operational semantics represents computation by means of deductive systems that turn the abstract machine into a system of logical inferences [19]. An alternative approach in operational semantics is known as natural semantics or big-step semantics. In natural semantics [10], the relationship between the initial and the final state of an execution is constructed. This method is mostly used for imperative languages but it has nice application also in area of domain specific languages, e.g. [2].

A further particular type of small-step semantics is Reduction Semantics with Evaluation Contexts (RSEC) [5], also known as contextual semantics. This method models an execution as a sequence of atomic rewrites of state, between each of which some small amount of time passes [3, 6].

Another approach to semantic methods is axiomatic semantics which models the relationship between pre- and post-conditions on program variables – it describes properties of program state, using the first-order logic [11, 14].

Action semantics [13, 24] is considered as a hybrid of denotational and operational semantics. Action semantics uses English phrases for defining the meaning of syntactic constructs (still being formal). It serves mostly as a very illustrative framework for teaching semantics [24].

In this approach, we present how to formulate and define the denotational semantics. However, denotational semantics is one of the oldest semantic methods [14], where only the contribution of each construct to the computational meaning of the enclosing program is modeled. This method defines the meaning of a program using functions/mappings defined over sets and/or lattices, respectively [18]. The intermediate states during the execution of the constructs are generally of no relevance and are thus not represented. One of its main aims is to provide a proper mathematical foundation for reasoning about programs and for understanding the fundamental concepts of programming languages. Therefore, denotational semantics plays an important *rôle* in the language design process.

[1] KKJ – konkatenacijski/kompozicijski jezik in Slovene or konkatenatívny/kompozičný jazyk in Slovak, both meaning concatenative/compositional language.

1.2 Background

In this paper, we focus on a simple programming language introducing only a handful of programming constructs while still offering a flexible and useful programming environment. The simplicity of the language, which for the purposes of this presentation we call KKJ, spurs a plethora of possibilities such as an option to clearly define semantics, to straightforwardly implement an interpreter or a compiler, and to design program analysis tools for the purpose of optimization or verification, etc. The so called concatenative/compositional nature of the language offers a great flexibility for decomposing a program into units than can be executed in parallel, and, as such, have a potential to be suitable for modern multi-core computer architectures as well as to serve as an intermediate code representation in compilers and interpreters. We provide a brief discussion on this later in the paper.

The proposed language abstracts away all the intricate details of any possible underlying hardware and processor architecture. Nevertheless, implementing programs in KKJ may occasionally resemble programming in an assembly language (without architectural details) since a programmer uses only simple basic "instructions" to arrive at a solution for a particular programming task.

On the other hand, the language also offers a programming construct representing first-class anonymous function which is usually not present in low-level languages and as such a programmer's perspective is raised to a level usually occurring in higher-level programming languages. In particular, programming in KKJ becomes similar to programming in functional programming languages without using variables, e.g., the *tacit* or *point-free* style of programming which is oftentimes aspired.

Many ideas found in this paper already appear in some similar form throughout the scientific literature. Already in 1977, John Backus in his Turing award lecture argued for simplification of conventional programming languages as well as proposed functional programming systems as an alternative. He explicated a *function-level* programming where new programs are written by putting together existing programs rather than by manipulating values and then abstracting from those values to produce programs. The result of his efforts is the FP language proposed in the lecture, which through many improvements involving several researchers evolved later into the FL [1], and other programming languages.

In contrast, conventional well-known *functional* programming languages such as Haskell and Lisp mostly base on the *lambda calculus* which puts forth *value-level* programming where new values are constructed from existing ones until the final result is obtained. The development of these has already greatly advanced from their inception and many modern functional programming languages offer both high expressiveness as well as execution speed.

Nevertheless, even though both function-level and value-level functional programming adopt somewhat different views on object manipulation their fundamental programming concept is still a function. In several other successful programming paradigms, such a concept may also be an object (i.e., object-oriented programming) or a relation (i.e., logic programming).

Our other source of inspiration is the area of *concatenative* programming languages. The term arises from the property that a (syntactic) concatenation of programs corresponds to the (semantic) composition of functions. Indeed, both aspects actually represent *monoid* algebraic structure. In particular, an empty program is a unit for program concatenation and program concatenation is clearly associative while an identity function is a unit for function composition and composition is associative as well.

The paradigm of concatenative languages is valuable for fundamental software engineering research (ideally for language experimentation and worth to be applied in software engineering because of their unique features) and might prove to be a suitable foundation for future programming [7, 21, 22] .

Unfortunately, the area received little attention from a scientific community and thus its treatment lack theoretical rigor and strictness. Moreover, many of the concatenative languages exist only as a prototype implementation, are not actively developed, and they lack financial support as well as any serious development environment and support. A list of several examples including brief descriptions is available on the http://concatenative.org website. Nevertheless, there are some prominent examples worth mentioning.

The Forth language [12] appeared in 1970 and is considered as a flexible, extensible, stack-based, procedural, concatenative programming language with many applications. Another is the Joy language [23] from 2001 which is considered as a purely functional, of theoretical interest and has established the term concatenative and had influenced many other concatenative languages. Finally, the Factor language [15, 25] which was conceived in 2003 and is an actively developed, dynamically typed, garbage collected language with a self-hosting compiler, interactive development environment, and large standard library. It supports both functional and object-oriented programming paradigm and is well used in practice.

Finally, we mention also several stack-based application virtual machines which often display many similarities with concatenative paradigm and they have already proved themselves to be successful and efficient in practice. Probably the most well-know examples are the Java Virtual Machine and Common Language Runtime as well as Erlang's BEAM runtime.

The language proposed in this paper resembles many of the above-mentioned languages in the way function composition is used, but it is simpler in order to enable theoretical consideration using formal methods. Our simplifications are partially in syntax (e.g., less syntactic domains) and also in semantics (e.g., no modules and information hiding), but the main concatenative features are present (e.g., function combinators).

The goal of the paper is two fold: first, to serve as a presentation of concatenative (compositional) programming constructs, and, second, to establish a firm basis for understanding the meaning of concatenative programs which is currently missing in the scientific literature. Indeed, current implementations of such programming languages are based on *ad hoc* definition of the concepts.

1.3 Informal description of KKJ

Since the corresponding concatenative programming paradigm of the proposed language is not well-known we start with a brief informal description of KKJ followed by a demonstration of the evaluation of an example program.

