Informatics Control Measurement
in Economy and Environment Protection

2/2021

Informatyka Automatyka Pomiary

w gospodarce i ochronie środowiska

www.e-IAPGOS.pl

ISSN 2083-0157

Kwartalnik Naukowo-Techniczny

fot. Oleh Kaptiukh

National University "Zaporizhzhia Polytechnic"
(Zaporizhzhia, Ukraine)

15th International Conference on Diagnostics of Processes and Systems

5 - 7 September 2022, Kashubia, Poland

Important dates

25th of March 2022
Full-length Paper Submission Deadline

25th of March 2022
Invited and Special Sessions Proposals

1st May 2022
Paper Acceptance Notification

15th May 2022
Final Paper Submission

20th May 2022
Early Registration Deadline

30th June 2022
Regular Registration Deadline

5th September 2022
Symposium Start Date

7th September 2022
Symposium End Date

Organizing committee

IPC Chair
Prof. Zdzisław Kowalczuk, Gdańsk University of Technology
(kova@pg.edu.pl)

Co-Chair
Prof. Józef Korbicz, University of Zielona Góra
(j.korbicz@issi.uz.zgora.pl)

Co-Chair
Prof. Jan M. Kościelny, Warsaw University of Technology
(jmk@mchtr.pw.edu.pl)

NOC Chair
PhD Marek Tatara, GUT
(martatar@pg.edu.pl)

NOC Co-Chair
DSc Anna Witkowska, GUT,
(anna.witkowska@pg.edu.pl)

DPS'2022 Secretary

Department of Robotics and Decision SystemsFaculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
ul. Narutowicza 11/12, 80-233 Gdańsk, Poland
phone: +48 58 347 2018, +48 58 347 2289
fax: +48 58 347 2018, +48 58 348 6373
email: dps2022@konsulting.gda.pl
http://www.konsulting.gda.pl/dps2022
(to be open)

Call for papers

The **15th International Conference on Diagnostics of Processes and Systems (DPS)**, hosted by the Gdańsk University of Technology, will be held in Kashubia (Pomerania, Poland), on September 5-7, 2022.

Conference history and motivation

In the last few decades, the diagnostics of processes and systems has developed significantly both in theory and in practice. Both the scope and applications of technical diagnostics cover many well-established topics, and numerous emerging advances in the fields of control and systems engineering, robotics and mechatronics, aerospace, applied mathematics, artificial intelligence, decision support, statistics and signal processing. More and more effective methods of fault diagnosis and their various applications are considered in the classical areas of electrical, mechanical, medical and chemical as well as in inter-sort or completely new fields of human activity. These methods are of key importance for the early identification of emerging failures and ensuring the safe, economic and ecological operation of machines, devices and systems. In the field of automation and robotics, special attention is paid to fault-tolerant control, which combines diagnostics with control methods to deal with faults intelligently and efficiently, thus reducing the risk to the safety of systems.

The International Conference on Diagnostics of Processes and Systems (initially called Diagnostics of Industrial Processes) has been organized since 1996 (usually every two years) by the Warsaw University of Technology, University of Zielona Góra and Gdańsk University of Technology.

This series of conferences is an excellent forum for the exchange of knowledge and experience, discussion of scientific and technological challenges, showing open problems and future research directions in the broad field of diagnostics and monitoring. The conference is also an opportunity for young people to acquire specialist knowledge, experience the atmosphere of science and present and discuss their own theoretical research results or practical applications.

An important task of this forum is the integration of scientists, engineers and managers from various industries and services, as well as developers of hardware and software for computer control systems, signal processing and diagnostics. The subjects of the DPS conference are in line with the topics of the IFAC Symposia on Fault Detection, Supervision and Safety of Technical Processes (SAFEPROCESS), as well as the International Conferences on Control and Fault-Tolerant Systems (SysTol).

Location and history: Kashubia (Pomerania):

Pomerania, including the Kashubian Cultural Area, is a geographic region in northern Poland. In its thousand-year history, after the territory was taken over by the Teutonic Knights in 1308-1454, it was then regained by Poland. In 1772 it was annexed to Prussia. Only a small part of it was assigned to Poland after World War I, but after World War II the entire Pomeranian region was returned to Poland.

Gdańsk Pomerania is a culturally rich and picturesque region of Poland. The indigenous people of the region are the Slavic Kashubians who speak the Kashubian dialect of the Pomeranian language, Kociewiacy and Borowiacy. Kashubia hide many secrets and undiscovered stories. You can admire unusual villages full of white houses cut with a brown chessboard beam, sculptures of local artists, cemeteries of ancient tribes, Lobel lakes and Pomeranian hydrotechnical monuments. Kashubia is not only a land with a unique history, traditions and customs cultivated for centuries, but also one of the most beautiful and picturesque natural regions of Poland. The central and southern part of Kashubia is covered with numerous lakes, dense forests and hills. At 329 m above sea level, Wierzyca is the highest peak in the entire European Lowlands. The northern part of Kashubia is dominated by a classic seaside landscape - beaches, lighthouses, fishermen's huts, seaside manors and palaces. Moving dunes in the Słowiński National Park are a must-see in the northern part of Kashubia. Historic Gdańsk, the spa town of Sopot and modern Gdynia constitute one unique Tri-City, where everyone will find something for themselves. The main city of the region – Gdańsk, is on the list of the 20 fastest growing cities in the EU.

Original and high-quality articles will be reviewed by the members of the International Program Committee and published in the conference proceedings in the Springer Verlag Lecture Notes series. They will also be submitted for indexing by the Web of Science. As before, Polish works will be published in the form of a book by PWNT.

We are looking forward to meeting you in Gdańsk

POLITECHNIKA GDAŃSKA POLSPAR tkp O.GDAŃSK

2/2021

kwiecień – czerwiec

Wydanie pod redakcją naukową
prof. dr hab. inż. Waldemara Wójcika

INFORMATYKA AUTOMATYKA POMIARY

W GOSPODARCE i OCHRONIE ŚRODOWISKA

Informatics Control Measurement in Economy and Environment Protection

p-ISSN 2083-0157, e-ISSN 2391-6761, www.e-iapgos.pl

EDITOR STAFF
ZESPÓŁ REDAKCYJNY

Editor-in-Chief
Redaktor naczelny

Paweł KOMADA
Lublin University of Technology, Lublin, Poland
p.komada@pollub.pl

Deputy Editors
Zastępcy redaktora

Jan SIKORA
Research and Development Center Netrix S.A., Lublin, Poland sik59@wp.pl

Dominik SANKOWSKI
Lodz University of Technology, Lodz, Poland
dsan@kis.p.lodz.pl

Pavel FIALA
Brno University of Technology, Brno, Czech Republic fialap@feec.vutbr.cz

Andrzej SMOLARZ
Lublin University of Technology, Lublin, Poland
a.smolarz@pollub.pl

Technical Editor
Redaktor techniczny

Tomasz ŁAWICKI
Lublin University of Technology, Lublin, Poland
t.lawicki@pollub.pl

Statistical Editor
Redaktor statystyczny

Ewa ŁAZUKA
Lublin University of Technology, Lublin, Poland
e.lazuka@pollub.pl

EDITORIAL OFFICE
REDAKCJA

Redakcja czasopisma
Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska
Katedra Elektroniki i Technik Informacyjnych
Politechnika Lubelska
ul. Nadbystrzycka 38A, 20-618 Lublin
tel. +48 81 53 84 309,
fax: +48 81 53 84 312
iapgos@pollub.pl
www.e-iapgos.pl
iapgos.pollub.pl
ph.pollub.pl/index.php/iapgos

PUBLISHER
WYDAWCA

Politechnika Lubelska
ul. Nadbystrzycka 38D
20-618 Lublin
tel. +48 81 53 84 100
www.pollub.pl
ph.pollub.pl

EDITORIAL BOARD
KOMITET REDAKCYJNY

Editor-in-Chief
Redaktor naczelny

Paweł KOMADA
Lublin University of Technology, Lublin, Poland
p.komada@pollub.pl

Topical Editors
Redaktorzy działowi

Electrical Engineering
Elektrotechnika

Jan SIKORA
Research and Development Center Netrix S.A., Lublin, Poland sik59@wp.pl

Computer Science
Informatyka

Dominik SANKOWSKI
Lodz University of Technology, Lodz, Poland
dsan@kis.p.lodz.pl

Electronics
Elektronika

Pavel FIALA
Brno University of Technology, Brno, Czech Republic fialap@feec.vutbr.cz

Automatic
Automatyka

Waldemar WÓJCIK
Lublin University of Technology, Lublin, Poland
waldemar.wojcik@ pollub.pl

Environmental Engineering
Inżynieria środowiska

Lucjan PAWŁOWSKI
Lublin University of Technology, Lublin, Poland
l.pawlowski@pollub.pl

Mechtronics
Mechatronika

Krzysztof KLUSZCZYŃSKI
Silesian University of Technology, Gliwice, Poland krzysztof.kluszczynski@polsl.pl

INTERNATIONAL PROGRAMME COMMITTEE
RADA PROGRAMOWO-NAUKOWA

Chairman
Przewodniczący

Waldemar WÓJCIK
Lublin University of Technology, Lublin, Poland

Deputy of Chairman
Zastępca przewodniczącego

Jan SIKORA
Research and Development Center Netrix S.A., Lublin, Poland

Members
Członkowie

Kazimierz ADAMIAK
University of Western Ontario, Ontario, Canada

Darya ALONTSEVA
D.Serikbaev East Kazakhstan State Technical University, Ust-Kamenogorsk, Kazakhstan

Shin-ichi AOQUI
Sojo University, Kumamoto, Japan

Javier BALLESTER
Universidad de Zaragoza, Saragossa, Spain

Yurii BOBALO
Lviv Polytechnic National University, Lviv, Ukraine

Oleksy BORYSENKO
Department of Elektronics and Computer Technics, Sumy, Ukraine

Hartmut BRAUER
Technische Universität Ilmenau, Ilmenau, Germany

Kathleen CURRAN
School of Medicine & Medical Science, Dublin, Ireland

Milan DADO
University of Žilina, Žilina, Slovakia

Jarmila DEDKOVA
Brno University of Technology, Brno, Czech Republic

Andrzej DEMENKO
Poznan University of Technology, Poznań, Poland

Pavel FIALA
Brno University of Technology, Brno, Czech Republic

Vladimir FIRAGO
Belarusian State University, Minsk, Belarus

Ryszard GOLEMAN
Lublin University of Technology, Lublin, Poland

Jan GÓRSKI
AGH University of Science and Technology, Cracow, Poland

Stanisław GRATKOWSKI
West Pomeranian University of Technology Szczecin, Szczecin, Poland

Antoni GRZANKA
Warsaw University of Technology, Warsaw, Poland

Jeni HEINO
Helsinki University of Technology, Helsinki, Finland

Oleksandra HOTRA
Lublin University of Technology, Lublin, Poland

Wojciech JARZYNA
Lublin University of Technology, Lublin, Poland

Mukhtar JUNISBEKOV
M.Kh. Dulaty Taraz State University, Taraz, Kazakhstan

Piotr KACEJKO
Lublin University of Technology, Lublin, Poland

Krzysztof KLUSZCZYŃSKI
Silesian University of Technology, Gliwice, Poland

Yurii KRAK
Taras Shevchenko National University of Kyiv, Kiev, Ukraine
Piotr KSIĄŻEK
Medical University of Lublin, Lublin, Poland
Piotr LESIAK
University of Economics and Innovation in Lublin Lublin, Poland
Volodymyr LYTVYNENKO
Kherson National Technical University, Kherson, Ukraine
Artur MEDVIED
Riga Technical University, Riga, Latvia
Paweł MERGO
Maria Curie-Skłodowska University, Lublin, Poland
Zbigniew OMIOTEK
Lublin University of Technology, Lublin, Poland
Andrzej NAFALSKI
University of South Australia, Adelaide, Australia
Il Han PARK
Sungkyunkwan University, Suwon, Korea
Lucjan PAWŁOWSKI
Lublin University of Technology, Lublin, Poland
Sergey PAVLOV
Vinnytsia National Technical University, Vinnytsia, Ukraine
Denis PREMEL
CEA Saclay, Gif-sur-Yvette, France
Jason RILEY
The Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, USA
Ryszard ROSKOSZ
Gdańsk University of Technology, Gdańsk, Poland
Tomasz RYMARCZYK
Research and Development Center Netrix S.A., Lublin, Poland
Dominik SANKOWSKI
Lodz University of Technology, Lodz, Poland
Stanislav SLOSARCIK
Technical University of Kosice, Kosice, Slovakia
Jan SROKA
Warsaw University of Technology, Warsaw, Poland
Bohdan STADNYK
Lviv Polytechnic National University, Lviv, Ukraine
Henryka Danuta STRYCZEWSKA
Lublin University of Technology, Lublin, Poland
Batyrbek SULEMENOV
Kazakh National Research Technical University after K.I.Satpayev, Almaty, Kazakhstan
Mirosław ŚWIERCZ
Bialystok University of Technology, Białystok, Poland
Stanisław TARASIEWICZ
Université Laval, Quebec, Canada
Murielle TORREGROSSA
University of Strasbourg, Strasbourg, France
Sławomir TUMAŃSKI
Warsaw University of Technology, Warsaw, Poland
Andrzej WAC-WŁODARCZYK
Lublin University of Technology, Lublin, Poland
Zygmunt WARSZA
Industrial Research Institute for Automation and Measurements, Warsaw, Poland
Sotoshi YAMADA
Kanazawa University, Kanazawa, Japan
Xiaoyi YANG
Beihang University, Beijing, China
Mykola YERMOSHENKO
International Academy of Information Sciences, Kiev, Ukraine
Athanasios ZACHAROPOULOS
University College London, London, United Kingdom
Ivan ZHARSKI
Belarusian National Technical University, Minsk, Belarus
Cao ZHIHONG
Institute of Soil Science Chinese Academy of Sciences, Nanjing, China
Paweł ŻUKOWSKI
Lublin University of Technology, Lublin, Poland

PRINTING HOUSE – DRUKARNIA

DjaF – Naświetlarnia B1+
ul. Kmietowicza 1/1
30-092 Kraków
http://www.djaf.pl
nakład: 100 egzemplarzy

OTHER INFORMATION – INNE INFORMACJE

Czasopismo jest indeksowane w bazach:

DOAJ:	doaj.org
BazTech:	baztech.icm.edu.pl
IC Journals Master List:	www.journals.indexcopernicus.com
Google Scholar	scholar.google.pl
POL-index	pbn.nauka.gov.pl
Sherpa RoMEO	www.sherpa.ac.uk

Czasopismo *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska* zostało objęte finansowaniem przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach programu *Wsparcie dla czasopism naukowych* w latach 2019-2020.

Czasopismo znajduje się w wykazie czasopism naukowych opublikowanym w Komunikacie Ministra Edukacji i Nauki z dnia 9 lutego 2021 r., Unikatowy Identyfikator Czasopisma: 200167 – z przypisaną liczbą punktów przyznawanych za publikację artykułu równą 20.

Zasady publikowania artykułów, przygotowania tekstów, zasady etyczne, procedura recenzowania, wykazy recenzentów oraz pełne teksty artykułów dostępne są na stronie internetowej czasopisma:

www.e-iapgos.pl

W celu zwiększenia oddziaływania czasopisma w środowisku naukowym redakcja zaleca:

- w artykułach publikowanych w IAPGOS cytować artykuły z renomowanych czasopism międzynarodowych (szczególnie indeksowanych w bazach Web of Science oraz Scopus) używając oficjalnych skrótów nazw czasopism,
- w artykułach publikowanych w innych czasopismach (zwłaszcza indeksowanych w bazach Web of Science oraz Scopus) cytować prace publikowane w IAPGOS – zwłaszcza posługując się numerami DOI, np.:
  Kluszczyński K. *Modelowanie – umiejętność czy sztuka?* Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska – IAPGOS, 1/2016, 4–15, DOI: 10.5604/20830157.1193833.