Syntactically a program in KKJ is just a sequence of *words*, i.e., numerals represent-

ing numbers and names representing functions, where words are separated with a whitespace. Additionally, there is also a programming construct called a *quotation* (i.e., abstraction) which encapsulates another program with brackets and it represents a definition of an anonymous function. The term quotation is also used in the Lisp programming language for a similar construct.
Observe the following two examples of programs in KKJ. The first one is without quotations and it consists of ten words:

```
3 4 add dup ispos 5 6 swap choose mul,                        (1)
```

and the second one contains two quotations:

```
14 {dup dup} {add add} compose apply.                         (2)
```

As we will see later, both programs evaluate to the same value, i.e., they are semantically equivalent. We notice also, that the defined syntax exhibits such a great simplicity that it is consequently very straightforward to implement an efficient parser to perform syntax analysis.
Now we focus our attention to the meaning of the above two programs. In what follows we show two different approaches for program evaluation. The first approach is based on term-rewriting (i.e., reduction semantics) where specific patterns in the program are found and being replaced until a *normal form* of the program is obtained. An example evaluation of the program 1 is shown in Table 1. We omit rule specification and rely on a reader's intuitive understanding.

Table 1
Evaluation of a program based on term-rewriting.

program	substitutions
`3 4 add dup ispos 5 6 swap choose mul`	`3 4 add → 7,`
	`5 6 swap → 6 5`
`7 dup ispos 6 5 choose mul`	`7 dup → 7 7`
`7 7 ispos 6 5 choose mul`	`7 ispos → true`
`7 true 6 5 choose mul`	`true 6 5 choose → 6`
`7 6 mul`	`7 6 mul → 42`
`42`	

Based on the example one can clearly see that the language is functional in a way that basic expressions represent built-in (or primitive) functions and a sequence of expressions forms a new expression representing a composition of functions; in other words, a computation is carried out entirely through the evaluation of expressions. Even though a functional point of view is a more appropriate one, built-in functions may also represent constructs usually found in the imperative paradigm of programming.
Another approach to evaluation is based on the stack data structure (i.e., state-transition semantics), where a program successively transforms the stack. Each word represents a function operating on the stack, where numerals represent functions pushing a number onto the stack. Again, an example evaluation of the program (1) is shown in Table 2.

Table 2
Evaluation of a program based on a stack.

program	stack
`3 4 add dup ispos 5 6 swap choose mul`	
`4 add dup ispos 5 6 swap choose mul`	3
`add dup ispos 5 6 swap choose mul`	3 4
`dup ispos 5 6 swap choose mul`	7
`ispos 5 6 swap choose mul`	7 7
`5 6 swap choose mul`	7 **true**
`6 swap choose mul`	7 **true** 5
`swap choose mul`	7 **true** 5 6
`choose mul`	7 **true** 6 5
`mul`	7 6
	42

Indeed, many concatenative programming languages are stack-based (i.e., operations manipulate the implicit stack), which may suggest an imperative view. However, in imperative languages the state is implicit, but it is explicitly manipulated (e.g. via assignments) whereas in stack-based concatenative languages the stack manipulation is considered implicit. Moreover, operations in these languages may also be regarded as unary transformations from stack to another stack.
Consider now the program (2). What value does it evaluate to? What is the meaning of a quotations {`dup dup`} and {`add add`}? We intuitively know that the first one is a function that duplicates the top element twice and the second one is a function that pops and adds three top values on the stack. Now, composing these two functions, i.e., the meaning of {`dup dup`} {`add add`} `compose`, results in a new function which given a number returns its triple. So, we simply calculate 3×14 here. Nevertheless, to clearly answer these questions we have to formally define a meaning function and that is the goal of the rest of the paper.

2 Syntax and semantics

In this section, we present a foundations of KKJ– we start with the definition of syntax. Then we define an abstraction of computer memory as memory states. After this step, we are ready to define a semantics of KKJ.

2.1 Syntax

Before delving into the semantics of KKJ, we define abstract syntax if its expressions of KKJ. First, we introduce syntactic domains:

- $i \in$ **IntNum** – integer numerals, i.e., strings of digits,
- $n \in$ **Name** – names, i.e., strings of alphanumeric characters,
- $E \in$ **Expr** – expressions.

Here, the elements $i \in \textbf{IntNum}$ represent integer numbers and they have no internal structure from the semantic point of view, but syntactically they can be represented with a regular expression $[0,\ldots,9]^{+}$. Similarly $n \in \textbf{Name}$ represent function names without any internal structure significant to defining semantics and its internal syntax is described with a regular expression $[\mathsf{a},\ldots,\mathsf{z}][\mathsf{a},\ldots,\mathsf{z},0,\ldots,9]*$.

To describe the syntax of expressions $E \in \textbf{Expr}$ in the programming language KKJ, we use the well-known BNF-notation:

$$E ::= \varepsilon \mid i \mid n \mid \{E\} \mid E\,E. \tag{3}$$

Here, ε stands for an empty expression, a numeral $i \in \textbf{IntNum}$ and a name $n \in \textbf{Name}$ are considered as expressions as well. Additionally, one can form a new expression by quoting (with brackets) an existing expression, i.e., a quotation $\{E\}$, to represent an anonymous function defined by the expression E. And, finally, a new expression can be formed by concatenating (in juxtaposition using only whitespace as a delimiter) two expressions, i.e., $E \;\; E$, simply to represent their composition.

We also note here that a name $n \in \textbf{Name}$ is considered to represent a built-in (also called primitive) operation such as `add`, `sub`, `pop`, `dup`, `compose`, and `apply`. We exactly specify these names in the following sections while specifying semantics. We assume that undefined names are not allowed in correct programs, whereas in practice they would cause the program to crash or raise an exception. However, we extend later the basic syntax of the language with a construct which allows us to assign functions to names.

2.2 Representation of states

The state is, in general, an abstraction of a computer memory (a kind of memory snapshot) and in our case, it is actually represented by a stack. In what follows we show how the stack is defined. First, we introduce a new semantic domain **Int** for integer values and **Bool** for Boolean values. They are defined as

$$\textbf{Int} = \mathbb{Z} \quad \text{and} \quad \textbf{Bool} = \mathbb{B} = \{\textbf{false}, \textbf{true}\}.$$

Second, we introduce a semantic domain **Stack** for representing the stack as well as a semantic domain **Fun** (a function space) for representing functions manipulating the stack, i.e.,

$$\textbf{Fun} = \textbf{Stack} \to \textbf{Stack}.$$

Now to define **Stack**, we first introduce a new domain **Value**, which represents values (i.e., elements, members) residing on the stack, i.e.,

$$\textbf{Value} = \textbf{Int} \cup \textbf{Bool} \cup \textbf{Fun}.$$

Finally, the type stack is represented with the following semantic domain

$$\textbf{Stack} = \textbf{Value}^{*} \cup \{\bot\},$$

where X^* represents Kleene's closure (or iteration) over X. Observe that, a stack $s \in \textbf{Stack}$ represents an abstraction (a snapshot) of the actual memory and thus represents a state in our semantics. A particular content of the stack may also be represented with an ordered sequence, i.e., $(x_1, x_2, \ldots, x_n) \in \textbf{Stack}$, where $x_i \in \textbf{Value}$ for each $1 \leq i \leq n$ and x_n is a topmost element of the stack.

Notational remark: We mostly use symbols $i, j \in \textbf{Int}$ for integers and $b, d \in \textbf{Bool}$ for Boolean values, $f, g, h \in \textbf{Fun}$ for functions, $x, y, z \in \textbf{Value}$ for value of any type, and $s, t \in \textbf{Stack}$ for stacks.

2.3 Semantics

Now let us describe denotational semantics for the language KKJ. To do this we specify a semantics of expressions $E \in \textbf{Expr}$, where their meaning can be summarized by a function from **Stack** to **Stack**, i.e.,

$$\mathscr{S} : \textbf{Expr} \rightarrow (\textbf{Stack} \rightarrow \textbf{Stack}). \tag{4}$$

The function $\mathscr{S}$ will be defined inductively in the following sections. In this paper, we often omit the symbol $\mathscr{S}$ and use the semantic bracket $[\![E]\!]$ around the syntactic parameter $E \in \textbf{Expr}$, when the notation is clear, e.g., $[\![E]\!] \equiv \mathscr{S}[\![E]\!]$.

When providing inductive definitions, we define semantic clauses for various syntactic constructs. Doing this we use several auxiliary functions defined as follows:

- newstack: $\rightarrow \textbf{Stack}$,
- id: $\textbf{Stack} \rightarrow \textbf{Stack}$,
- push: $\textbf{Value} \rightarrow \textbf{Stack} \rightarrow \textbf{Stack}$.

Here, the newstack is an initial function which *ex nihilo* creates a new empty stack, i.e., newstack $= ()$, the id function (identity) leaves a stack unchanged, i.e., id $s = s$, while the push function appends an element to a stack, i.e.,

$$\begin{aligned} \textsf{push}\ x\ (x_1, x_2, \ldots, x_n) &= (x_1, x_2, \ldots, x_n, x), \\ \textsf{push}\ x\ \bot &= \bot, \end{aligned}$$

where $x, x_i \in \textbf{Value}$ and $1 \leq i \leq n, n \geq 0$.

In the rest of the paper, we often write $s = (x_1, x_2, \ldots, x_n)$ and use the symbol $\cdot$ as an infix operator representing push, i.e., $s \cdot x \equiv$ push $x\ s$. Moreover, we also use $\cdot$ in pattern matching, and thus define pop $s \cdot x = s$ as a function that returns the stack without its top element, and top $s \cdot x = x$ as the function that returns the top element of the stack. Notice also, that we define push as a curried function.

Generally, we define function $[\![E]\!]$ by defining it on each expression from eq. (3) as follows:

$$\begin{aligned} [\![\varepsilon]\!]\ s &= s, \\ [\![i]\!]\ s &= s \cdot \mathbf{i} && \forall i \in \textbf{IntNum} \\ [\![\{E\}]\!]\ s &= s \cdot [\![E]\!], \\ [\![E_1 E_2]\!]\ s &= ([\![E_2]\!] \circ [\![E_1]\!])\ s. \end{aligned}$$

The semantics of an expression $[\![n]\!]\ s$ depends on concrete name $n \in \mathbf{Name}$: the language KKJ uses concrete names for arithmetic operations, Boolean operations, operations for stack manipulation, functions, condition and loop expressions. We explain the details of these definition in sections that follow.

2.4 Expression concatenation

Let us begin with a presentation of the semantics for expression concatenation. The semantics of an empty program is the identity function and the concatenation of two programs corresponds to a composition of the semantic functions corresponding to the programs. The semantic clauses are given in Table 3.

Table 3
Semantics of the empty expression and expression concatenation

$$[\![\varepsilon]\!] = \mathsf{id} \qquad [\![E_1\ E_2]\!] = [\![E_2]\!] \circ [\![E_1]\!]$$

where

$$f \circ g\ s = \begin{cases} \bot, & \text{if } s = \bot \vee g\ s = \bot; \\ f\ (g\ s), & \text{otherwise.} \end{cases}$$

The latter rule also gives a rationale for the term *concatenative* for naming this sort of programming languages, since concatenation of valid programs results in a new valid program. Moreover, the new program semantics is defined as a composition of the semantics of the original programs. Hence, the term *compositional* languages may also be used [9] analogously to the term *applicative* which is sometimes used for conventional functional programming languages.

Now, consider a sequence of expressions, we state that the exact order of how the expression concatenation rule is applied is not important from the viewpoint of semantics. First we write the following theorem.

Theorem 1. Let $E_1, E_2, E_3 \in \mathbf{Expr}$. We have

$$[\![E_3]\!] \circ [\![E_1\ \ E_2]\!] = [\![E_2\ \ E_3]\!] \circ [\![E_1]\!].$$

Proof. First, observe that the function composition as defined in Table 3 is associative. Then, on the left-hand side of the equation we have

$$[\![E_3]\!] \circ [\![E_1\ \ E_2]\!] = [\![E_3]\!] \circ ([\![E_2]\!] \circ [\![E_1]\!]) = [\![E_3]\!] \circ [\![E_2]\!] \circ [\![E_1]\!],$$

and on the right-hand side we have

$$[\![E_2\ \ E_3]\!] \circ [\![E_1]\!] = ([\![E_3]\!] \circ [\![E_2]\!]) \circ [\![E_1]\!] = [\![E_3]\!] \circ [\![E_2]\!] \circ [\![E_1]\!]$$

which are both obviously equal. □

When there are three or more concatenated expressions the rule can be applied in multiple ways. For example, having three expressions we may decompose $E_1\ E_2\ E_3$ into either $\llbracket E_1\ \ E_2\ \ E_3 \rrbracket = \llbracket E_3 \rrbracket \circ \llbracket E_1\ \ E_2 \rrbracket$ or $\llbracket E_1\ \ E_2\ \ E_3 \rrbracket = \llbracket E_2\ \ E_3 \rrbracket \circ \llbracket E_1 \rrbracket$, yet still obtaining the same semantical result. In a more general case with n concatenated expressions, we $n-2$ times use Theorem 1. We can state the following corollary.