CONTENTS – SPIS TREŚCI

http://doi.org/10.35784/iapgos.2596

A STEP TOWARDS THE MAJORITY-BASED CLUSTERING VALIDATION DECISION FUSION METHOD

Taras Panskyi, Volodymyr Mosorov
Lodz University of Technology, Institute of Applied Computer Science, Lodz, Poland

Abstract. *A variety of clustering validation indices (CVIs) are aimed at validating the results of clustering analysis and determining which clustering algorithm performs best. Different validation indices may be appropriate for different clustering algorithms or partition dissimilarity measures; however, the best suitable index to use in practice remains unknown. A single CVI is generally unable to handle the wide variability and scalability of the data and cope successfully with all the contexts. Therefore, one of the popular approaches is to use a combination of multiple CVIs and fuse their votes into the final decision. This work aims to analyze the majority-based decision fusion method. Thus, the experimental work consisted of designing and implementing the NbClust majority-based decision fusion method and then evaluating the CVIs performance with different clustering algorithms and dissimilarity measures to discover the best validation configuration. Moreover, the authors proposed to enhance the standard majority-based decision fusion method with straightforward rules for the maximum efficiency of the validation procedure. The result showed that the designed enhanced method with an invasive validation configuration could cope with almost all data sets (99%) with different experimental factors (density, dimensionality, number of clusters, etc.).*

Keywords: clustering, clustering validation index, decision fusion method

KROK W KIERUNKU METODY FUZJI DECYZJI OPARTEJ NA WIĘKSZOŚCI DLA WALIDACJI WYNIKÓW KLASTERYZACJI

Streszczenie. *Różnorodne indeksy walidacji klasteryzacji (CVI) mają na celu walidację wyników analizy skupień i określenie, który algorytm klasteryzacji działa najlepiej. Różne indeksy walidacji mogą być odpowiednie dla różnych algorytmów klasteryzacji lub miar niepodobieństwa podziału; jednak najlepszy walidacyjny indeks do zastosowania w praktyce pozostaje nieznany. Pojedynczy CVI na ogół nie jest w stanie poradzić sobie z dużą zmiennością i skalowalnością danych oraz z powodzeniem poradzić sobie we wszystkich kontekstach. Dlatego jednym z popularnych podejść jest użycie kombinacji wielu CVIs i połączenie ich głosów w ostateczną decyzję. Celem tej pracy jest analiza metody fuzji decyzji opartej na większości. W związku z tym prace eksperymentalne polegały na zaprojektowaniu i wdrożeniu metody NbClust fuzji decyzji opartej na większości, a następnie ocenianie wydajności CVIs za pomocą różnych algorytmów klasteryzacji i miar niepodobieństwa w celu odkrycia najlepszej konfiguracji walidacji. Ponadto autor zaproponował rozszerzenie standardowej metody fuzji decyzji oparta na większości o proste reguły dla maksymalnej efektywności procedury walidacji. Wynik pokazał, że zaprojektowana ulepszona metoda z inwazyjną konfiguracją walidacji może poradzić sobie z prawie wszystkimi zbiorami danych (99%) z różnymi eksperymentalnymi parametrami (gęstość, wymiarowość, liczba klastrów itp.).*

Słowa kluczowe: klasteryzacja, indeks walidacji klasteryzacji, metoda fuzji decyzji

Introduction

Clustering is a process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have a high natural association among themselves while remaining relatively distinct from each other [3]. In general, the essence of cluster analysis assumes that little or nothing is known about the grouping structure which underlies the data set. The operational objective, in this case, is to discover the grouping data structure which is frequently described as a problem of finding "natural groups".

Many methods for cluster analysis have been developed in recent years and many of these methods have shortcomings and limitations in their practical use. It is difficult to provide a clear categorization of clustering methods because these categories may overlap so that a method may have features from several categories. Nevertheless, the major fundamental clustering methods can be classified into the following categories [9]: hierarchical methods formed [27, 47, 60], partitioning methods [15, 56, 62], density-based methods [10, 34], graph-based methods [1, 67], and grid-based methods [14, 17].

Different clustering algorithms usually lead to different partitions of data; even for the same algorithm, the selection of different input parameters may greatly affect the clustering results. Thus, effective evaluation standards and criteria are critically important to give the researcher confidence regarding the clustering results. The procedure of evaluating the correctness of clustering results is called cluster validation [32] and for a long time, it has been recognized as one of the vital problems essential to the success of data clustering.

It is usual to classify the cluster validation techniques under two groups — internal and external [35, 53]. External validation indices use external information not presented in the data to estimate the extent to which the clustering structure discovered by a clustering algorithm matches a certain external structure. On the other hand, internal indices evaluate the correctness of the clustering structure without reference to external information. Both external and internal validation indices are crucial for many application scenarios. However, there are still scenarios where clustering validation indices have limitations in estimating the correctness of clustering results. Examples include the case when external criteria are not available and internal indices are not robust enough. Moreover, despite the vast amount of expert endeavors spent on this issue, there is no consistent and conclusive solution to cluster validation. The multitude of different validation approaches creates an added difficulty, since results obtained using different methods cannot be compared in the same validation framework. Also, the relationship between different validation indices is not clear and has not been fully established.

A variety of indices aimed at validating the results of clustering analysis and determining which clustering algorithm performs best. However, the choice of the best or the most appropriate clustering validation index is strikingly similar to the dilemma of comparing and selecting the best classifiers in pattern recognition, where the no free lunch theorem rules that there is no universally best classifier [44]. Moreover, given the fact that different validation indices may be appropriate for different clustering algorithms or partition dissimilarity measures, the best suitable index to use in practice remains unknown. In the recent work by Gurrutxaga et al. [33], the authors accepted that there is no single way of establishing the quality of a partition by selecting the optimal validation index which would be more robust than the rest in all contexts and under different conditions. Therefore, Yera et al. [69] suggest using decision fusion validation strategies to obtain a more stable behavior which would make it possible for the user to avoid having to choose a different validation index for each particular environment.

In this work, the authors drew inspiration from the works of Arbelaitz et al. [2], Gurrutxaga et al. [33], Yera et al. [69], and the decision fusion method developed by Charrad et al. [16]. Since Charrad et al. [16] did not substantiate the use of a particular data, clustering algorithm, and a dissimilarity measure for the majority-based decision fusion voting procedure, the authors provide a comparative study of these factors. In the first part of the article, the authors propose to analyze Charrad's majority-based decision fusion method (MBDFM) using a *non-invasive configuration*. A non-invasive configuration relies on selecting the best MBDFM clustering algorithm and dissimilarity measure that works correctly in every experimental environment. Thereby, the authors try to achieve the best possible results for Charrad's MBDFM without changing the internal validation algorithm, but only its input validation parameters.

In the second part of this work, the authors propose to use an *invasive configuration* of Charrad's MBDFM. The authors hypothesize that the non-invasive configuration of MBDFM can be better than the default configuration of MBDFM (*k*-means algorithm and Euclidean distance), however, the authors suspect that this may not be enough to achieve the task, that is the revealing of the largest number of "true" clusters in the experimental setup. *What if the non-invasive configuration of MBDFM even with the best clustering algorithm and dissimilarity measure does not provide the expected satisfactory results*? The underlying idea of an invasive configuration is to change the MBDFM by interfering and modifying the internal Charrad's validation algorithm. Finally, the authors will show the difference between the MBDFM voting approaches with default, non-invasive and invasive configurations.

According to Arbelaitz et al. [2], there is no standard terminology and formalization for clustering validity indices; therefore, in this article, the abbreviation of CVI will be used for Cluster Validity Index. The next section discusses other works related to CVI comparison, in particular the examples of decision fusion methods found in the literature.

Since testing for revealing the data structure is the main objective of this article, the problem of choosing between the attribute space and the problem of discovering the optimal number of clusters will not be considered.

1. Related works

Most of the works that compare CVIs use the same approach: a set of CVIs is used to estimate the number of clusters in a set partitioned by several algorithms. Despite this widely used approach, most of the works are not comparable since they differ with respect to the compared CVIs, used data sets, or analysis results.

The paper published by Milligan and Cooper [52] compared 30 CVIs. The experiments were conducted using hierarchical clustering algorithms. They used 108 synthetic data sets with a varying number of non-overlapped clusters (2, 3, 4, or 5), dimensionality (4, 6, or 8), and cluster sizes. The same tabular format was used by Dubes [23]. Bezdek et al. [7] published a paper comparing 23 CVIs based on 3 runs of the EM algorithm and 12 synthetic data sets. Another study that compared 15 CVIs was performed by Dimitriadou et al. [22], based on 100 runs of the *k*-means algorithm for 162 data sets with binary attributes. Recently, Brun et al. [12] have compared 8 CVIs using several clustering algorithms: *k*-means, fuzzy *c*-means, SOM, single-linkage, complete-linkage, and EM, using 600 synthetic data sets with varying dimensionality (2 or 10), cluster shape, and number of clusters (2 or 4). Shim et al. [64] followed the Milligan and Cooper experiment but added certain CVIs or extended the study. Other CVI comparisons can be found where new CVIs are proposed; however, the experiments are usually limited to similar data sets comparing 5 or 10 CVIs [18, 37, 71]. The exception is a work by Arbelaitz et al. [2] based on the same Milligan and Cooper CVI framework, but with an extensive set of configurations (dissimilarity measure, data density, noise, overlapping clusters, etc.).

Since there is no universal CVI to always make a correct decision, many authors [44, 69] agreed to use multi-criteria solutions to reach the best and adequate results. Multi-criteria solutions assume the adoption of several CVIs to achieve greater certainty and correctness of clustering results. Bezdek and Pal [8] suggested a combined decision-based fusion strategy for all CVIs used in a validation process. This research includes the following decision-making methods and their rules: the mean, median, and mode rules. According to those rules, the final validation decision is made by the mean, median, or mode of CVIs that participated in the voting procedure. Later, the authors [44] showed a comparison of different fusion techniques of multiple CVIs. Moreover, Kryszczuk and Hurley [44] pointed that the best-performing scheme was the mean-rule decision fusion scheme. The recent work by Yera et al. [69] also discusses the use of decision fusion strategies for cluster validation purposes. The authors suggested two types of voting strategies: Global Voting and Selective Voting. Global Voting is a simple vote that fuses the decisions of all CVIs presented in the study. Selective Voting uses a limited group of CVIs for decision fusion. Moreover, three approaches were developed, each limiting the group of voting CVIs based on the following criteria: the global performance of the CVIs, their factor dependence success rate, and the impact the CVIs have on the results. The Yera et al. [69] decision fusion techniques presented above have certain critical disadvantages. The Global Voting approach was not even used in the comparative study due to the weakness of the archived results. This type of decision fusion technique is similar to the decision-making methods developed by Kryszczuk and Hurley. The Selective Voting technique with the global performance of CVIs did not meet the authors' expectations either, since the best vote could not beat the success rate of the best individual CVI. The Selective Voting technique with the factor dependence success rate of the CVIs beat the overall success rate of the best individual CVI. However, the improvement was slightly more robust than the best CVI involved in the voting procedure. Furthermore, the work uses a limited number of experimental factors (three clusters, at least 100 numbers of objects in each cluster, three dimensions, etc.) which can significantly affect the presented decision fusion approaches. Finally, these decision fusion strategies require weighting CVIs votes, which reduces the precision of estimated CVIs decisions.

Charrad et al. [16] suggested that other decision-making fusion methods should be used which a group of CVIs may use to seek a satisfying solution; namely, the authority rule and the majority rule. The authority rule refers to groups that have a leader, i.e. the main CVI which has the authority to make the ultimate decision for the entire group. Although the method can generate a final decision quickly, it does not encourage the maximization of the strengths of individual CVIs in the group [46]. The majority rule depends on an individual decision of each CVI, where the final decision is made by the majority of the total CVIs votes. This method delivers fast solutions and follows a clear rule of using independent CVIs in the validation process.

In light of the decision fusion techniques presented above, taking all of the above multi-criteria methods into consideration, the MBDFM has been chosen as a major scheme for further analysis and improvement.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

2. Tools for multi-criteria decision fusion clustering validation

A significant amount of software is available for data clustering validation purposes. Interestingly, most of the sophisticated software on clustering validation is open-source software, which is freely available at different Web sites. On the other hand, most of the commercial software comprises implementations of simpler and more classical validation solutions. This is because open-source software is often in the form of research prototypes, created by researchers, which reflect more recent advances in the clustering validation field. The clustering validation procedure has been implemented as packages in many software applications, such as *SAS*, *RapidMiner*, *MATLAB*, and *R*. However, only the *R* programming environment offers a large number of unique CVIs and different validation approaches. Moreover, in addition to the object-oriented nature of the language, implementing the CVIs within the *R* statistical programming framework provides the additional advantage in that it can interface with numerous clustering algorithms in existing *R* packages, and accommodate new algorithms as they are created and coded into R libraries.

There are several *R* packages that perform clustering validation and are available from https://www.r-project.org/. Indeed, Milligan and Cooper [52] examined thirty CVIs, with simulated artificial data, where the number of clusters was known beforehand. Eleven CVIs among them are available in the *R cclust* package [21], eight CVIs in the *clusterSim* package [68], two CVIs in *clv* [54]. The *clValid* package [11] includes 3 internal CVIs, 4 stability CVIs (special versions of internal indices), and 2 biological CVIs. The *cl_validity()* function in the *clue* package [42] performs validation for both partitioning and hierarchical methods using 3 CVIs, and the *fcclusterIndex()* function in package *e1071* [50] has built-in 7 fuzzy CVIs. The *cluster.stats()* function in the *fpc* package [41] uses 8 CVIs for clustering validation purposes. The package *NbClust* [16] gathered the 26 CVIs, several clustering algorithms with corresponding dissimilarity measures together to provide an exhaustive list of CVIs. Currently, NbClust is only one package that offers such a variety of CVIs, however, some indices examined in the Milligan and Cooper study were not implemented due to a lack of detailed CVI's explanation. Moreover, unlike the rest packages, the NbClust is quite flexible and offers the possibility of a fairly broad change of input parameters for further clustering validation purposes. Therefore, Charad's NbClust is chosen as the basic MBDFM for further modifications. For clarification purposes, instead of Charad's NbClust MBDFM the abbreviation MBDFM will be used in the remainder of this article.

3. Majority-based decision fusion method notation

This note studies a method of the CVI's decision formation and aims to explain the stylized fact that the support for one out of k clusters at stake often shows a high degree of CVI's heterogeneity and persistent cross-sectional variance that is only partly explained in clustering conditions. An intuitive explanation of this stylized fact is that each CVI may show a tendency to conform to the vote of what it perceives to be the best opinion. As postulated before that the behavior of each CVI in formulating the overall decision could be described by the majority rule. Although the authors do not model the voting process of CVI's in detail, but only show that the reduced forms are consistent with an explicit clustering validation foundation. The clustering validation theory does not explain how each CVI votes to conform to the majority. Moreover, the theory does not describe the growing or reducing CVI's tendency towards consensus in the decision scheme. Therefore, the first objective is to formulate the simple model of the majority-based decision fusion rule.

The CVIs decision formation process is defined as follows. Suppose the clustering algorithm run over the data set $\mathbb{X}$ with a set of m different values for the k parameter $K = \{k_1, k_2 \dots, k_m\}$, and let $\mathcal{P} = \{P_1, P_2, \dots, P_m\}$ be the k partitions. Suppose the clustering algorithm reveals the k_{est} to be the best number of clusters with the corresponding P_{est} partition. The true number of clusters k_{true} is known beforehand. Let CVI= $\{CVI_1, CVI_2, \dots, CVI_n\}$ be a set of n clustering validation indices CVI_i, where $i = 1 \dots n$ which are to be analyzed. Each CVI_i return the value of $CVI_i(P)$ for the proposed partition over all the partition $\mathcal{P}$. Moreover, the returned $CVI_i(P)$ value indicates the specific k_j used as an input parameter for the clustering algorithm. Each $CVI_i(P)$ value is counted as a vote for a particular k_j. Thereafter, the method counts the CVIs votes for each k_j and forms the CVI's decision groups DEC $= \{Dec_1, Dec_2, \dots, Dec_k\}$, where Dec_j is a sum of CVI's votes for particular k_j. From among the CVI's decision groups is formed the biggest group or the majority $Maj_{est} = max_{Dec \in \mathbb{N}}\{Dec_1, Dec_2, \dots, Dec_k\}$ that should identify the "true" estimated number of clusters $k_j = k_{est}$. However, only if the Maj_{est} identifies the $k_{est} = k_{true}$ clusters, the majority justifies that the estimated number of clusters is the "true" one, and hence $Maj_{est} = Maj_{true}$.