Corollary. *Let $E_1, E_2, \ldots, E_n \in$ **Expr** be n expressions. We have*

$$\llbracket E_1\ \ E_2\ \ \ldots\ \ E_n \rrbracket = \llbracket E_n \rrbracket \circ \cdots \circ \llbracket E_2 \rrbracket \circ \llbracket E_1 \rrbracket.$$

2.5 Arithmetic and Boolean operations

In this section, we consider several functions representing arithmetic and Boolean operations. In particular, these functions deal only with integer or Boolean values, which may during the operation be consumed from the stack or produced and pushed onto it.

See semantic clauses listed in Table 4 for semantic definitions of the selected basic arithmetic and Boolean operations. In the specification, the meaning of $\llbracket i \rrbracket\ s$ is to push the number $\mathbf{i} \in \mathbf{Int}$ corresponding to the numeral i on the stack s. For details on how to define an additional semantic function determining the number for a given numeral see [14].

Table 4
Semantics of arithmetic and Boolean operations

$\llbracket i \rrbracket\ s = s{\cdot}\mathbf{i}$	$\llbracket \mathtt{sub} \rrbracket\ s{\cdot}i{\cdot}j = s{\cdot}(i-j)$
$\llbracket \mathtt{add} \rrbracket\ s{\cdot}i{\cdot}j = s{\cdot}(i+j)$	$\llbracket \mathtt{mul} \rrbracket\ s{\cdot}i{\cdot}j = s{\cdot}(i \times j)$
$\llbracket \mathtt{false} \rrbracket\ s = s{\cdot}\mathbf{false}$	$\llbracket \mathtt{not} \rrbracket\ s{\cdot}b = s{\cdot}\begin{cases} \mathbf{true}, & \text{if } b = \mathbf{false}; \\ \mathbf{false}, & \text{otherwise.} \end{cases}$
$\llbracket \mathtt{true} \rrbracket\ s = s{\cdot}\mathbf{true}$	$\llbracket \mathtt{and} \rrbracket\ s{\cdot}b{\cdot}d = s{\cdot}\begin{cases} \mathbf{true}, & \text{if } b = d = \mathbf{true}; \\ \mathbf{false}, & \text{otherwise.} \end{cases}$
$\llbracket \mathtt{cmp} \rrbracket\ s{\cdot}i{\cdot}j = s{\cdot}\mathrm{sgn}(i-j)$	$\llbracket \mathtt{isneg} \rrbracket\ s{\cdot}i = s{\cdot}\begin{cases} \mathbf{true}, & \text{if } i < 0; \\ \mathbf{false}, & \text{otherwise.} \end{cases}$
	$\llbracket \mathtt{ispos} \rrbracket\ s{\cdot}i = s{\cdot}\begin{cases} \mathbf{true}, & \text{if } i > 0; \\ \mathbf{false}, & \text{otherwise.} \end{cases}$

The table also includes basic arithmetic operations such as addition (`add`), subtraction (`sub`), and multiplication (`mul`) as well as operations to produce Boolean values (`true` and `false`) together with operations for logical negation (`not`) and conjunction (`and`).

Additionally, we also include the operation for comparing (`cmp`) two integer numbers producing -1,0, or 1 on the stack if the first number is lower, the numbers are equal, or the second number is lower, respectively. And finally, operations to determine whether the top number on the stack is negative (`isneg`) or is positive (`ispos`).

Observe also, that the listed operations are defined only when the "input types match" as indicated by the use of variable names in the stack notation. For example, the operation $[\![\texttt{add}]\!]\ s{\cdot}i{\cdot}j$ is only defined when the top two elements i and j of the stack belong to the **Int** domain while it is not defined in all other cases, e.g., when the top element $j \in \mathbf{Bool}$, etc. Hence, the meaning of `7 true add` is evaluated as

$$[\![\texttt{7 true add}]\!]\ s = [\![\texttt{add}]\!]\ s{\cdot}7{\cdot}\mathbf{true} = \bot.$$

In this paper, we do not delve into details of type checking issues; we assume that the expressions are always correct regarding types. See also [16] for an in-depth discussion on types.

Several other important arithmetic and integer comparison operations can easily be formed using the basic operations from Table 4. To demonstrate this we give some examples in Table 5. Note: In the examples, we also use the operations `dup` (top of stack duplication) and `swap` (exchange of the top two elements) which are both defined in the next subsection.

Table 5
Several derived arithmetic and logical operations

`pred`	≡ `1 sub`	...	predecessor
`succ`	≡ `1 add`	...	successor
`neg`	≡ `0 swap sub`	...	negation
`iszero`	≡ `dup isneg not swap ispos not and`	...	is it zero?
`lt`	≡ `cmp isneg`	...	$<$
`le`	≡ `cmp dup isneg swap iszero or`	...	$\leq$
`eq`	≡ `cmp iszero`	...	$=$
`ne`	≡ `eq not`	...	$\neq$
`ge`	≡ `lt not`	...	$\geq$
`gt`	≡ `le not`	...	$>$
`or`	≡ `not swap not and not`	...	logical disjunction
`square`	≡ `dup mul`	...	square

2.6 Stack manipulation

In a programming language based on the function application parameters given to a function are explicitly specified by a programmer, but in a language based on function composition, parameters are implicitly set on a data stack and must there also be put into a specific order as required by the corresponding operation. Consequently, the programmer must be able to explicitly manage the stack by using special operations for manipulating the values on the stack. Such operations are occasionally also called *rewiring* operations.

We present definitions of semantic clauses for several stack manipulation operators in Table 6. Here, `clear` empties the stack, `id` represent the identity function, `pop` removes the top element, `dup` duplicates the top element, `over` duplicates the element just below the top of stack, `swap` exchanges the top two elements, and `rotl` rotates the top three element to the left.

Table 6
Semantics of basic stack manipulation operators

$\llbracket\texttt{clear}\rrbracket\ s = \mathsf{newstack}$	$\llbracket\texttt{over}\rrbracket\ s{\cdot}x{\cdot}y = s{\cdot}x{\cdot}y{\cdot}x$
$\llbracket\texttt{id}\rrbracket = \mathsf{id}$	$\llbracket\texttt{swap}\rrbracket\ s{\cdot}x{\cdot}y = s{\cdot}y{\cdot}x$
$\llbracket\texttt{pop}\rrbracket = \mathsf{pop}$	$\llbracket\texttt{rotl}\rrbracket\ s{\cdot}x{\cdot}y{\cdot}z = s{\cdot}y{\cdot}z{\cdot}x$
$\llbracket\texttt{dup}\rrbracket\ s{\cdot}x = s{\cdot}x{\cdot}x$	

Using the operations in Table 6 we can clearly perform the removal of arbitrary number of top stack elements, e.g., pop2 ≡ `pop pop`, pop3 ≡ `pop pop pop`, etc. We can also duplicate the top two elements, e.g., dup2 ≡ `over over`, but we cannot duplicate three or more top elements of the stack.

Let us notice also that `rotl` operation enables us to obtain any permutation of the top three stack elements. See Table 7 for definitions of operations which, given a stack $s{\cdot}x{\cdot}y{\cdot}z$ with top three elements x, y, and z, produce a particular permutation.

Table 7
Several derived stack manipulation operations

`id`		... $s{\cdot}x{\cdot}y{\cdot}z$
`swap`		... $s{\cdot}x{\cdot}z{\cdot}y$
`swapOver`	≡ `rotl swap`	... $s{\cdot}y{\cdot}x{\cdot}z$
`rotl`		... $s{\cdot}y{\cdot}z{\cdot}x$
`rotr`	≡ `rotl rotl`	... $s{\cdot}z{\cdot}x{\cdot}y$
`mirror`	≡ `rotl rotl swap`	... $s{\cdot}z{\cdot}y{\cdot}x$

2.7 Functions

In this subsection, we continue with a semantic clause for a quotation which represents an anonymous function. We also define some useful operations for manipulating functions; such operations are usually called *combinators*. See Table 8 for the list of quintessential semantic clauses appearing in concatenative programming languages.

Here, the semantics of quotation is to push the enclosing function on the stack. Operation `apply` takes an existing function from the stack and applies the function

Table 8
Semantics of quotations and function operations

$[\![\{E\}]\!]\ s = s\cdot[\![E]\!]$	$[\![\texttt{compose}]\!]\ s\cdot f\cdot g = s\cdot(g\circ f)$
$[\![\texttt{apply}]\!]\ s\cdot f = f\ s$	$[\![\texttt{applyOver}]\!]\ s\cdot f\cdot x = (f\ s)\cdot x$
$[\![\texttt{quote}]\!]\ s\cdot x = s\cdot\mathsf{push}\ x$	

on the remaining stack, quote takes a value from the stack and produces a function that pushes that value on the stack. Next, we have the compose operation which takes two functions and produces their composition, and applyOver which acts similarly to apply but it preserves the top element of the stack.

As an example, let us now evaluate the program (2) from the introduction.

$$\begin{aligned}
&[\![\texttt{14 \{dup dup\} \{add add\} compose apply}]\!]\ s = \\
&\quad = [\![\texttt{apply}]\!]\circ[\![\texttt{compose}]\!]\circ[\![\texttt{\{add add\}}]\!]\circ[\![\texttt{\{dup dup\}}]\!]\circ[\![\texttt{14}]\!]\ s \\
&\quad = [\![\texttt{apply}]\!]\circ[\![\texttt{compose}]\!]\circ\mathsf{push}\ [\![\texttt{add add}]\!]\circ\mathsf{push}\ [\![\texttt{dup dup}]\!]\circ[\![\texttt{14}]\!]\ s \\
&\quad = [\![\texttt{apply}]\!]\circ[\![\texttt{compose}]\!]\ s\cdot 14\cdot[\![\texttt{dup dup}]\!]\cdot[\![\texttt{add add}]\!] \\
&\quad = [\![\texttt{apply}]\!]\ s\cdot 14\cdot([\![\texttt{add add}]\!]\circ[\![\texttt{dup dup}]\!]) \\
&\quad = [\![\texttt{add add}]\!]\circ[\![\texttt{dup dup}]\!]\ s\cdot 14 = [\![\texttt{add}]\!]\circ[\![\texttt{add}]\!]\circ[\![\texttt{dup}]\!]\circ[\![\texttt{dup}]\!]\ s\cdot 14 \\
&\quad = [\![\texttt{add}]\!]\circ[\![\texttt{add}]\!]\ s\cdot 14\cdot 14\cdot 14 = s\cdot 42
\end{aligned}$$

Now consider a function twice $\equiv$ dup compose apply which applies a function twice. We have $[\![\texttt{twice}]\!]\ s\cdot f = (f\circ f)\ s$. However, maybe contrary to the intuition, the input and output arity of the function f need not match. For example, let $f = \{\texttt{dup}\}$ which takes zero elements from the stack and produces one, thus $[\![\texttt{\{dup\} twice}]\!]\ s\cdot x = [\![\texttt{dup}]\!]\circ[\![\texttt{dup}]\!]\ s\cdot x = s\cdot x\cdot x\cdot x$.

2.8 Conditional expression

Now let us introduce a simple conditional expression. It consumes three elements from the stack: one Boolean value and two more elements. The Boolean value represents a condition, based on which one of the other two elements is pushed back to the stack. We call this operation choose and its semantic definition is given in Table 9.

To define choose, we also introduced a special utility function cond which has the following signature

$$(\mathbf{Stack}\to\mathbf{Stack})\times(\mathbf{Stack}\to\mathbf{Stack})\times(\mathbf{Stack}\to\mathbf{Stack})\to(\mathbf{Stack}\to\mathbf{Stack}).$$

The function cond takes three stack manipulating functions and combines them into a new one. Here, the idea of applying $\mathsf{cond}(f,g,h)$ to the given stack s is as follows. First, the function f is applied to s, and the top element of the resulting stack s' is checked: if it is **true** $\in$ **Bool** or **false** $\in$ **Bool** then g or h is applied on s', respectively.