4. Experimental setup

Before analyzing and modifying the majority-based decision fusion method the experimental setup should be outlined. In this section, the authors describs the experiment setup including 24 CVIs: *Calinski-Harabasz index* [13], *J-index* [24], *pseudo T-squared* [25], *C-index* [43], *F-ratio* [6], *CCC criterion* [61], *Ptbiserial index* [51], *DB index* [19], *Frey index* [30], *Harigan index* [40], *Ratkowsky* and *Lance index* [57], *Scott* and *Symons index* [63], *Marriot index* [48], *Ball* and *Hall index* [4], *TrCovW* and *TraceW indices* [52], *Friedman* and *Rubin indi*ces [31], *McClain* and *Rao i*ndex [49], *KL index* [45], *Silhouette in*dex [59], *Dunn index* [26], *Halkidi indices* [38, 39]. Since the CVIs are compared in a wide variety of configurations, an experiment with several factors has been designed.

The authors' proposal follows, to a certain extent, the traditional problem of estimating the number of clusters in a data set, which was described well in Arbelaitz et al. [2]: to run a clustering algorithm over a data set with a set of different values for the k input parameter, to obtain a set of different partitions, and to evaluate each particular CVI for all obtained partitions. The detected number of clusters in the target partition yielding satisfactory results is considered a decision of the CVI for that particular data set. However, the decision is considered successful only if it justifies that the estimated number of clusters is "true".

Eight agglomerative hierarchical clustering algorithms were used to compute partitions from the data sets: *Ward*, *single-linkage*, *complete-linkage*, *average-linkage*, *mcquitty*, *median*, and *centroid*. The *k*-means, one of the most commonly adopted partitioning algorithms, has also been used. These clustering algorithms are well known; moreover, it is easy to obtain different partitions by modifying the input parameter that controls the number of clusters of an output partition. Each clustering algorithm will be used to compute a set of partitions with the number of clusters ranging from 2 to 10. From the perspective of dissimilarity measures, the comparison analysis will also be conducted. Five dissimilarity measures for each particular clustering algorithm will also be used: *Euclidea*n, *maximum*, *Manhattan*, *Canberr*a, and *Minkowski* distances.

To evaluate the performance of the 24 CVIs, 90 artificially generated data sets will be created. Most of the synthetic numerical data sets will be generated using the mixture models of the Gaussian distribution but with different parameters. Furthermore, 10 benchmark data sets (the true number of clusters is known a priori from the literature) drawn from the literature sources as well as from available *UCI* and *Kaggle* repositories will also be analyzed (see Table 2). The synthetic data sets were created to cover a large number of factor combinations such as: the number of clusters (K), the minimum (n_{min}) and maximum (n_{max}) number of objects in a data set, cluster density (den), and dimensionality (dim). The values of the parameters used to create the synthetic data sets are shown in Table 1.

Table 1. Values of the parameters used for generating the synthetic data sets

Parameter	Value
n_{min}	100
n_{max}	6000
K	2…10
dim	2…4
den	1…4

Since 90 synthetic data sets will be created, 4050 configurations have been obtained by multiplying this value by 5 partition dissimilarity measures and 9 clustering algorithms. In the case of benchmark data sets, the experiment is based on 450 configurations — 10 data sets, 9 clustering algorithms, and 5 partition dissimilarity measures. Considering the synthetic and benchmark data sets and taking into account the different number of partitions computed for each data set, each of the 24 CVIs should be computed for 40500 partitions.

Data sets

In synthetic numerical data sets, the clusters are non-overlapping represented as multivariate finite mixtures. The synthetic data sets were created without introducing overlapping, noise, or missing data objects. Imprecise and noisy data with overlapping clusters could be distorted as compared to human intuition [55]; therefore, noise and overlap level factors are excluded from this experimental setup.

Table 2. Characteristics of the benchmark data sets

Data set	Number of clusters
Steinley [65]	5
G2-set [29]	2
Unbalance 1 [58]	3
Unbalance 2 [58]	5
Square1 [36]	4
Triangle1 [36]	4
AD_5_2 [5]	5
AD_10_2 [5]	10
Haberman-survival (Kaggle)	4
Iris (UCI)	3

The “true” number of clusters in synthetic data sets ranges from 2 to 10 depending on the set, with cluster sizes from 50 to 3000 data objects per cluster. Furthermore, 90 synthetic data sets were generated with an uneven number of clusters per data set, namely: 4 data sets with 2 clusters, 8 data sets with 3 clusters, 22 data sets with 4 clusters, 20 data sets with 5 clusters, 10 data sets with 6 and 7 clusters, 6 data sets with 8 and 9 clusters, 4 data sets with 10 clusters. Afterward, about half of them (47%), which is 42 data sets, were generated with 4 or 5 clusters.

5. Common majority-based decision fusion method dubious scenarios

Passing through the testing phase repeatedly, there is not always a clear distinguishing line between all the majority situations. Taking into consideration the experimental results, the most common 4 cases of MBDFM controversial situations are presented. All scenarios were obtained by the default clustering validation configuration, i.e. *k*-means clustering and Euclidean distance.

Scenario 1: The decision is made by the relative CVI’s majority, and the nearest alternative is 50% votes less than half of the majority one. The relative majority points to the “true” number of clusters. The example of Scenario 1 and the accompanying data set presented to enhance reader understanding of CVIs voting is shown in Figure 1a.

Scenario 2a: Situation when the decision is taken by the CVI’s relative majority, and the nearest alternative is 50% votes more than half of the majority one. The relative majority points to the “true” number of clusters. The example of Scenario 2a is shown in Figure 1b. This scenario requires additional MBDFM verification for the final statement. In a validation configuration presented in Figure 1b, an almost equal number of CVIs voted for the 3 and 6 clusters. This scenario shows the controversial situation with no clear-cut majority. However, following the hard logic, the relative CVI’s majority points to the 6 clusters to be the “true” ones.

Scenario 2b: The scenario where the relative CVI’s majority points to the incorrect number of clusters. The nearest alternative group of CVIs in turn shows the „true” number of clusters. The example of the Scenario 2b case is shown in Figure 1c. The most critical of all the previous scenarios requires complete and precise MBDFM verification. Due to the limited facility of the classifier (Euclidean distance) and the crisp nature of the *k*-means algorithm, the results are completely misleading. This scenario shows the majority of CVIs voted for 3 clusters, and the nearest alternative voted for 4 numbers of clusters to be the “true” ones. The MBDFM is data-dependent since different CVIs behave differently on different data sets in various environments. The majority evaluation works on the fundamental assumption that the clustering algorithm works correctly. If this assumption does not hold, there could be a “fake” majority that identifies a false “true” number of clusters. Moreover, the lack of knowledge of the “true” number of clusters has a detrimental effect on clustering quality. Clustering in real-life applications is executed in a black-box fashion. The analyst is usually unable to correctly determine the “true” number of clusters beforehand. Therefore, Scenario 2 has been divided into two sub-cases.

Scenario 3: The data sets for which no majority prevails. In situations when no majority exists and two equal groups of CVIs voted for a different number of clusters to be the “true” ones, however, only one group of CVIs is correct and the other is misleading. The example of Scenario 3 is shown in Figure 1d. The presented scenario is a natural CVIs bias in favor of the status quo. However, according to the MBDFM, the “true” number of clusters among two equal groups of CVIs, is the one that is the first in the list. In this case, the function decided that the majority of CVIs voted for 2 as the best number of clusters despite the same number of CVIs cast the vote for 4 clusters.

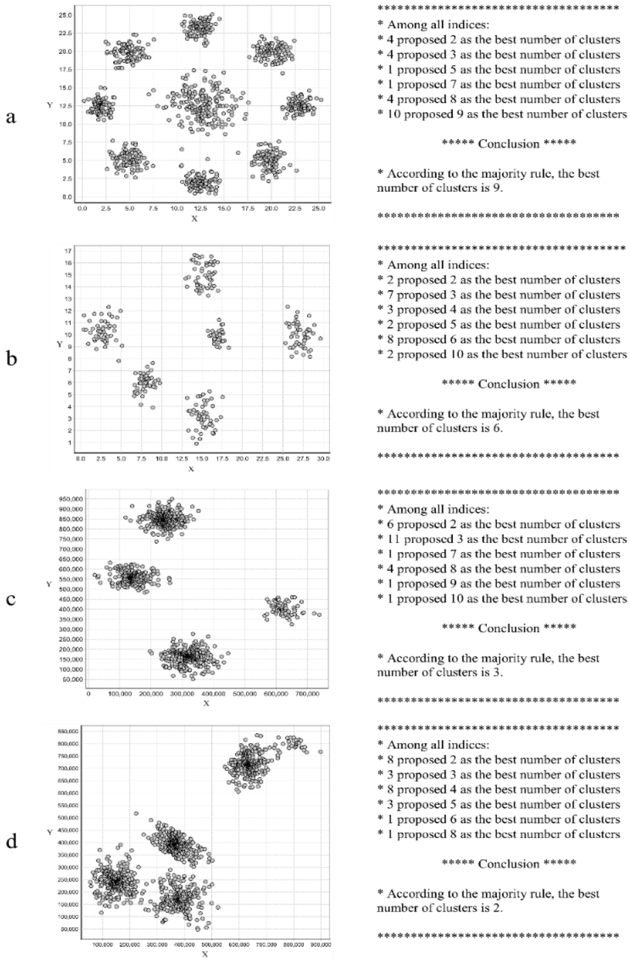


Fig. 1. Two-dimensional plot of synthetic data sets used in the experiment broken down by the default MBDFM configuration: (a) corresponds to the Scenario 1 situation with nine globular-form “true” clusters. (b) corresponds to the Scenario 2a situation with six globular-form “true” clusters. (c) corresponds to the Scenario 2b situation with four globular-form “true” clusters, however, according to the default MBDFM configuration, the revealed number of clusters is equal to three. (d) corresponds to the Scenario 3 situation with four globular and elongate form “true” clusters, however, according to the default MBDFM configuration the revealed number of clusters is equal to two

Summarizing all the above scenarios and CVI's results, doubts are expressed whether the MBDFM with default configuration is feasible in all clustering situations. Moreover, even a clear-cut data grouping structure (compact and far from other clusters) can easily deceive the 24 CVIs. Interestingly, there are less than 15% (15 data sets) of all the cases in the experimental setup, where the absolute majority indicates the "true" number of clusters. Thus, the cause must be sought in the NbClust majority rule, in the expediency, the classifier, and clustering algorithm, but do not question the quality and the correctness of each of the 24 CVIs. The MBDFM with the default validation configuration should maintain the basis for further clustering improvement. Thus, a non-invasive configuration of MBDFM is proposed to replace the default configuration, for all controversial scenarios.

6. Non-invasive validation configuration

6.1. Clustering method

Recent research focuses on clustering analysis to understand the strengths and weaknesses of various clustering algorithms in terms of data factors. As has been mentioned before, certain data characteristics may strongly affect clustering analysis, including high dimensionality, noise, types of attributes, and scales [66]. That being said, the authors have studied the clustering validation procedure by answering the question: *How to choose the best clustering algorithm appropriate for the MBDFM*? Considering that there are numerous clustering algorithms proposed in the literature, especially after an algorithm boom in the data mining area, it is arguable which clustering algorithm is the most suitable for the MBDFM.

Fig. 2 shows the percentage of correct guesses achieved by all 24 CVIs, which are sorted by the success score. Notice that this percentage refers to 194400 configurations: 100 synthetic and benchmark data sets, 9 clustering algorithms, 9 partitions, and 24 CVIs. All presented clustering algorithms are examined with the Euclidean dissimilarity measure as a distance metric. Correct guesses are considered as CVI votes that identify the "true" number of clusters. In brief, the Boolean value for each CVI is obtained; a correct CVI guess corresponds to 1 and an invalid one to 0. The sum of correct guesses forms the majority. Furthermore, only scenarios 1 and 2a are in favor of proper situations, and scenarios 2b and 3 are considered to be invalid. Indeed, Scenario 2a is initially classified along with Scenario 1 as a part of the group that forms the overall success score. Moreover, after the comparative analysis, all Scenario 1 data sets identify the true number of clusters correctly, and therefore, the MBDFM does not require any additional adjustment.

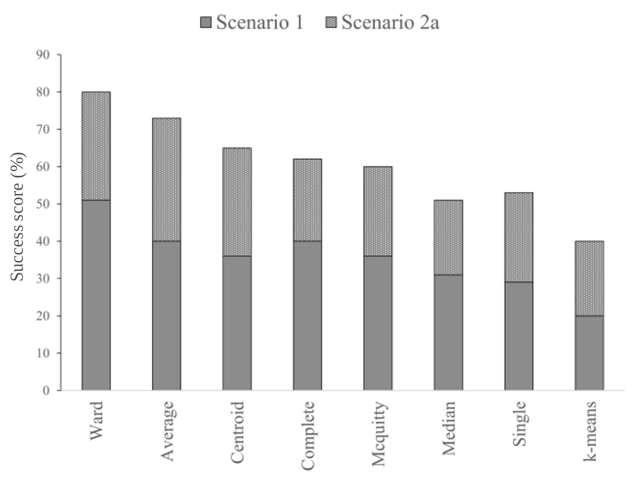


Fig. 2. Overall success score for data sets broken down by a clustering algorithm

Although we can find a clear pattern, it seems that the overall comparative results are severely affected by the used clustering algorithm. Assuming that all the potential candidates from Scenario 2a become a fully-fledged majority that point to the "true" number of clusters, most CVIs obtain their worst results for the *k*-means algorithm, i.e. 40% (Scenario 1 – 20% + Scenario 2a –20%), while Ward shows the highest success score of 80% (Scenario 1 – 51% + Scenario 2a – 29%). If we focus on *k*-means and Ward only, this factor shows drastically different results and an accuracy difference of 40%. On the other hand, the result for the centroid algorithm of 65% (Scenario 1 – 36% + Scenario 2a – 29%), the complete-linkage algorithm of 62% (Scenario 1 – 40% + Scenario 2a – 22%), and the mcquitty algorithm of 60% (Scenario 1 – 36% + Scenario 2a – 24%) reduce the differences between the CVIs decisions and balance the overall success score.

Table 3. Overall success-failure score (%) of the majority-based decision fusion method for all data sets with the Euclidean dissimilarity measure broken down by clustering algorithms

	Scenario 1	Scenario 2a	Scenario 2b	Scenario 3
Ward	51%	29%	9%	11%
Single	29%	24%	47%	0%
Complete	40%	22%	31%	7%
Average	40%	33%	22%	5%
Mcquitty	36%	24%	36%	4%
Median	31%	20%	40%	9%
Centroid	36%	29%	26%	9%
k-means	20%	20%	49%	11%

The situation becomes more interesting and clearer after the analysis of all scenarios (see Table 3). According to the previous definition of success, failure is defined as an incorrect decision of the CVI. Thus, the failure score is a total sum of CVI scores that form Scenario 3 situations, or wrong and "fake" majority (Scenario 2b) that differs from the "true" number of clusters. With respect to the failure score, Scenario 3 with the Ward clustering algorithm has reached (11%); the next are the median and centroid clustering algorithms with (9%) each and the last is the single-linkage clustering algorithm (0%). The results of Scenario 2b are the most valuable and controversial at the same time. As can be seen, *k*-means shows the highest failure score (49%) and Ward – the lowest (9%) one. Summing up, according to the presented results, the Ward clustering algorithm is the only obvious rational choice for further validation research.