Table 9
Semantics of the choose conditional operator

$$[\![\texttt{choose}]\!]\ s{\cdot}b{\cdot}x{\cdot}y = \mathsf{cond}(\mathsf{push}\ b, \mathsf{push}\ x, \mathsf{push}\ y)\ s$$

where

$$\mathsf{cond}(f,g,h)\ s = \begin{cases} g\ s', & \text{if } f\ s = s'{\cdot}\textbf{true};\\ h\ s', & \text{if } f\ s = s'{\cdot}\textbf{false};\\ \bot, & \text{otherwise.} \end{cases}$$

Notice that, the function cond may not be defined when the function f does not leave a Boolean value on the top of the stack; however, we are only interested in cases when it does. Now, let us define a total function and consider cases when cond is total.

Definition 1. *A function f is total if $f\ s = \bot$ if and only if $s = \bot$.*

Lemma 1. *Let f, g, and h be total functions on **Stack**, $f,g,h : \textbf{Stack} \to \textbf{Stack}$. If the application of f always produces a Boolean value from **Bool** on the top of stack, then the function $\mathsf{cond}(f,g,h)$ is also total.*

Proof. Consider the definition of cond from Table 9. Since, f is total we have $s \neq \bot \implies f\ s \neq \bot$, and, by assumption, either $f\ s = s'\ {\cdot}\textbf{true}$ or $f\ s = s'\ {\cdot}\textbf{false}$. In the former, we have $\mathsf{cond}(f,g,h)\ s = g\ (\mathsf{pop}\ (f\ s))'$, and, in the latter, we have $\mathsf{cond}(f,g,h)\ s = h\ (\mathsf{pop}\ (f\ s))'$. □

Since, the function push b (used in the definition of choose) always pushes a Boolean on the stack, we have the following corollary.

Corollary. *The semantic clause $[\![\texttt{choose}]\!]\ s{\cdot}b{\cdot}x{\cdot}y$ is a total function.*

Proof. Observe, that all push b, push x, and push y are total functions. Moreover, push b leaves $b \in \textbf{Bool}$ on the top of the stack. Now, use Theorem 2.8. □

We can also observe this if we simplify the definition of the choose operation like this:

$$[\![\texttt{choose}]\!]\ s{\cdot}b{\cdot}x{\cdot}y = \mathsf{cond}(\mathsf{push}\ b, \mathsf{push}\ x, \mathsf{push}\ y)\ s = \begin{cases} s{\cdot}x, & \text{if } b = \textbf{true};\\ s{\cdot}y, & \text{otherwise.} \end{cases}$$

Obviously, the cond function is quite versatile and consequently used as a basis in other conditional and looping programming constructs. For example, let us introduce an operation $\texttt{if} \equiv \texttt{choose apply}$ and its semantics as

$$\begin{aligned}
[\![\texttt{if}]\!]\ s{\cdot}b{\cdot}f{\cdot}g &= [\![\texttt{choose apply}]\!]\ s{\cdot}b{\cdot}f{\cdot}g = [\![\texttt{apply}]\!] \circ [\![\texttt{choose}]\!]\ s{\cdot}b{\cdot}f{\cdot}g \\
&= [\![\texttt{apply}]\!] \circ \mathsf{cond}(\mathsf{push}\ b, \mathsf{push}\ f, \mathsf{push}\ g)\ s = \mathsf{cond}(\mathsf{push}\ b, f, g)\ s \\
&= \begin{cases} f\ s, & \text{if } b = \textbf{true}; \\ g\ s, & \text{otherwise.} \end{cases}
\end{aligned}$$

We also observe similar corollary.

Corollary. *The semantic clause* $[\![\texttt{if}]\!]\ s{\cdot}b{\cdot}f{\cdot}g$ *is a total function if* f *and* g *are total functions.*

Proof. Let $b \in \textbf{Bool}$ be a Boolean expression. From the semantics of $[\![\texttt{if}]\!]$ we see that either $[\![\texttt{if}]\!]\ s{\cdot}b{\cdot}f{\cdot}g = f\ s$ and $[\![\texttt{if}]\!]\ s{\cdot}b{\cdot}f{\cdot}g = g\ s$, which are both total if f and g are total. □

2.9 Iteration

In imperative languages, one of the more general programming constructs supporting iteration is a while loop. In this section, we consider a similar construct for our language.

Intuitively the `while` operation expects two functions on the stack: a loop condition followed by a loop body. It then executes the condition, which should push a Boolean value on the stack. The top of the stack is then checked and consumed: if it equals **false** the iteration ends, otherwise if it equals **true** the body is executed and the process is repeated.

To define a semantics of the `while` operation we employ similar idea as is presented in [14]. In particular, we first rewrite `while` using `if` operation, i.e.,

$$\{C\}\ \{B\}\ \texttt{while} = C\ \{B\ \{C\}\ \{B\}\ \texttt{while}\}\ \{\}\ \texttt{if}$$

and proceed as follows

$$\begin{aligned}
[\![\{C\}\ \{B\}\ \texttt{while}]\!]\ s &= [\![C\ \{B\ \{C\}\ \{B\}\ \texttt{while}\}\ \{\}\ \texttt{if}]\!]\ s \\
&= [\![\texttt{if}]\!] \circ [\![\{\}]\!] \circ [\![\{B\ \{C\}\ \{B\}\ \texttt{while}\}]\!] \circ [\![C]\!]\ s \\
&= [\![\texttt{if}]\!]\ ([\![C]\!]\ s){\cdot}[\![B\ \{C\}\ \{B\}\ \texttt{while}]\!]{\cdot}\mathrm{id} \\
&= [\![\texttt{if}]\!]\ ([\![C]\!]\ s){\cdot}([\![\{C\}\ \{B\}\ \texttt{while}]\!] \circ [\![B]\!]){\cdot}\mathrm{id} \\
&= \mathrm{cond}([\![C]\!], [\![\{C\}\ \{B\}\ \texttt{while}]\!] \circ [\![B]\!], \mathrm{id})\ s
\end{aligned}$$

Unfortunately, we cannot use this equation as a denotational clause because it is not a compositional definition, but if we denote $h = [\![\{C\}\ \{B\}\ \texttt{while}]\!]$, we can rewrite it as

$$h\ s = \mathrm{cond}([\![C]\!], h \circ [\![B]\!], \mathrm{id})\ s$$

and see that h is a least fixed point of a functional F defined by

$$F\ h = \mathrm{cond}([\![C]\!], h \circ [\![B]\!], \mathrm{id}).$$

Now note that $[\![\{C\}\ \{B\}\ \texttt{while}]\!]\ s = [\![\texttt{while}]\!]\ s \cdot [\![C]\!] \cdot [\![B]\!]$ and summarize the semantics of `while` in Table 10. The functionality of the auxiliary function FIX is

$$\mathsf{FIX} : ((\mathbf{Stack} \to \mathbf{Stack}) \to (\mathbf{Stack} \to \mathbf{Stack})) \to (\mathbf{Stack} \to \mathbf{Stack}).$$

The $\mathsf{FIX}\ F$ function thus returns the least fixed point of F, i.e., $F(\mathsf{FIX}\ F) = \mathsf{FIX}\ F$ and if $F\ g = g$ then $\mathsf{FIX}\ F$ is smaller than g. We refer the reader to [14] for the details.