6.2. Dissimilarity measure

Using an explicit dissimilarity measure to guide the validation process is a very popular approach, adopted by many widely-used clustering algorithms. Unfortunately, there are no definitive rules on which measure to choose for a particular problem. Dissimilarity measures should be considered in the context of the study where they are to be used, including the nature of data and the type of analysis. However, certain general guidelines do exist, i.e. the nature of data should strongly influence the choice of the dissimilarity measure; the choice of dissimilarity measure should depend on the scale of the attributes; the clustering algorithm should influence the choice of the dissimilarity measure. It could be considered a fatal defect in the validation procedure if too many dissimilarity measures have to be taken into consideration; however, it might be felt that a wide variety of possible measure choices is an advantage making the validation procedure usefully flexible.

It is hard to choose the most appropriate dissimilarity measure for a given clustering task without a preliminary experiment. Various dissimilarity measures presented in this article can be considered for use with all the presented clustering algorithms that are flexible enough not to be tied to a particular measure. It makes it possible to choose carefully based on the available domain knowledge and to verify the effects of several candidate measures experimentally.

Instead, a comparative study of 5 dissimilarity measures has to be conducted in the clustering verification process. The analysis will be focused on an appropriate choice of a dissimilarity measure in Ward's algorithm since the rest of the clustering algorithms have previously been rejected.

As in the case of selecting the clustering algorithm, scenarios 1 and 2a are the proper situations and scenarios 2b and 3 are considered to be incorrect. Figure 3 shows that the selected partition dissimilarity measure moderately affects the behavior of the CVIs, not as extremely as in the case of the clustering algorithm. Two of the presented dissimilarity measures, i.e. Minkowski and Euclidean, follow the overall pattern with 80% of correct guesses. The maximum dissimilarity measure shows slightly better results with an 81% success score. Furthermore, the Canberra dissimilarity measure yields extremely good results – 85% of correct guesses.

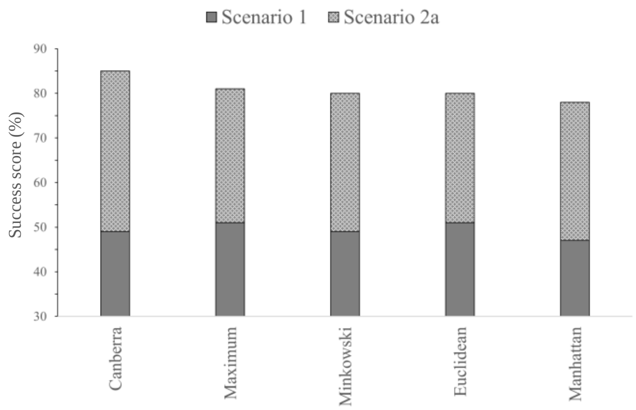


Fig. 3. The overall score for data sets broken down by a dissimilarity measure

In terms of the failure score, the results show that the difficulty imposed by the bias situations (Scenario 3) could be seen to a relatively small extent (9% – 11%) in all dissimilarity measures (see Table 4). The Canberra dissimilarity measure notifies the 5% of the detected Scenario 3 cases. Considering the contribution of Scenario 2b situations to the overall results, the Euclidean, maximum, and Minkowski dissimilarity measures should be mentioned as ones with the lowest (9%) failure scores. Mostly, false decisions were made using the Manhattan dissimilarity measure (13%). The Canberra dissimilarity measure shows slightly better results (10%) than Manhattan one.

Table 4. Overall success-failure score (%) of the majority-based decision fusion method for all data sets with the Ward clustering algorithm broken down by a dissimilarity measure

	Scenario 1	Scenario 2a	Scenario 2b	Scenario 3
Canberra	49%	36%	10%	5%
Maximum	51%	30%	9%	10%
Minkowski	49%	31%	9%	11%
Euclidean	51%	29%	9%	11%
Manhattan	47%	31%	13%	9%

In conclusion, the experiments show sufficiently moderate evidence for choosing a dissimilarity measure that is significantly better than the rest. However, the Ward clustering algorithm with the Canberra dissimilarity measure is recommended as the best for non-invasive MBDFM configuration settings.

7. Invasive validation configuration

Despite all attempts to improve MBDFM using the non-invasive configuration by way of altering the dissimilarity measure and the clustering algorithm, the best overall success score remains at the 85% with the Ward clustering algorithm in conjunction with the Canberra dissimilarity measure. Moreover, attempts to improve the result, the validation of the CVIs number were also carried out. Unfortunately, adding 2 more CVIs to 26 or subtracting 2 CVIs to obtain 22 did not change the overall success score. Of course, if the number is significantly changed from 24 to 10 CVIs, the result will also change. However, then another problem appears, namely the expediency and application correctness of each CVI individually. The smaller the number of CVIs the greater responsibility and user trust lies with each of them. In this context, the first step is to justify and select the best group of CVIs, and only then hold the voting procedure. The disadvantage of this approach is the very fact of CVIs division into best and worst. Furthermore, changing the input data can dramatically turn the situation and the best ones may become worst and vice versa. Taking all of the above into consideration, the main way to improve the results is to modify the MBDFM using the invasive configuration.

The non-invasive MBDFM relies on selecting the best clustering algorithm and dissimilarity measure to ensure optimal validation results. These two input MBDFM parameters could be tuned by the researcher. Moreover, the researcher could tune the third parameter – cluster range. The "true" number of clusters k_{true} are located across the range $K = \{2, \dots, 10\}$. However, the cluster range should be changed with great caution, as ill-considered change can lead to critical consequences. Such a change could eliminate the "true" number of clusters from the validation procedure and further the researcher, without knowing it, will look for the "true" number of clusters in a range of knowingly fake ones. All the following partitions will reveal the fake number of clusters, moreover, all of the CVIs will be forced to vote for the fake number of "true" clusters which will lead to erroneous MBDFM decisions. In real-world validation issues, the "true" number of clusters is unknown a priori and the researcher without knowing this fact forcibly restricts the CVI's possible decisions in the frames of cluster range. Whether the CVI is good or its decision is far from optimal, it should cast the vote only for the particular number of clusters in the prescribed cluster range. The role of the initial cluster range is extremely high as a broad cluster range gives more freedom to each of CVIs in their voting, however, the validation procedure is becoming fuzzy. On the other hand, the excessive compression of cluster range gives less freedom of CVI's votes cast for a particular cluster, but the validation procedure is becoming crisp.

The authors found a clear pattern in cluster range modification which allows to safely reduce it. The cluster range remains the same at the beginning of the MBDFM procedure. Figure 1d describes the operation of MBDFM with decision groups of CVIs that cast votes for the particular cluster. The authors noticed, that none of the CVIs cast a vote for 9 and 10 clusters. None of the 24 CVIs presented in the experimental setup even with possessed decision capabilities doesn't vote for these clusters. Thereby, the upper bound of the cluster range could be safely reduced – $K = \{2, \dots, 8\}$. Moreover, if none of the CVIs cast votes for clusters in the lower bound, it allows reducing the cluster range on the other side. This procedure will efficiently distribute the CVI's decisions across the optimized cluster range. The authors hypothesized, if none of the CVIs cast votes for clusters in the upper or lower bound of the cluster range, it is reliably confirmed that these clusters cannot become the candidates to be the "true" ones. However, it is not always possible to optimize the cluster range. If one of the CVIs cast one single vote for the particular cluster in the lower or upper bound of the cluster range, it cannot be taken lightly (see Figure 1b, 1c). That is, at this stage, all CVI's votes are taken into account and the cluster range is reduced by cutting its lower and upper bounds i.e. the clusters for which none of the CVIs cast a vote.

It should also be noted that only the lower and upper bound of the cluster range should be reduced without dividing the cluster range into two or more subranges. Figure 1c shows that none of the CVIs cast votes for the 4, 5, and 6 clusters. However, these clusters are located within the initial cluster range, and subtracting them leads to the cluster range division into $K_1 = \{2 \dots 3\}$ and $K_2 = \{7, \dots, 10\}$ subranges. Analyzing the cluster subranges separately gives the researcher two or more "true" clusters where only one is "true" and the others – fake "true". Moreover, Figure 1c will show the erroneous CVI results, since the researcher will subtract the chance of CVIs to vote for the four clusters to be the "true" one. Therefore, in this stage, the initial cluster range should remain integral, even if none of the CVIs cast a single vote to the cluster located within the cluster range. The cluster range optimization procedure applies to the clusters without any votes cast by CVIs only in the lower and upper bound of the initial cluster range.

The cluster range optimization procedure is not always possible to conduct. Figure 1a shows that CVI's votes are cast for clusters that completely cover the initial cluster range $K = \{2, \ldots, 10\}$. This, in turn, does not allow to carry out optimization procedure in the way described above, however, gives the chance to consider another cluster range optimization strategy. This strategy is based on the CVI's votes cast for the relative majority $Maj_{est}^{\hat{}}$ and its nearest alternative Maj_{est}^{*}. The relative majority $Maj_{est}^{\hat{}}$ estimated by the MBDFM with the non-invasive configuration, could point to the "true" number of clusters (Scenario 1, 2a) but also may indicate the fake number (Scenario 2b). The nearest alternative Maj_{est}^{*} is the second-largest decision CVI's group which is closest to the "true" number of clusters. If the first strategy optimizes the cluster range by means of cutting its lower and upper bound and subtracting the cluster with no CVI's vote, the second one is entirely based on the CVI's majority and its nearest alternative. The authors' hypothesizes, that $Maj_{est}^{\hat{}}$ and Maj_{est}^{*} with its corresponding $k_{est_j^{\hat{}}}$ and $k_{est_j^{*}}$ should become the upper/lower bound of K. In Figure 1b in the second optimization step, the cluster range will become $K = \{3, \ldots, 6\}$, where the relative majority $Maj_{est}^{\hat{}}$ will become its upper bound with corresponding $k_{est_5^{\hat{}}} = 6$ and the nearest alternative Maj_{est}^{*} will become its lower bound with corresponding $k_{est_2^{*}} = 3$ clusters. That is, at this stage, all CVI's votes that were excluded from the validation procedure by means of cluster range reduction will forcibly cast the votes only for clusters within the new optimized cluster range. The authors assume that all CVI's votes excluded from the initial cluster range will strengthen the final decision – the majority that points to the "true" number of clusters in the optimized cluster range. Furthermore, even if MBDFM with a non-invasive configuration will return the final decision of fake "true" number of clusters (Scenario 2b), the optimization procedure will help the CVI's votes to steer their decisions in the direction of the "true" number of clusters. This optimization strategy assumes that the "true" number of clusters should be located within the new cluster range (with Maj_{est}^{*} and $Maj_{est}^{\hat{}}$ the upper/lower bound of K), and all CVI's votes that were excluded from the optimized cluster range will be forcibly asked to cast their votes only for clusters within the new range K.

Each CVI formulates a vote that favors one of the 9 clusters at stake. The dynamic process that characterizes each CVI vote formation is based on the idea that the CVI's majority reveals the particular clusters probabilistically. The formation process of the CVI's votes strongly depends on the cluster range. It is assumed that the optimization procedures described above could change the CVI's votes in favor of the "true" number of clusters, even if before the optimization procedures some of CVIs could cast their votes for the fake number of "true" clusters.

These optimization procedures are carefully collected and written in the form of the majority MBDFM rule (*i*), which aims at revealing the biggest number of "true" clusters in the experimental setup. Using Scenario 1 2a and 2b cases from non-invasive MBDFM configuration as our input data, the enhanced majority rule can be written as follows.

Rule (*i*):

1. Run the MBDFM with a non-invasive configuration. Let DEC= $\{Dec_1, Dec_2, \ldots, Dec_k\}$ and Scenario 1, 2a, or 2b is considered.
2. Reveal the number of clusters $k_{est_j^{\hat{}}}$ and $k_{est_j^{*}}$ that corresponds to the relative majority $Maj_{est}^{\hat{}}$ and its nearest alternative Maj_{est}^{*} respectively.
3. Optimize the cluster range by means of reducing the upper and lower bound of K when some of the decision groups Dec_i do not reveal any number of clusters. If no majority prevails, use rule (*ii*).
4. Repeat step (2). If MBDFM reveals the absolute majority $Maj_{est}^{\hat{}}$, assume it corresponds to the "true" number of clusters, then $k_{est_j^{\hat{}}} = k_{true}$, otherwise, if the absolute majority was not achieved move to the next step.
5. Change the cluster range, where $Maj_{est}^{\hat{}}$ and Maj_{est}^{*} and its corresponding $k_{est_j^{\hat{}}}$ and $k_{est_j^{*}}$ becomes the upper/lower bound of K.
6. Rep`eat step (2). If the MBDFM reveals $k_{est_j^{\hat{}}} = k_{true}$, the estimated number of clusters is the "true" one; otherwise, if the $k_{est_j^{\hat{}}} \neq k_{true}$, the majority $Maj_{est}^{\hat{}}$ identifies the fake "true" number of clusters. Moreover, if no majority prevails, rule (*i*) did not give the expected results and should not be used in the MBDFM invasive configuration.

The overall success score is fully justified and confirmed, since rule (*i*) of MBDFM with an invasive configuration applied to the Scenario 1, 2a, and 2b cases (49% for scenario 1, 36% for scenario 2a, 10% for scenario 2b) approves the 95% of correct guesses. All Scenario 2b controversial situations (10%) have been solved in favor of the "true" number of clusters. The rule (*i*) shows sufficiently strong evidence to adopt it into the MBDFM's default clustering validation decision scheme to enhance the NbClust majority voting procedure.

Nevertheless, there is a group (Scenario 3) of about 5% of all data sets that seems to show the questionable bias situations even under the MBDFM invasive validation configuration applied with the rule (i). It should be emphasized that Scenario 3 mirrors the situation where no majority prevails when the two biggest decision groups $Dec_{est}^{1} = Dec_{est}^{2}$ have equal votes cast by CVIs. In this case, only one decision group among them specifies the "true" number of clusters. Scenario 3 situations require not only the confirmation of the relative majority correctness (as in Scenario 2a data sets) but thorough analysis and modification of the decision-making scheme in general. Due to the work limitations which cannot embrace every data case, the contentious situations, therefore, could appear for other data sets not examined in the experimental setup.

The MBDFM with an invasive configuration and rule (*i*) should be applied to Scenarios 1,2a and 2b, where the majority of CVI's votes prevails. However, rule (*i*) could not be used in the case of Scenario 3. The strategy of optimization the cluster range, by means of excluding the CVIs that cast no votes in the upper or lower bound of the initial cluster range, remains the same. However, the strategy of optimizing the cluster range by means of revealing the majority and its nearest alternative will be modified. Rule (*ii*) is created to cope with scenario 3 situations. The authors hypothesize, that Dec_{est}^{1} and Dec_{est}^{2} with their corresponding $k_{est_j^{\hat{}1}}$ and $k_{est_j^{\hat{}2}}$ should become the upper/lower bound of K. In Figure 1d this optimization procedure will produce the new cluster range $K = \{2, \ldots, 4\}$, where the Dec_{est}^{1} will become its lower bound with corresponding $k_{est_1^{\hat{}1}} = 2$ and the Dec_{est}^{2} will become its upper bound with corresponding $k_{est_3^{\hat{}2}} = 4$ clusters. All CVI's votes excluded from the initial cluster range will be forcibly asked to cast their votes for clusters in the optimized range. The optimization procedure in Scenario 3 cases mainly aimed at revealing the majority between two equal groups of CVIs. Therefore, rule (*ii*) aims to effectively imbalance the votes divided between the two biggest groups of CVIs and tips the scales in favor of the decision group that points to the "true" number of clusters. Finally, the optimized cluster range applied in a rule (*ii*) should strengthen the final CVI's decision – the majority with the corresponding "true" number of clusters.

Rule (*ii*) along with rule (*i*) make up an integral part of the validation procedure and, therefore, both of them become appropriate tools for the invasive MBDFM configuration scheme. Furthermore, for the best revealing of the "true" number of clusters, the MBDFM with invasive configuration should be performed based on the validation results produced by the non-invasive configuration. That is, the MBDFM with non-invasive configuration should always precede the MBDFM with invasive configuration. Finally, the invasive configuration aims to strengthen the non-invasive to efficiently cope with all data sets presented in the experimental majority-based validation procedure.

Using Scenario 3 cases from the non-invasive MBDFM configuration as our input data, the enhanced majority rule can be written as follows.

Rule (*ii*):

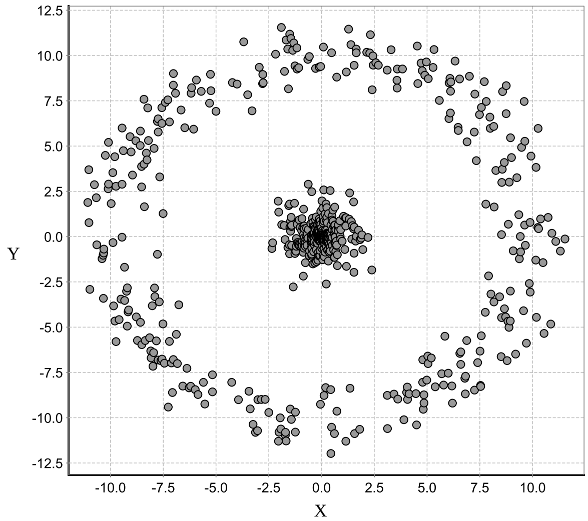
1. Run the MBDFM with a non-invasive configuration. Let DEC= $\{Dec_1, Dec_2, \ldots, Dec_k\}$ and Scenario 3 is considered.
2. Reveal the number of clusters $k_{est_j^{\hat{}1}}$ and $k_{est_j^{\hat{}2}}$ that corresponds to the two equal groups of CVIs, Dec_{est}^1 and Dec_{est}^2 respectively.
3. Optimize the cluster range by means of reducing the upper and lower bound of K when some of the decision groups Dec_i do not reveal any number of clusters.
4. Repeat step (2). If MBDFM reveals an absolute majority, assume it corresponds to the "true" number of clusters, then $k_{est_j^{\hat{}}} = k_{true}$, otherwise, if no majority prevails, move to the next step. If the MBDFM reveals the relative majority $Maj_{est}^{\hat{}}$ use step (5) of rule (*i*).
5. Change the cluster range, where Dec_{est}^1 and Dec_{est}^2 and its corresponding $k_{est_j^{\hat{}1}}$ and $k_{est_j^{\hat{}2}}$ becomes the upper/lower bound of K.
6. Repeat step (2). If the MBDFM reveals $k_{est_j^{\hat{}}} = k_{true}$, the estimated number of clusters is the "true" one; otherwise, if the $k_{est_j^{\hat{}}} \neq k_{true}$, the majority $Maj_{est}^{\hat{}}$ identifies the fake "true" number of clusters. Moreover, if no majority prevails, rule (*ii*) did not give the expected results and should not be used in this MBDFM invasive configuration.

With regard to Scenario 3 situations, rule (*ii*) had a strong impact on the CVI majority voting procedure. The number of successes is considerably increased when rule (*ii*) is adopted to the MBDFM with an invasive configuration. In particular, the overall success score of 95% without rule (*ii*) is exceedingly improved to 99%.

Another remarkable and surprising fact is that 1% of experimental data sets (1 data set) show wrong results even with the application of MBDFM with the invasive configuration. This is due to a more complex data group structure presented in the data set.

Figure 4 shows the informally called the "crater" data set that misled rule (*i*) and the MBDFM with the invasive configuration in general. This synthetic "toy" data set consists of 2 clusters, one of them being a globular form dense cluster and the other being a ring cluster that surrounds the first one. This case corresponds to the Scenario 2c situations. This scenario was not included in the main list of the most frequent dubious scenarios. This case cannot be included in Scenario 2a, since the relative majority $Dec_3 = Maj_{est}^{\hat{}} \neq Maj_{true}$ and the number of corresponding clusters $k_{true} \neq k_{est_3^{\hat{}}} = 6$ did not point to the "true" number of clusters. Moreover, this case cannot be included in Scenario 2b, since the nearest alternative group of CVIs and the number of corresponding clusters $k_{est_2^*} = 3$ did not show the „true" one $Dec_2 = Maj_{est}^* \neq Maj_{true}$. This case shows that neither the relative majority $Maj_{est}^{\hat{}}$ nor the neighboring alternative Maj_{est}^* indicated the correct "true" number of clusters $k_{true} = 2$. In this particular data set, the MBDFM with an invasive configuration did not provide the expected results, moreover, the method reveals the misleading fake number of "true" clusters.

The behavior of MBDFM becomes unpredictable for a number of reasons which are not directly related to the correctness of the proposed approach. For such cases, it is necessary to separately select a clustering algorithm and a dissimilarity measure that would be well adapted to such data. For such data sets [20] suggested using the special clustering algorithms. These types of clustering algorithms are robust to noise and the "touching problems" [72] including the "neck problem" [70] and the ''adjacent problem'' [28]. Moreover, some of CVIs may not be able to cope with such data a priori.



****
* Among all indices:
* 5 proposed 2 as the best number of clusters
* 6 proposed 3 as the best number of clusters
* 8 proposed 6 as the best number of clusters
* 2 proposed 7 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters
***** Conclusion *****
* According to the majority rule, the best number of clusters is 6

****

Fig. 4. Two-dimensional plot of synthetic "toy" data set used in the experiment broken down by the non-invasive MBDFM. The data set shows two "true" clusters of different densities, however, according to the non-invasive MBDFM configuration the revealed number of clusters is equal to six

8. The performance of validation approaches

In this section, the authors will show the difference between the MBDFM voting approaches with default, non-invasive and invasive configurations. Moreover, the authors will compare the MBDFMs to the individual CVI (Silhouette) with *k*-means and Ward clustering with Euclidean and Canberra dissimilarity measures. The authors [21] claim that the Silhouette is the best individual CVI that achieves the best overall validation results for synthetic and real data sets broken down by the number of clusters, dimensionality, cluster overlap, and density experimental factors. Therefore, with full confidence in accordance with results presented by Arbelaitz et al. [21], the Silhouette has been chosen as the best individual CVI for clustering validation comparison reasons. Table 5 lists the overall success-failure score (%) of the MBDFMs compared to the individual CVI.

Table 5. Overall success-failure score (%) of the MBDFM with non-invasive, invasive, and default configurations compared with the result revealed by the individual Silhouette CVI

	Success score (%)		Failure score (%)		
Scenario:	1	2a	2b	2c	3
1. Silhouette with *k*-means clustering and Euclidean distance	26%	16%	41 %	10 %	7 %
2. Silhouette with *k*-means clustering and Canberra distance	21%	20%	39 %	9 %	11 %
3. Silhouette with Ward clustering and Euclidean distance	28%	22%	36 %	7 %	7 %
4. Silhouette with Ward clustering and Canberra distance	23%	27%	43 %	3 %	4 %
5. MBDFM with default configuration	20%	20%	49 %	0 %	11 %
6. MBDFM with non-invasive configuration	49%	36%	10 %	0 %	5 %
7. MBDFM with invasive configuration	99%	0%	0 %	1 %	0 %

As it can be observed, the MBDFM with default configuration cannot beat the Silhouette index for all data sets in a different configuration. The best success score has been achieved using the Silhouette CVI with the Ward clustering algorithm and Canberra or Euclidean distance 50%. The Silhouette CVI with the *k*-means clustering and Canberra distance has achieved the smallest success score – 41% (Scenario 1 – 21% + Scenario 2a – 20%). Moreover, the MBDFM with default configuration achieved an equal success score in comparison with the Silhouette CVI with *k*-means clustering and Euclidean distance – 40%. In conclusion, the Silhouette CVI with different configurations achieved a higher or equal individual success score than MBDFM with the default configuration.

Considering the MBDFM with non-invasive and invasive configurations, both of these decision fusion methods beat the overall success score of the Silhouette CVI for all data sets presented in an experimental setup. In particular, the improvement over this CVI and the best configuration was 35% for the MBDFM with the non-invasive configuration and 49% for the MBDFM with the invasive configuration and adapted rules. The analysis showed that the design of decision fusion strategies requires careful choice of the validation configurations. Finally, the MBDFM with default configuration showed no improvement in performance, whereas both voting methods MBDFM with non-invasive and invasive configurations showed to perform better than single Silhouette CVI.

9. Conclusions

The experimental results demonstrated the appealing performance of MBDFMs in searching and justifying the “true” number of clusters and thus confirmed the potential approach of integrating MBDFMs into the clustering framework. The MBDFMs and overall clustering validation schema could be iterative and researchers seek a “true” number of clusters each time. Depending on the task’s requirements or/and the level of acceptance with the validation results, the MBDFM with default configuration can be quite satisfactory. However, for detailed and sophisticated data analysis, the authors propose a more refined MBDFM with invasive configuration, where the information of all previously uncovered “true” clusters by means of MBDFM with non-invasive configuration will be used as background knowledge to derive a precise final decision. Moreover, if researchers wish to determine the pros and cons of other existing or novel CVIs, clustering algorithms, or data sets in the future, this benchmarking framework can be applied to make a thorough comparison.

In light of the results achieved, the authors consider that MBDFMs are a successful path to obtain the best partition for each context, which is the key issue in the data clustering field. Thus, the authors believe that new contributions on MBDFMs clustering validation can help to reduce the uncertainty about the suitability of the partitions generated by the algorithms. This work also raises some questions and, therefore, suggests some future work. The authors consider that, even though they performed an extensive comparison, there is still room for extending it to include more CVIs, data sets, clustering algorithms, dissimilarity measures, cluster range, high dimensionality, etc. In this context noise and overlap would appear to be the most interesting factors to analyze in greater depth. Moreover, the work is limited to binary crisp CVI’s decisions, so a fuzzy CVI’s comparison would be a natural continuation.

References

[1] Akoglu L., Tong H., Koutra D.: Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery 29(3), 2015, 626–688.

[2] Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J., Perona I.: An extensive comparative study of cluster validity indices. Pattern Recognition 46(1), 2013, 243–256.

[3] Bailey K.D.: Typologies and Taxonomies: An introduction to classification techniques (quantitative applications in the social sciences). SAGE Publications, Thousand Oaks 1994.

[4] Ball G.H., Hall D.J.: ISODATA, a Novel Method of Data Analysis and Pattern Classification. Stanford Research Institute 1965.

[5] Bandyopadhyay S., Maulik U: Nonparametric genetic clustering: comparison of validity indices. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 31(1), 2001, 120–125.

[6] Beale E.M.L.: Cluster Analysis. Scientific Control Systems, London 1969.

[7] Bezdek J., Li W., Attikiouzel Y., Windham M.: A geometric approach to cluster validity for normal mixtures. Soft Computing – A Fusion of Foundations, Methodologies and Applications 1(4), 1997, 166 –179.

[8] Bezdek J., Pal N.: Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics) 28(3), 1998, 301–315.

[9] Berkhin P.: A Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data. Springer, Berlin 2006.

[10] Braune C., Besecke S., Kruse R.: Density Based Clustering: Alternatives to DBSCAN, Partitional Clustering Algorithms. Springer, Cham 2014.

[11] Brock G., Pihur V., Datta S., Datta S.: clValid: An R Package for Cluster Validation. Journal of Statistical Software 25(4), 2008, 1–22.

[12] Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.: Model-based evaluation of clustering validation measures. Pattern Recognition 40(3), 2007, 807–824.

[13] Calinski T., Harabasz J.: A dendrite method for cluster analysis. Communications in Statistics – Theory and Methods 3(1), 1974, 1–27.

[14] Cannataro M., Congiusta A., Mastroianni C., Pugliese A., Talia D., Trunfio P.: Grid-Based Data Mining and Knowledge Discovery. Intelligent Technologies for Information Analysis. Springer, Berlin 2004.

[15] Celebi M.: Partitional clustering algorithms. Springer, Cham 2015.

[16] Charrad M., Ghazzali N., Boiteau V., Niknafs A.: NbClust: AnRPackage for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software 61(6), 2014, 1–36.

[17] Cho K., Lee J.: Grid-Based and Outlier Detection-Based Data Clustering and Classification. Communications in Computer and Information Science. Springer, Berlin 2011.

[18] Chou C., Su M., Lai E.: A new cluster validity measure and its application to image compression. Pattern Analysis and Applications 7(2), 2004, 205–220.

[19] Davies D., Bouldin D.: A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1(2), 1979, 224–227.

[20] Deng M., Liu Q., Cheng T., Shi Y.: An Adaptive Spatial Clustering Algorithm Based On Delaunay Triangulation. Computers, Environment and Urban Systems 35, 2011, 320–332.

[21] Dimitriadou E.: cclust: Convex Clustering Methods and Clustering Indexes. R package version 0.6-18, 2014.

[22] Dimitriadou E., Dolňicar S., Weingessel A.: An examination of indexes for determining the number of clusters in binary data sets. Psychometrika 67(1), 2002, 137–159.

[23] Dubes R.: How many clusters are best? – An experiment. Pattern Recognition 20(6), 1987, 645–663.

[24] Duda R., Hart P: Pattern classification and scene analysis. Wiley, New York 1973.

[25] Duda R, Hart P., Stork D.: Pattern classification. Wiley, New York 2001.

[26] Dunn J.: Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics 4(1), 1974, 95–104.

[27] Embrechts E., Gatti C., Linton J., Roysam B.: Hierarchical Clustering for Large Data Sets. Advances in Intelligent Signal Processing and Data Mining. Springer, Berlin 2013.

[28] Estivill-Castro V., Lee I.: Argument Free Clustering For Large Spatial Point-Data Sets Via Boundary Extraction From Delaunay Diagram. Computers, Environment and Urban Systems 26, 2002, 315–334.

[29] Fränti P., Mariescu-Istodor R., Zhong C.: XNN Graph, Lecture Notes in Computer Science, 10029, 2016, 207–217.

[30] Frey T., van Groenewoud H.: A Cluster Analysis of the D 2 Matrix of White Spruce Stands in Saskatchewan Based on the Maximum-Minimum Principle. The Journal of Ecology 60(3), 1972, 873–886.

[31] Friedman H., Rubin J.: On Some Invariant Criteria for Grouping Data. Journal of the American Statistical Association 62(320), 1967, 1159–1178.

[32] Granichin O., Volkovich Z., Toledano-Kitai D.: Cluster Validation. Intelligent Systems Reference Library. Springer, Berlin 2015.

[33] Gurrutxaga I., Muguerza J., Arbelaitz O., Pérez J., Martín J.: Towards a standard methodology to evaluate internal cluster validity indices. Pattern Recognition Letters 32(3), 2011, 505–515.

[34] Halim Z., J. Khattak J.: Density-based clustering of big probabilistic graphs. Evolving Systems 10, 2019, 333–350.

[35] Halkidi M., Batistakis Y., Vazirgiannis M.: On Clustering Validation Techniques. Journal of Intelligent Information Systems 17(2/3), 2001, 107–145.

[36] Handl J., Knowles J.: Multi-Objective Clustering and Cluster Validation. Studies in Computational Intelligence. Springer, Berlin 2006.

[37] Halkidi M., Vazirgiannis M.: A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters, 29(6), 2008, 773–786.

[38] Halkidi M., Vazirgiannis M.: Clustering validity assessment: finding the optimal partitioning of a data set. Proceedings 2001 IEEE International Conference on Data Mining. IEEE, San Jose 2001.

[39] Halkidi M., Vazirgiannis M., Batistakis Y.: Quality Scheme Assessment in the Clustering Process. Lecture Notes in Computer Science. Springer, Berlin 2000.