Table 10
Semantics of the `while` operation

$[\![\texttt{while}]\!]\ s{\cdot}f{\cdot}g = \mathsf{FIX}\ F$
where
$F\ h = \mathsf{cond}(f, h \circ g, \mathsf{id})$

3 Extensions and discussion

In this subsection, we discuss some features of and possible extensions to the proposed programming language. Our language only includes values of three types, i.e., **Int**, **Bool**, and **Fun**. However, an additional type could straightforwardly be added similarly as we introduced these three by defining a new semantic domain and corresponding primitive operations.

For example, to support a list data structure in KKJ, we can define a new semantic domain $\mathbf{List} = \mathbf{Value}^*$ with additional primitive operations such as `head`, `tail`, `cons` defined similarly as in many functional programming languages.

Instead of this, we would rather propose another research direction where functions take the role of lists. In particular, the content of the list represented by a function are the elements which are pushed to the stack if the function is applied. Different functions may represent the same list, e.g., {`1 2`} and {`1 dup dup add`} both push 1 and 2 on the stack.

Concatenation of lists thus corresponds to a composition of functions, i.e., the `compose` operation. Furthermore, to prepend an element to a list we define

$$\texttt{cons} \equiv \texttt{swap quote swap compose}.$$

Its semantics is evaluated as

$$\begin{aligned}
[\![\texttt{cons}]\!]\ s{\cdot}x{\cdot}f &= [\![\texttt{swap quote swap compose}]\!]\ s{\cdot}x{\cdot}f \\
&= [\![\texttt{quote swap compose}]\!]\ s{\cdot}f{\cdot}x \\
&= [\![\texttt{swap compose}]\!]\ s{\cdot}f{\cdot}\mathsf{push}\ x \\
&= [\![\texttt{compose}]\!]\ s{\cdot}\mathsf{push}\ x{\cdot}f \\
&= s{\cdot}(f \circ \mathsf{push}\ x)
\end{aligned}$$

Clearly, the resulting list (function on the stack) first pushes the prepended element x and afterwards also the elements represented by f.

Another useful programming operation is to take one or more elements from the stack and store them in a list for later manipulation. We refer to such operations as *stack packing* and we already introduced one such operation, i.e., `quote`.

Using functions as lists we can easily build operation which packs an arbitrary number of the top stack elements. The idea lies in the repeated use of `cons`. For example: `quote2` $\equiv$ `quote cons` and `quote3` $\equiv$ `quote cons cons` pack two and three elements from the stack into a list, respectively. Again let us evaluate the semantics of `quote2`

$$\begin{aligned}
[\![\texttt{quote2}]\!]\ s{\cdot}x{\cdot}y &= [\![\texttt{quote cons}]\!]\ s{\cdot}x{\cdot}y \\
&= [\![\texttt{cons}]\!]\ s{\cdot}x{\cdot}\mathsf{push}\ x \\
&= s{\cdot}(\mathsf{push}\ y \circ \mathsf{push}\ x)
\end{aligned}$$

Besides that, we can also access an arbitrary element of the stack and push it on the top. We refer to this as *stack picking* and we already have operations to pick the top and the element below the top, i.e., `dup` and `over`, respectively. To construct the operation which picks the element two positions below the stack top, we simply use `pick2` $\equiv$ `quote2 over applyOver`, and similarly `pick3` $\equiv$ `quote3 over applyOver` to pick the element three positions below the stack top. Here, the semantics of `pick2` is

$$\begin{aligned}
[\![\texttt{pick2}]\!]\ s{\cdot}x{\cdot}y{\cdot}z &= [\![\texttt{quote2 over applyOver}]\!]\ s{\cdot}x{\cdot}y{\cdot}z \\
&= [\![\texttt{over applyOver}]\!]\ s{\cdot}x{\cdot}(\mathsf{push}\ z \circ \mathsf{push}\ y) \\
&= [\![\texttt{applyOver}]\!]\ s{\cdot}x{\cdot}(\mathsf{push}\ z \circ \mathsf{push}\ y){\cdot}x \\
&= s{\cdot}x{\cdot}y{\cdot}z{\cdot}x
\end{aligned}$$

Most of the programming languages also support the assignment of values to names to ease the task of programming to the users. A somewhat standard technique to do this is to extend the state to contain also the definitions of assignments, i.e., **State** $=$ **Stack** $\times$ **Defs** where **Defs** $=$ **Name** $\to$ **Fun**. Similarly, the semantic clauses of existing programming constructs must be extended to work on the **State**, i.e., on its first component, while a new programming construct to support assignment must

also be defined. We refer the reader to [20] for a more detailed presentation on the technique supporting also variable scoping in an imperative language.
An interesting further research direction is to exploit the parallelism which is inherently present in concatenative programs. Both concatenation and composition are associative, hence, any order of execution, as well as parallel execution, produce the same result. For example

$$\begin{aligned}&[\![\texttt{3 4 add dup ispos 5 6 swap choose mul}]\!]\, s = \\ &\quad = [\![\texttt{5 6 swap choose mul}]\!] \circ [\![\texttt{3 4 add dup ispos}]\!]\, s\end{aligned}$$

Clearly, all operations are considered to be pure, i.e., without side effects. Associativity may also be exploited for program optimization and refactoring.

Finally, the implementation of a developer toolchain for concatenative language is quite straightforward. The syntax exhibits great simplicity and thus support easy parsing. Here, an interesting example is the Forth programming language parser which supports constructs to change itself. Additionally, stack-based evaluation is also straightforward as the stack is a standard structure directly supported by almost all modern computer architectures. Moreover, even the anonymous functions syntactically represented by quotations are easily represented and manipulated by pointers. To support these claims we developed a parser and evaluation engine of the KKJ language with some extensions (e.g. more primitive operations and support for function definitions) in the Haskell programming language. The corresponding source code is freely available under a permissive license at https://www.github.com/jurem/kkj-lang.