[40] Hartigan J.A.: Clustering Algorithms. John Wiley & Sons, New York 1975.
[41] Hennig C.: Methods for merging Gaussian mixture components. Advances in Data Analysis and Classification 4, 2010, 3–34.
[42] Hornik K.: A CLUE for CLUster Ensembles. Journal of Statistical Software 14(12), 2005, 1–25.
[43] Hubert L., Levin J.: A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin 83(6), 1976, 1072–1080.
[44] Kryszczuk K., Hurley P.: Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. Lecture Notes in Computer Science, Springer, Berlin 2010.
[45] Krzanowski W., Lai Y.: A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. Biometrics 44(1), 1988, 23–34.
[46] Lu J., Zhang G., Ruan D., Wu F.: Multi-objective group decision making: methods, software and applications with fuzzy set techniques. Imperial College Press, London 2007.
[47] Maalel W., Zhou K., Martin A., Elouedi Z.: Belief Hierarchical Clustering, Belief Functions: Theory and Applications. Lecture Notes in Computer Science. Springer, Cham 2014.
[48] Marriott F.: Practical Problems in a Method of Cluster Analysis. Biometrics 27(3), 1971, 501–514.
[49] McClain J., Rao V.: CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects. Journal of Marketing Research 12(4), 1975, 456–460.
[50] Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F.: E1071: Misc Functions of the Department of Statistics, Probability Theory Group. R package version 1.6-8, 2017.
[51] Milligan G.: An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45(3), 1980, 325–342.
[52] Milligan G., Cooper M.: An examination of procedures for determining the number of clusters in a data set. Psychometrika 50(2), 1985, 159–179.
[53] Nerurkar P., Pavate A., Shah M., Jacob S.: Performance of Internal Cluster Validations Measures for Evolutionary Clustering. Advances in Intelligent Systems and Computing. Springer, Singapore 2018.
[54] Nieweglowski L.: clv: Cluster Validation Techniques. R package version 0.3-2.1, 2014.
[55] Oliveira J., Pedrycz W.: Advances in fuzzy clustering and its applications. John Wiley & Sons Ltd, Chichester 2007.
[56] Peng Q., Wang Y., Ou G., Tian Y., Huang L., Pang W.: Partitioning Clustering Based on Support Vector Ranking. Lecture Notes in Computer Science. Springer, Cham 2016.
[57] Ratkowsky D.A., Lance G.N.: A Criterion for Determining the Number of Groups in a Classification. Australian Computer Journal 10(3), 1978, 115–117.
[58] Rezaei M., Fränti P.: Set Matching Measures for External Cluster Validity. IEEE Transactions on Knowledge and Data Engineering 28(8), 2016, 2173–2186.
[59] Rousseeuw P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 1987, 53–65.
[60] Roux M.: A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. Journal of Classification 35(2), 2018, 345–366.
[61] Sarle W.S.: Cubic Clustering Criterion, SAS Technical Report A-108. SAS Institute Inc, Cary 1983.
[62] Saemi B., Hosseinabadi A., Kardgar M., Balas V., Ebadi H.: Nature Inspired Partitioning Clustering Algorithms: A Review and Analysis. Advances in Intelligent Systems and Computing. Springer, Cham 2017.
[63] Scott A., Symons M.: Clustering Methods Based on Likelihood Ratio Criteria. Biometrics 27(2), 1971, 387–397.
[64] Shim Y., Chung J., Choi I.: A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm. International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). IEEE, Vienna 2005.
[65] Steinley D., Henson R.: OCLUS: An Analytic Method for Generating Clusters with Known Overlap. Journal of Classification 22(2), 2005, 221–250.
[66] Tan P., Steinbach M., Kumar V.: Introduction to data mining. Pearson, 2005.
[67] Vathy-Fogarassy A., Abonyi J.: Graph-Based Clustering and Data Visualization Algorithms. Springer, London 2013.
[68] Walesiak M., Dudek A.: clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.43-4, 2014.
[69] Yera A., Arbelaitz O., Jodra J., Gurrutxaga I., Pérez J., Muguerza J.: Analysis of several decision fusion strategies for clustering validation. Strategy definition, experiments and validation. Pattern Recognition Letters 85, 2017, 42–48.
[70] Zahn C.: Graph-Theoretical Methods For Detecting And Describing Gestalt Clusters. IEEE Transactions on Computers C-20, 1971, 68–86.
[71] Žalik K., Žalik B.: Validity index for clusters of different sizes and densities. Pattern Recognition Letters 32(2), 2011, 221–234.
[72] Zhong C., Miao D., Wang R.: A Graph-Theoretical Clustering Method Based On Two Rounds Of Minimum Spanning Trees. Pattern Recognition 43, 2010, 752–766.

M.Sc. Eng. Taras Panskyi
e-mail: tpanski@kis.p.lodz.pl

T. Panskyi received his M.Sc. from the Lviv Polytechnic National University, Institute of Telecommunications, Radioelectronics, and Electronic Engineering. Currently, he is a Ph.D. student at the Lodz University of Technology, Institute of Applied Computer Science. His areas of interest are data clustering, clustering validation indices, clusterability, etc. He has published more than 20 technical articles.

http://orcid.org/0000-0002-0416-8711

Prof. D.Sc. Eng. Volodymyr Mosorov
e-mail: w.mosorow@kis.p.lodz.pl

V. Mosorov received his Ph.D. in 1998 from the Lviv Polytechnic National University, Ukraine. He received his habilitation degree from AGH University of Science and Technology in Krakow Poland in 2009. Currently, he holds a position as an associate professor at the Institute of Applied Computer Science of Lodz University of Technology, Poland. His research interests include data mining, clustering, etc. He has published more than 110 technical articles.

http://orcid.org/0000-0001-6016-8671

otrzymano/received: 11.03.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2597

FUZZY APPROACH TO DEVICE LOCALIZATION BASED ON WIRELESS NETWORK SIGNAL STRENGTH

Michał Socha, Wojciech Górka, Marcin Michalak
Research Network Łukasiewicz, Institute of Innovative Technologies EMAG, Katowice, Poland

Abstract. *The paper presents an original approach to device location detection in a building. The new method is based on a map of individual interiors, drawn up based on the measurements of the strength of wireless network signals for each building venue. The device is initially assigned to all venues whose descriptions sufficiently correspond with the current measurements taken by the device. A fuzzy assignment level for each of the potentially considered venues depends on the difference between the averaged network strengths for the venue and the signal strengths currently measured with the device for localization purposes. Ultimately, the device is assigned to the venue with the highest level of assignment.*

Keywords: wireless networks, fuzzy sets, device location

ROZMYTE PODEJŚCIE DO LOKALIZACJI URZĄDZEŃ NA PODSTAWIE SIŁY SYGNAŁU SIECI BEZPRZEWODOWYCH

Streszczenie. *W pracy przedstawiono oryginalną metodę lokalizowania urządzeń w budynku. Nowa metoda bazuje na mapie poszczególnych pomieszczeń, stworzonej w oparciu o pomiary sygnałów sieci bezprzewodowych zmierzonych w tych pomieszczeniach. Wstępnie urządzenie przypisywane jest do tych pomieszczeń, których opis w odpowiednim stopniu pokrywa się z pomiarami dokonanymi przez urządzenie. Stopień rozmytej przynależności do każdego z wstępnie wytypowanych pomieszczeń zależy z kolei od różnicy pomiędzy uśrednionymi wartościami sygnałów sieci bezprzewodowych i aktualnie zmierzonymi do celów lokalizacji. Ostatecznie urządzeniu przypisywane jest to pomieszczenie, dla którego stopień przynależności jest największy.*

Słowa kluczowe: sieci bezprzewodowe, zbiory rozmyte, lokalizowanie urządzeń

1. Introduction

The use of location tracking is becoming increasingly popular in many services used on a daily basis. Traffic navigation or location context services that can be used outdoors are mainly based on GPS [5]. There are many possible uses of indoor location tracking. A good example can be guidance apps at museums [8, 9]. A museum app can contextually present descriptions of nearby exhibits. It can also be an indoor map supported with information about the user's current position. Location tracking can be also useful for disabled people and health care buildings [2, 14]: monitoring people entering or leaving restricted areas and venues, equipment location, searching for people, guiding services. Unfortunately, GPS localization does not work properly indoors. There is a strong need to provide indoor location tracking in a different manner. There have been many solutions for indoor localization developed so far. Just to mention systems based on infrared sensors [15, 18], ultrasound [16], or magnetic fields [4, 6]. In recent years we have seen the growing popularity of Bluetooth-based location systems named iBeacons [17]. However, these systems require the installation of additional devices.

On the other hand, WiFi-based location tracking is also becoming popular: as there are many access points in offices using WiFi, such location tracking should be costless. Since RADAR [1], many WiFi-based indoor localization solutions were proposed and developed [2, 7, 9, 11–14]. In general, WiFi location tracking uses a map as a reference. There are many types of maps, which can be divided into discrete and continuous maps [10]. Continuous maps allow determining the position as a point on the map. Such precision in WiFi location tracking is possible to obtain under certain conditions: homogeneous devices, defined obstacles, no interference from people. It is useful when we have to know the precise position, e.g. to control a robot [11]. A discrete map usually defines certain zones which are treated as positions on the map. Such location tracking is less precise (because we cannot determine the position, only the zone) but higher location precision cannot be reliably achieved using the most popular WiFi devices (different radio parameters, different measurement techniques).

The continuous approach usually does not require the venue map preparation stage. The system should have information about the location of access points, the structure of obstacles, the strength of the signal in the current position, and based on this information it determines the location in the venue [11]. However, the calculation is vulnerable to errors due to many factors, such as weather conditions, interference from human bodies, different equipment [7].

Discrete positioning usually requires building a pre-learned set of fingerprints to infer the position of a device. There are many solutions to this issue: hidden Markov models [12], Bayesian filtering [21], clustering techniques [19], and genetic algorithms [3].

WiFi-based localization is currently easy to implement in many buildings using existing infrastructure. Discrete positioning seems to be a good choice considering the variety of WiFi devices, the precision of signal strength measurement, and the most common needs. The main contribution of this paper is to present a new discrete, WiFi signal-based positioning algorithm with the support of an SQL database and specified queries. The new method is based on a pre-learned set of fingerprints of defined locations. It uses fuzzy logic to determine the location and the SQL database and queries directly on the collected data. Additionally, it is possible to regularly update the data – location fingerprints – and the operation does not require restarting the system. In opposition to the mentioned solutions, the location tracking is based on the rooms specified by the system operator, not on the zones calculated by the algorithm. As a result, it is possible to correct a location tracking procedure or a venue map definition manually. From the functional side, we receive specific information about whether the user is in a given room, if the user is missing, or their location cannot be determined.

The paper presents only the logical aspects of the SQL-based fuzzy location procedure: the process of building the map and the algorithm of fuzzy location. The other aspects of the device location such as software and hardware requirements, database structure, and SQL queries are not provided.

The paper is organized as follows: it starts with a short description of existing approaches to device location tracking issues with the discussion of the necessity of a new solution; afterward, the definition of the building map is presented, which is based on the wireless network's signal strength measured in the venues in question; the next part presents the fuzzy location tracking procedure in details and with a simple example incorporating artificial data; finally, the description of experiments and their results are presented. The paper ends with conclusions and possibilities of future works.

2. Related works

As was already mentioned in the previous section, there are many known implementations of WiFi-based indoor location tracking procedures. One of the first and most frequently mentioned is a system called RADAR [1]. In the context of WiFi location tracking, a k-NN classifier was used to decide if a person is in a discrete location (close to the previously measured point). The authors are also considering the direction of the localized person in relation to the transmitter. Eventually, the direction is considered unnecessary.

Another work on WiFi localization is related to SLAM (Simultaneous Localization and Mapping) systems [11]. The authors aimed to build a solution that does not need to use fingerprint maps (building a fingerprint map may be an expensive task). Their solution has to precisely obtain the position which is used to navigate the robot. The position is evaluated based on the measured signal strength of several WiFi sensors. Such algorithms also need to evaluate the signal loss due to propagation and obstacles.

The fingerprint solution is a hierarchical topology based on WiFi Indoor Localization [10]. The test environment is built and then split into smaller sub-zones with a reduced number of WiFi Access Points (AP) and reference positions to be identified. The hierarchical partition of the map is created using a KMeans clustering algorithm and the Calinski–Harabasz Index. The authors try to use different classifiers: K-NN, SVM, FURIA. According to the authors, the results are superior to those of the RADAR system. The hierarchical approach makes the solution scalable from small to rather huge environments (several floors, many venues).

The authors of the Fuzzy Logic Based System [7] focus on measurement issues. The signal level can be influenced by weather conditions, nearby devices, obstacles, people in rooms, different radio signal frequencies. These all lead to drops and peeks in the signal strength of access points. The authors propose several pre-processing and post-processing techniques to improve the quality of measurements. They also propose the fuzzy calculations approach to position detection.

3. Building map

The natural solution to the issue of determining the current position is to measure distances from the reference points of known coordinates. Based on these data and after some necessary calculations the location can be specified. Generalizing the problem, it can be stated that the reference points impose (constitute) a local coordinate system in relation to which furthe analyses will be carried out. In the defined area, it is possible to measure the distance to reference points and in this way determine the coordinates of a point in the local coordinate system.

A generalized localization case consists of two stages. The first stage is to define the local coordinate system and determine the reference points appropriate to the local system. The second stage is to compute the relation between the measuring point and the reference points. Finally, after making the calculations, the measuring point receives the coordinates in the local system coordination system.

The first stage involves installing the infrastructure, e.g. setting reference points. This was called defining a map of a building covered by the location system.

The process of making the map involves selecting rooms in the building. Not all rooms in the building were considered to be attached to the building map. The selection of rooms was conducted according to the following criteria:

- proximity of the rooms on the same floor, horizontal distance;
- proximity to the rooms on different floors, vertical distance;
- different types of walls separating the rooms;
- distant rooms with multiple separating walls;
- rooms between which there is visibility through the windows;
- large rooms.

A large room was divided into 4 subzones. This selection of rooms aimed at using the results collected during the experiment for later analysis.

The data collection procedure was as follows:

1) A person enters the room and stops near its centre, e.g. at the reference point. (The room divided into 4 subzones is an exception: the procedure was performed separately for each subzone).
2) All WiFi signals detected by the phone are measured (this step required a dedicated application for Android OS).
3) Measurements (WIFIs, strengths, room, device id, etc.) are stored in the database.

Additionally, the order of measurements in the location and the order of locations are saved.

The sample data are presented in Table 1, where:

- id – database technical row identifier,
- device_id – identifier of the device used to provide measurements,
- BSSID – network identifier,
- point_id – reference point identifier which represents a room,
- strength –signal strength in dB measured at the reference point,
- m_in_point – order number, grouping rows in one measure probe and numbering according to the reference point,
- m_in_total – order number, the same as m_in_point but numbering according to all gathered samples.

For better generalization of measurements, five different devices operating under Android OS were used. Data acquisition was performed at different times and at various intervals to avoid measurement result contamination with particular environmental features in which the measurements were made. Additionally, the measurements were taken manually so the position of the measuring device relative to the designated reference point in the room changed slightly with each measurement. Over 30 measurements were made at each point. It was the consequence of the strength variability of the measured WiFi signal resulting from many factors independent of the measurement process and of the WiFi technology itself.

After 30 individual measurements in all reference points have been taken, each of them was described by the following dataset:

- BSSID,
- average signal strength,
- standard signal strength deviation.

This dataset defines the coordinates of the reference point in the map space. It should be emphasized that due to the attenuation of the WiFi signal, the sets describing reference points contain a different number of BSSIDs. Of course, the average signal strength and standard deviation are also different. Table 2 presents sample descriptions of two reference points.

Finally, after collecting all the measurements for each reference point, the coordinates identifying the point in the building space were determined using the WiFi signal strength. For each reference point, the average and standard deviation of the WiFi signal strength of each WiFi network available at that point were determined. The networks were identified by BSSID. As a result, each reference point is identified by the coordinates resulting from all the WiFi networks available at the point. Additionally, for each average signal strength, the standard deviation is calculated that will be used in the localization procedure.

Table 1. Sample data describing WiFi signal strength in different locations measured with different devices

id	device_id	BSSID	point_id	strength	m_in_point	m_in_total
1800	1	06:d6:aa:35:1b:5f	12	-42	6	70
1801	1	00:17:c5:e3:75:5e	12	-66	6	70
1802	1	00:17:c5:e3:75:5d	12	-66	6	70
1803	1	00:17:c5:e3:76:e9	12	-75	6	70
1804	1	00:17:c5:e3:76:ea	12	-75	6	70
1805	1	00:17:c5:e4:42:9b	12	-82	6	70
1806	1	00:17:c5:e4:42:9c	12	-83	6	70
3249	1	06:d6:aa:35:1b:5f	12	-26	7	105
3251	1	00:17:c5:e3:75:5e	12	-65	7	105
3255	1	00:17:c5:e3:75:5d	12	-66	7	105
3258	1	00:17:c5:e3:76:e9	12	-66	7	105
3261	1	00:17:c5:e3:76:ea	12	-66	7	105
3287	1	06:d6:aa:35:1b:5f	12	-30	8	106
3289	1	00:17:c5:e3:75:5e	12	-67	8	106
3291	1	00:17:c5:e3:75:5d	12	-67	8	106
3293	1	00:17:c5:e3:76:e9	12	-76	8	106
3295	1	00:17:c5:e3:76:ea	12	-76	8	106
3298	1	68:a0:f6:17:32:fa	12	-82	8	106
3303	1	00:17:c5:e4:42:9b	12	-83	8	106
3307	1	00:17:c5:e4:42:9c	12	-86	8	106

Table 2. Sample descriptions of two reference points

point_id	BSSID	strength_avg	strength_stddev
13	06:d6:aa:35:1b:5f	-46.48	9.25
13	0c:37:dc:b7:96:f6	-80.54	4.33
13	e8:de:27:b7:6a:fe	-83.79	2.42
13	4c:5e:0c:7e:ca:ed	-61.09	7.83
13	56:67:11:e4:40:f4	-89.80	4.21
13	92:5c:14:56:13:31	-90.00	3.46
13	92:5c:14:ac:20:26	-90.20	1.30
13	90:5c:44:ac:20:26	-90.89	1.05
13	54:67:51:e4:40:f4	-88.00	4.56
13	9c:b2:b2:5f:b1:67	-84.00	7.44
13	68:a0:f6:17:32:fa	-70.50	6.15
14	00:1d:7e:bc:45:af	-87.33	6.35
14	00:17:c5:e3:71:2f	-80.83	4.46
14	00:17:c5:e3:71:fe	-68.41	7.27
14	00:17:c5:e3:76:ea	-92.17	2.71
14	00:17:c5:e3:77:b0	-93.00	0.00
14	00:17:c5:e3:81:01	-82.96	3.78
14	00:17:c5:e3:f3:c1	-88.40	3.51
14	00:17:c5:e3:f7:2a	-77.91	4.07
14	00:17:c5:e3:71:2e	-80.87	4.20

4. Fuzzy location procedure with an example

In the research, the fuzzy representation of venue assignment was assumed. The notion of fuzziness – proposed over 50 years ago in [20] – is one of the most common approaches of uncertainty representation. Such ability of uncertainty processing is a significant advantage when variable signals – like the WiFi network strength – are being analyzed.

The positioning of the device is similar to the map development procedure described earlier. Unlike map creation, positioning is an unattended process and is performed once for each location determination attempt. At the time of the measurement, the position of the device (the person holding the telephone) is unknown and the measurement is performed once or several times at short intervals. Repeating the signal strength measurement in several second intervals is intended to reduce the impact of WiFi signal variability. As a result, the networks whose signals are very week (or cannot be detected by the devices) are eliminated. Data concerning networks and signal strength collected by the device that wants to know its location are compared with data describing reference points. The similarity of an unknown-location point to the reference points (for which the location is known) is determined. As a result of the algorithm, we get a list of similarities. After sorting, at the top of the list is the reference point nearest to which the measuring device (smartphone) is located. This reference point is considered the device position. The presented fuzzy location procedure will now be described in more detail.

All visible WiFi networks are read at the unknown location point. The networks are identified thanks to BSSID. For each network, the signal power that reaches the measuring device from the access point is read. An example dataset is provided below in Table 3.

Table 3. Detected WiFi networks and their signal strength

BSSID	signal strength
4e:8e:ff:8c:5a:8f	-31
4c:9e:ff:8c:5a:8e	-33
4c:9e:ff:8c:5a:8f	-43
4e:8f:ff:8c:5a:80	-43
4c:9e:ff:8e:1c:9a	-54
4e:1f:ff:8e:1f:10	-59
4e:20:ff:8e:1f:21	-63
4c:9e:ff:8e:1f:20	-64
4e:9a:ff:8e:1c:9b	-67
4c:9e:ff:8e:1c:9b	-67
4e:9b:ff:8e:1c:9c	-67

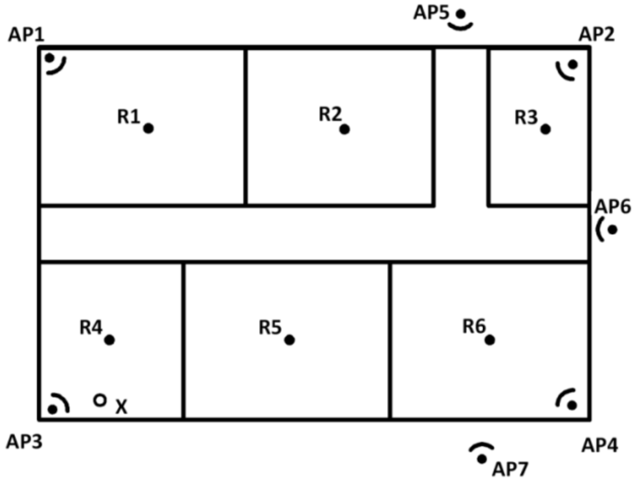


Fig. 1. A sample building floor plan

Table 4. Detected WiFi networks and their signal strength

Description of point R1		
Net ID	Average signal strength (s)	Standard deviation (std)
AP1	-42	3
AP2	-68	7
AP3	-55	6
AP5	-91	4

Description of point R2		
Net ID	Average signal strength (s)	Standard deviation (std)
AP1	-62	5
AP2	-59	6
AP3	-69	5
AP4	-72	7
AP5	-82	9
AP6	-94	6

Description of point R3		
Net ID	Average signal strength (s)	Standard deviation (std)
AP1	-70	8
AP2	-38	6
AP4	-65	7
AP5	-74	9
AP6	-71	7

Description of point R4		
Net ID	Average signal strength (s)	Standard deviation (std)
AP1	-57	9
AP2	-98	7
AP3	-41	6
AP4	-79	8
AP6	-95	5
AP7	-87	9

Description of point R5		
Net ID	Average signal strength (s)	Standard deviation (std)
AP1	-61	7
AP2	-89	8
AP3	-54	5
AP4	-68	8
AP6	-89	7
AP7	-78	5

Description of point R6		
Net ID	Average signal strength (s)	Standard deviation (std)
AP1	-85	9
AP2	-78	7
AP3	-79	6
AP4	-39	8
AP5	-87	5
AP6	-74	9
AP7	-77	7

The process of measuring and determining the position of the device can be presented in several steps described below. Prior development of the map is necessary for determining the position, as described above. In order to illustrate the position determination process, we assume that there is a map prepared for the building fragment as shown in Fig. 1.

Points R1 to R6 are the map reference points, example descriptions of these points are presented in Table 4. Point X is the measuring point for which the location will be determined, an example measurement made here is shown in Table 5. Points AP1 to AP7 represent the location of the wireless network access points. AP5 to AP7 are located outside the analyzed area, but the signal of these access points is visible in the area covered by the map. In the example tables, instead of the full identifier of the BSSID network, short identifiers were included to increase the transparency of the example. Similarly, most numbers are represented as integers.

The fuzzy location procedure consists of the following steps:

Step 1: A person who wants to determine their position in the building, goes to any place covered by the map. Using the mobile device, the person measures WiFi networks visible there. Together with the BSSID network, the strengths of individual signals are read. An example measurement taken in room 4 at point X can provide the following results as presented in Table 5.

Step 2: The collected results are compared with the building map. The quantity condition of the network is checked first. For each reference point, we compare a set of network identifiers assigned to that point with a set of network identifiers visible at point X. Next we determine a common part of these two sets.

Table 5. Results of measurement made at location X

Net ID	signal strength (s)
AP1	-68
AP3	-38
AP4	-78
AP7	-75

Using the sample data, it can be seen that the measurement point X and the reference point R3 have three common network identifiers: AP1, AP4. In the case of X and R4, common identifiers are: AP1, AP3, AP4, AP6.

Step 3: For each reference point, a network factor is determined that reflects the ratio of found networks (in reference to the number of networks assigned to the reference point). If the factor takes on a value of more than 0.5, the reference point is taken into account in the next steps. Otherwise, the reference point is rejected and is not included in further calculations. This approach reduces the computational complexity of position determination but with the possible cost of decreased location accuracy (or the time of location increase).

The networks factor for the pair measurement point and the reference point is determined by the formula:

$$f_{X,R_j} = \frac{n_{X,R_j}}{n_{R_j}} \tag{1}$$

where f_{X,R_j} is the network factor, n_{X,R_j} is the number of common network identifiers for sets describing the X measurement point and reference point, R_j, and n_{R_j} means the number of network identifiers describing the reference point R_j.

For sample data, the factor summary is presented in Table 6.

Table 6. Network factors for a reference point, based on measurement from location X

j	n_{X,R_j}	n_{R_j}	f_{X,R_j}
1	2	4	0.50
2	3	6	0.50
3	2	5	0.40
4	4	6	0.67
5	4	6	0.67
6	4	7	0.57

Based on network factors, reference points R1, R2 and R3 will be rejected from further processing.

Then, for the remaining reference points, the strengths of the network signals from the measurement point set are analyzed, along with the mean signal strengths and the standard deviation of the network in the set for each reference point. For each pair of sets the distance factor is determined according to the following formula:

$$d_{X,R_j} = \sum_{i=1}^{n_{X,R_j}} rm_i(X,R_j) \cdot c_i(X,R_j) \tag{2}$$

For each considered network – i – the multiplication of two functions is performed. The first can be called the reference point assignment and is defined as the fuzzy assignment of the measurement point X to the reference point R_j:

$$rm_i(X,R_j) = \begin{cases} 1, & |s(R_j,i) - s(X,i)| < 1 \\ 0.5, & |s(R_j,i) - s(X,i)| \in [std(R_j,i), 2 \cdot std(R_j,i)) \\ 0.25, & |s(R_j,i) - s(X,i)| \in [2 \cdot std(R_j,i), 3 \cdot std(R_j,i)) \\ 0, & |s(R_j,i) - s(X,i)| \geq 3 \cdot std(R_j,i) \end{cases} \tag{3}$$

where $s(R_j,i)$ is the average signal strength for the network from the description of the reference point, R_j is the measured signal strength for the network at the measuring point R_j, and $std(R_j,i)$ is the standard deviation of the signal strength for the network i from the description of the reference point R_j.

In the early stages of location procedure development, many different fuzzy assignment functions were taken into consideration: triangular, piece-wise linear, trapezoidal, trigonometrical (cosine and arcus tangent-based), and many more. However, from the point of view of location accuracy and implementation complexity (SQL queries) the presented solution gave the best results.

The second function c represents the signal strength correction factor and is defined as follows:

$$c_i(X,R_j) = \begin{cases} 0.2, & s(R_j,i) \in (-100;-90] \\ 0.5, & s(R_j,i) \in (-90;\ -80] \\ 0.8, & s(R_j,i) \in (-80;-55] \\ 1, & s(R_j,i) \in (-55;-35] \\ 0.8, & s(R_j,i) \in (-35;0] \end{cases} \tag{4}$$

The results of using distance only based on reference point assignment functions were unsatisfactory. This was mostly due to the fact that the radio signal strength results from many other factors such as transmitter power, the output of the transmission and reception antennas, attenuation on the receiver and transmitter, and signal reflections (amplifying or weakening the signal in the receiver). This implied introducing an additional correction factor whose final form, presented above, was obtained through trial and error.

Step 4: After performing the above calculations, the determined distance factor is corrected with the previously calculated network factor f_{X,R_j}. The correction takes place for each reference point.

This is expressed by the formula:

$$f_{X,R_j} = \frac{n_{X,R_j}}{n_{R_j}} \tag{5}$$

As a result, after performing the above calculations, we obtain a measure matching measurement point X to individual reference points. As the obtained all room assignment levels have positive, non-limited values, they must be normalized. The range $[0; \max_j m_{X,R_j}]$ is scaled linearly to [0;1].

The results showing the measure of matching the X measuring point to the map reference points are presented in Table 7.

Table 7. Raw levels of assignment and normalized levels of assignment

j	m_{X,R_j}	normalized m_{X,R_j}
4	1.63	1.00
5	1.07	0.65
6	0.60	0.37

After sorting the table according to the value of the measure of matching from the highest to the lowest value, in the first place we obtain the location of the measurement point. In the example, this is the reference point, i.e. based on the performed calculations it can be concluded that the measurement was conducted in room 4. The obtained results can be filtered by rejecting the values of the matching measure which did not exceed the limit value. Such filtering, however, requires additional research that will allow to deliberately set a results rejection threshold. Without the use of such filtering, the order in the list shows the distance of the measuring point from the reference point. The measuring point is located closest to the reference point in the first position and farthest from the reference point in the last position.

5. Experiments

5.1. Data acquisition

The building, in which the procedure was developed and applied, is a three-storey one with several dozen rooms on each floor. In the experiments, 20 of them were taken into consideration. One of them was the conference hall which was divided into four separate areas, so the final number of locations was 23. The criterion for selecting these rooms was the frequency of changes of the monitored people's locations.

As it occurred, inside these rooms over 100 different wireless network IDs were detected. In each room, 30 measurements of wireless network signal levels were conducted. These measurements were taken using several different types of mobile phones. The highest number of detected signals in a room was almost 35. This implies that the raw data were rather sparse. It also confirms that in the presented problem a non-resistant method could not be applied for missing values.

Finally, the collected dataset consisted of 690 objects and over 100 features. The measured signal strength was expressed in dB and, as it is assumed in the wireless network area, the signal strength is a negative number, and as the level of the signal decreases, the value of the strength decreases too.

A sample part of the collected data is presented in Table 8.

Table 8. Sample data acquisition results

mes_id	dev_id	n001	…	n012	n013	n014	n015	…	n100	ven_id
595803	2				-71	-71				16
595804	1			-33	-63	-63	-73			16
595812	2									17
595813	2									17
595814	1			-63	-62	-63	-61			17
595815	1									17
595831	2				-80	-80				14
595832	1				-67	-66	-76			14
595833	1			-80						14

The mes_id column represents the specific measurement identifier, the dev_id column is connected with the specific measurement device, columns from n001 up to n100 deal with wireless networks, and finally – ven_id – is the venue identifier in the building. The collected data, due to the nature of the environment (dozens of networks and hundreds of measurements detecting only several networks), are characterized by a very high number of missing values. The total amount of non-missing ones was approximately 6%. For that reason, the application of the most popular multiclass classifiers was impossible.

5.2. Simple experiments

All experiments were carried out in the leave-one-out/stratified cross-fold validation method. In each of the 30 iterations, exactly one measurement from each venue was moved to the test set. This ensured that the fuzzy maps of venues were built on 29 measurements (each of the maps). On the other hand, the proportion of venue representation in the training and test set was equal. The proposed scheme also ensured that the prediction accuracy, measured as the fraction of correctly and incorrectly classified positions, is equal to the class-weighted accuracy. Finally, it is possible to present the prediction results in one confusion matrix as each measurement is only taken once as the testing object.