Conclusions

In this paper, we proposed and examined a simple yet powerful programming language which exhibits properties of both low-level, e.g., assembly, and, high-level, e.g., functional languages. Additionally, the term-rewriting view on the evaluation puts the language among functional ones while stack-based view puts it among imperative ones. We strongly believe that these dichotomies represent a great advantage to such languages and many features and properties are yet waiting to be discovered.
The discussed language syntactically and semantically belongs to a group of concatenative and compositional programming languages, respectively. We formally defined the syntax as well as the semantics of the language, and, hence, formed a basis for further theoretical investigation of concatenative programming languages whose formal treatment received little attention in the scientific literature.
Despite this, various variants of such languages are already present (however, usually missing the quotation construct) in the mainstream industry mostly as intermediate languages or as virtual-machine byte-code. In particular Java Virtual Machine (JVM), Common Language Infrastructure (CLI) and the Python virtual machine (PVM) are typical examples of common stack-based virtual machines. Even, until version 5.0, Lua's virtual machine was a stack-based machine; however, 5.0's virtual machine is a register machine. Another example is the GraalVM (virtual machine allowing polyglot features for JVM, Python and other languages).

Our main goal in the paper was to select the prominent features of concatenative languages and formally describe their meaning using tools and techniques from the field of denotational semantics. As we feel that the area deserves more scientific attention we also identified and proposed several possible research directions.

Acknowledgement

The first author was supported by the Slovak Academic Information Agency under the National Scholarship program during his research stay in 2018. The second and third authors were supported by the project KEGA 011TUKE-4/2020: "A development of the new semantic technologies in educating of young IT experts".

References

[1] A. Aiken, J. H. Williams, and Wimmers. "The FL project: The design of a functional language", 1991.

[2] M. Benčík and L. Dedera. "Natural semantics of battle management languages". In *2019 Communication and Information Technologies (KIT)*, pp. 1-4, 2019.

[3] O. Danvy and L. R. Nielsen. "Refocusing in reduction semantics". BRICS, Department of Computer Science, University of Aarhus, 2004.

[4] S. Diehl, P. Hartel, and P. Sestoft. "Abstract machines for programming language implementation". *Future Generation Computer Systems*, 16(7):739-751, 2000.

[5] A. K. Wright and M. Felleisen. "A syntactic approach to type soundness". *Journal Information and Computation*, 115(1):38-94, 1994.

[6] P. Haller and H. Miller. "A reduction semantics for direct-style asynchronous observables". *Journal of Logical and Algebraic Methods in Programming*, 105:75-111, 2019.

[7] D. Herzberg and T. Reichert. "Concatenative programming – An Overlooked Paradigm in Functional Programming". In *Proceedings of the 4th International Conference on Software and Data Technologies*, pp. 257-262, 2009.

[8] J.M.E. Hyland and C.-H.L. Ong. "On full abstraction for PCF: I, II, and III". *Information and Computation*, 163(2):285-408, 2000.

[9] T. Jones and M. Homer. "The practice of a compositional functional programming language". In *Proceedings of the 16th Asian Symposium on Programming Languages and Systems*, 2018.

[10] G. Kahn. "Natural semantics". In *Proceedings of the 4th Annual Symposium on Theoretical Aspects of Computer Science*, STACS '87, page 22-39, Berlin, Heidelberg, 1987. Springer-Verlag.

[11] R. A. Kemmerer. "Hoare's axiomatic semantics". In *roceedings of ACM SIGSOFT International Symposium on Software Testing and Analysis*, Clearwater Beach, Florida, 1997. ACM Press.

[12] C. Moore. "Forth: a new way to program a mini-computer". *Astronomy and Astrophysics Supplement*, Vol. 15, pp. 497-511, 1974.

[13] P. D. Mosses. "Theory and practice of action semantics". In *MFCS '96, Proc. 21st Int. Symp. on Mathematical Foundations of Computer Science*, pages 37-61. Springer-Verlag, 1996.

[14] H. R. Nielson and F. Nielson. *"Semantics with Applications: An Appetizer"*. Springer-Verlag London, 2007.

[15] S. Pestov, D. Ehrenberg, and J. Groff. "Factor: A dynamic stack-based programming language". In *Proceedings of the 6th Symposium on Dynamic Languages*, DLS '10, ACM, New York, NY, USA, pp. 43-58, 2010.

[16] B. C. Pierce. *"Types and Programming Languages"*. The MIT Press, 1st edition, 2002.

[17] G. D. Plotkin. "The origins of structural operational semantics". *J. Log. Algebr. Program.*, Vol. 60-61, pp. 3-15, 2004.

[18] D. A. Schmidt. *"Denotational semantics. Methodology for Language Development"*. Allyn and Bacon, 1986.

[19] K. Slonneger and B. Kurtz. *"Formal Syntax and Semantics of Programming Languages: A Laboratory Based Approach"*. Addison-Wesley Longman Publishing Co., Inc., USA, 1st edition, 1995.

[20] W. Steingartner, V. Novitzká, M. Bačíková, and Š. Korečko. "New approach to categorical semantics for procedural languages". *Computing and Informatics*, 36(6):1385-1414, 2017.

[21] S. Szymoniak. "Security protocols analysis including various time parameters". *Mathematical Biosciences and Engineering*, 18(2): 1136-1153, 2021.

[22] O. Siedlecka-Lamch, S. Szymoniak, M. Kurkowski, I. El Fray. Towards Most Efficient Method for Untimed Security Protocols Verification. In: *Proceedings of the 24th Pacific Asia Conference on Information Systems: Information Systems (IS) for the Future, PACIS 2020.* 2020

[23] M. von Thun. "Joy: Forth's functional cousin". In *In Proceedings from the 17th EuroForth Conference*, 2001.

[24] D. A. Watt. "Action Semantics in Retrospect". In Palsberg J. (eds) *Semantics and Algebraic Specification. Lecture Notes in Computer Science*, Vol. 5700, Springer, Berlin, Heidelberg, 2009.

[25] A. L. Zackery, S. Perugini. "An Introduction to Concatenative Programming in Factor". *J. Comput. Sci. Coll.* 35(5):70-77, 2019. Consortium for Computing Sciences in Colleges, Evansville, IN, USA.