The results of the localization procedure are presented in Table 9.

Table 9. Confusion matrix for the classified objects

	predicted venue																						
original venue	1	2	3	4	5	6	7	8	9	10	12	13	14	17	18	19	20	21	22	24	25	26	27
1	21	3	6																				
2	2	20	1	5					2														
3	1		28						1														
4	3	13	1	13																			
5					30																		
6					1	28													1				
7	3			1	1		13	2						6	1				3				
8							1	28											1				
9									28	2													
10									3	26									1				
12											25						1		4				
13									2			27							1				
14			2										23				4		1				
17					9	1	1							8	9	2							
18					1										27				1				1
19		1							1						5	19	3		1				
20													1			1	28						
21															10	1		15	4				
22			1								2						4		22	1			
24											1				1				4	24			
25									1										3		26		
26	1									2						1	3	4	12			7	
27															1								29

The total accuracy reaches a level of almost 75% (74.64%). The least class prediction accuracy was at the level of 23.33% but it should rather be considered the outlier because the median is 83.33%, which is quite more than the average. It is also worth noticing that none of the objects was unclassified. In other words: each wireless network signal level measurement was classified to one of the possible venues.

Let us remember that the first four places can be interpreted as one location (the conference hall divided into four regions). That simplification increases the prediction accuracy up to 78.63% and the median up to 86.66%. However, from the perspective of utility, we would not consider the class aggregation in further experiments.

5.3. Haste makes waste

Typically in the case of classification or regression tasks, a situation where an object is not classified to any known class (it is not possible to predict the real value of the dependent variable) is usually considered the misclassification (the error of predicting a known value). This results from the fact that the environment of independent variables remains unchanged. This is very easy to explain: in the case of character recognition, the image of acquired pixels does not change in time; while in the case of real values prediction, the constant current value of the predicted variable should be constant due to constant independent variables values. However, in the case in question, when a person enters the room, we can agree that as long as that person stays in the same room the prediction may be done at a certain cost in terms of prediction duration time. In other words: in such a case – haste makes waste. We agree for the longer time of prediction provided that the prediction accuracy also increases. This issue will be explained later in this section.

The maximum number of networks in the venue maps is 33. This leads to the remark that we may define the condition of the minimal percentage of venue map networks to be recognized by the measuring device (network factor) as the necessary condition for taking the map into consideration. For example: if the map is built on the strength of 10 network signals and the assumed minimal percentage of the measured signal is 25%, at least three signal levels must fulfill the map criteria to begin to consider the location pointed by that map.

Let us consider all possible signal threshold levels, varying from 0 to 1. The smallest sensible value of the threshold is 0.03 as there is no map built on more than 30 network signals. Then, similar experiments were carried out for each considered threshold. The results are presented in Figure 2.

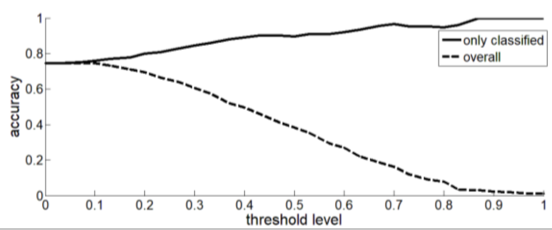


Fig. 2. Data classification accuracy

The solid line represents how the classification accuracy of the map covered the sample changes due to the threshold increase. Starting from the value of 0.7464 (which is the same as in the previous experiments) it increases to 1.0 (for the threshold of 0.8700). The dotted line also represents the classification accuracy but in this case, unclassified objects are considered wrongly classified. We can observe the intuitive effect: as the threshold increases, the accuracy decreases. The decrease in the accuracy is caused by the decrease in the data coverage. In the next figure (Figure 3) we can compare both tendencies.

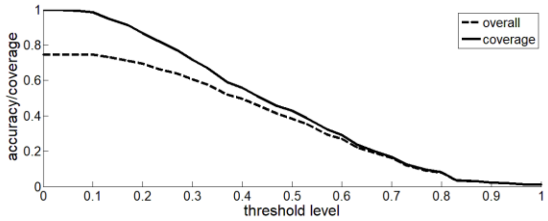


Fig. 3. Comparison of accuracy and coverage decrease

The intuitive reflection about the increasing range of unclassified objects says that the model degrades with the threshold level.

It is a well-known procedure of parametrized classifier quality description to present the chart called ROC – Receiver Operator Curve. In terms of the classification issue, it shows how the true positive and false positive rates change due to the classifier parameters modification. In our case, it is worth taking into consideration two other criteria. We are interested in high accuracy and low non-coverage (difference between a set of objects and a set of objects classified by the model/map - the measure complement to the coverage). Let us check how this dependence looks like in the case of our predictor. ROC is presented in Figure 4.

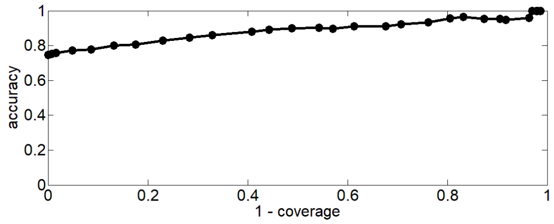


Fig. 4. Comparison of accuracy and coverage decrease

The presented figures confirm that in this specific case, with changing environmental conditions, the "lazy" decision is preferred due to its correctness and it is worth considering the threshold increase in the real application. The following section will be devoted to proper threshold level selection.

5.4. Network factor level selection

As it was mentioned in the previous section, the main goal of the issue is to predict the proper location even at the cost of time spent to obtain this prediction. So it is important to define the appropriate compromise between the coverage of the model and the accuracy of the model predictions (limited only to the final statements – unclassified objects do not degrade the prediction accuracy).

One of the most intuitive approaches is to compare the coverage decrease and the accuracy increase due to the threshold increase. It is easy to find the threshold level in which the mentioned two measurements intersect. The comparison of data coverage and covered data prediction accuracy is presented in Fig. 5.

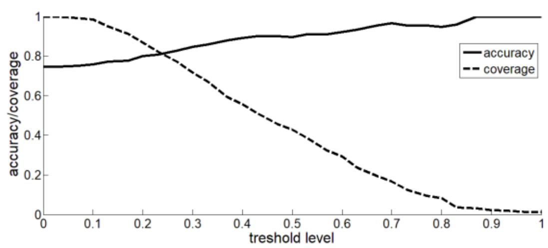


Fig. 5. Coverage and covered data accuracy

The analysis of that figure should imply the proper threshold level on the value equal to 0.23. A higher level will cause the coverage to be lower than the accuracy. With the assumed threshold, the accuracy would be equal to 0.807.

Another approach is to build the average of mentioned measures. The average result is presented in Fig. 6.

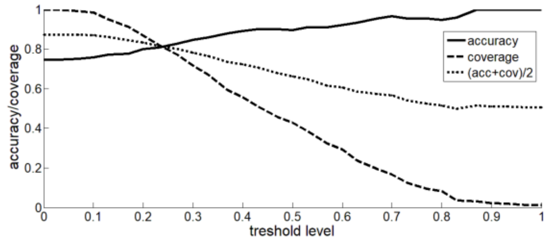


Fig. 6. Coverage and covered data accuracy together with their average

It occurs that such averaged measure decreases at the threshold level equal to just 0.1. That means that the optimal threshold value is a step-before one (0.066) and this situation corresponds with the initial prediction accuracy: 0.7518.

The mentioned two approaches are based on the raw quality measurement values. However, the most important quality criterion should take into consideration the real application results. The increase of the threshold value simultaneously increases the accuracy and decreases the data coverage, which was presented in the previous section. The crucial question is ”How long can the threshold be increased to assure all venues be covered by the prediction results?’’. In other words, what is the maximum threshold level that assures at least one correct classification into each of the possible venues?

This leads us to the results presented in Fig.7.

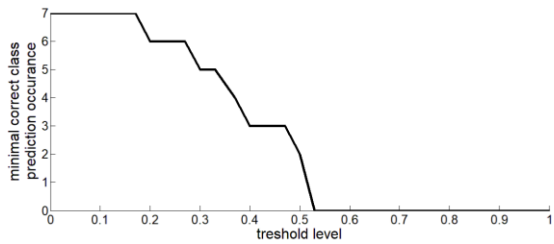


Fig. 7. The minimum number of objects classified correctly to the class

It occurs that the maximum threshold level that assures at least one correct object classification (for each possible object class) is 0.50. Up to this level, it is only a matter of time to obtain the correct location for the considered device. Over this level, there is a possibility that we are in a venue that we will never be classified into. The classification accuracy at this threshold level is 0.8953.

6. Conclusions and further works

The paper presents a novel approach to the location issue based on the wireless network signal measurement. A single measurement observed with the mobile phone is compared with the previously built description of all venues in the building. The final venue assignment is made based on the fuzzy assignment to each of the considered rooms.

The developed strategy of building the description (map) and the fuzzy location procedure is now applied in the people location tracking system developed at the Institute of Innovative Technologies EMAG.

The presented solution is adaptable but it requires an advanced implementation phase: the process of creating the map of the building based on averaged measurements. Additionally, the analysis of the proper network factor should be performed for each building separately.

Our later works will focus on further fuzzy location improvement – especially on using other wireless signal measures to improve the location accuracy (e.g. beacons).

The data acquired during experiments are also a valuable source of information on how the location procedure duration depends on the network factor: it was shown that increasing this factor improves the classification accuracy but only if the condition of the minimum number of detected networks is fulfilled. A more detailed analysis of these durations may help to define a longer time of procedure application with the advantage of better location accuracy.

References

[1] Bahl P., Padmanabhan V. N.: RADAR: An in-building RF-based user location and tracking system. Proceedings IEEE INFOCOM 2000, 775–784.

[2] Benavente-Peces C. et al.: Global System for Localization and Guidance of Dependant People: Indoor and Outdoor Technologies Integration. Lecture Notes in Computer Science 5597, 2009, 82–89.

[3] Chintalapudi K., PadmanabhaIyer A., Padmanabhan V.: Indoor localization without the pain. Proceedings of the sixteenth annual international conference on Mobile computing and networking, 2010, 173–184.

[4] Chung J. et al.: Indoor Location Sensing Using Geo-Magnetism. Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services, 2011, 141–154.

[5] Enge P., Misra P.: Special issue on Global Positioning System. Proceedings of the IEEE 87(1), 1999, 3–15.

[6] Galván-Tejada C. E., Carrasco-Jimenez J. C., Brena R.: Location Identification Using a Magnetic-Field-Based FFT Signature, Lecture Notes in Computer Science 8276, 2013, 9–16.

[7] Garcia-Valverde, T., Garcia-Sola, A., Hagras, H., Dooley, J. A., Callaghan, V., Botia, J. A.: A fuzzy logic-based system for indoor localization using WiFi in ambient intelligent environments. IEEE Transactions on Fuzzy Systems 21(4), 2013, 702–718.

[8] Górka W., Piasecki A. Socha M.: Mobile application supporting universal access to culture, taking into account the needs of disabled people. Proceedings of the 11th Scientific Conference Internet in the Information Society, 2016, 191–201.

[9] Hammadi O.A., Hebsi A. A., Zemerly M. J., Ng J. W. P.: Indoor localization and guidance using portable smartphones. Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2012, 337–341.

[10] Hernández N., Alonso J. M., Ocaña M.: Hierarchical Approach to Enhancing Topology-based WiFi Indoor Localization in Large Environments. Journal of Multiple-Valued Logic and Soft Computing 26(3-5), 2016, 221–241.

[11] Herranz, F., Llamazares, Á., Molinos, E., Ocaña, M., Sotelo, M. A.: WiFi SLAM algorithms: An experimental comparison. Robotica, 34(4), 2016, 837–858.

[12] Krumm J., Horvitz E.: Locadio: Inferring motion and location from Wi-Fi signal strengths. Prpceedings of The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2004, 4–13.

[13] Liu H. et al: Survey of wireless indoor positioning techniques and systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 37:6, 2007, 1067–1080.

[14] López E., Barea R., Bergasa L. M., Escudero M. S.: A human-robot cooperative learning system for easy installation of assistant robots in new working environments. Journal of Intelligent and Robotic System 40(3), 2004, 233–265.

[15] Polito S. et al.: Performance evaluation of active RFID location systems based on RF power measures. IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, 2007, 1–5.

[16] Priyantha N. B., Chakraborty A., Balakrishnan H.: The cricket location support system. Proceedings of the Annual International Conference on Mobile Computing and Networking, 2000, 32–43.

[17] Steinbuss S., Holtkamp B., Opriel S.: HANDELkompetent – Situation Aware Learning in Retail. Procedia Manufacturing 9, 2017, 245–253.

[18] Want R., Hopper A., Falcao V., Gibbons J.: The Active Badge Location System. ACM Transactions on Information Systems 10(1), 1992, 91–102.

[19] Youssef M., Agrawala A.: The Horus WLAN location determination system. Proceedings of the 3rd international conference on Mobile systems, applications, and services, 2005, 205–218.

[20] Zadeh L.: Fuzzy sets. Information and Control 8(3), 1965, 338–353.

[21] Zaruba G., Huber M., Kamangar F., Chlamtac I.: Indoor location tracking using RSSI readings from a single Wi-Fi access point. Wireless Networks 13(2), 2007, 221–235.

M.Sc. Michał Socha
e-mail: Michal.Socha@emag.lukasiewicz.gov.pl

Michał Socha was born in Poland in 1972. He received his M.Sc. Eng. in computer science from the Silesian University of Technology in 1997 and has begun Ph.D. degree study in 2020 at the same university. His scientific interests cover machine learning, geographic information systems, software architecture design. He is an author and co-author of over 20 scientific papers.

https://orcid.org/0000-0002-7439-4001

M.Sc. Wojciech Górka
e-mail: Wojciech.Gorka@emag.lukasiewicz.gov.pl

Wojciech Górka was born in Poland in 1979. He received his M.Sc. Eng. in computer science from the Silesian University of Technology in 2003. Since 2006 he has been working at the Łukasiewicz EMAG research institute (former ITI EMAG). His scientific interests are natural language processing, technology supporting the blind and deaf, and mobile devices and their ability to locate in indoor areas.

https://orcid.org/0000-0003-2124-7734

Ph.D. Marcin Michalak
e-mail: Marcin.Michalak@emag.lukasiewicz.gov.pl

Marcin Michalak was born in Poland in 1981. He received his M.Sc. Eng. in computer science from the Silesian University of Technology in 2005 and a Ph.D. degree in 2009 from the same university. His scientific interests cover machine learning, data mining, rough sets, and biclustering. He is an author and co-author of over 90 scientific papers.

https://orcid.org/0000-0001-9979-8208

otrzymano/received: 15.03.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2647

APPLICATION OF DIGITAL IMAGE PROCESSING METHODS FOR OBTAINING CONTOURS OF OBJECTS ON ULTRASOUND IMAGES OF THE HIP JOINT

Pavlo Ratushnyi, Yosyp Bilynsky, Stepan Zhyvotivskyi
Vinnytsia National Technical University, Department of Electronics, Vinnytsia, Ukraine

Abstract. *In this work, the problems of research of ultrasonic images of joints are formulated. It is that for early diagnosis of developmental disorders of the hip joints needs to take frequent pictures, and the least harmful to health is ultrasound. But the quality of such images is not sufficient for high-quality automated measurement of geometric parameters and diagnosis of deviations. The ultrasound image of the hip joint is evaluated by quantifying the exact values of the acetabular angle, the angle of inclination of the cartilaginous lip, and the location of the center of the femoral head. To get these geometric parameters, you need to have clear images of objects. And for the operation of automated computer measurement systems, it is necessary to use such methods of pre-digital image processing, which would give clear contours of objects. Known and available image processing algorithms, in particular contour selection, face problems in processing specific medical images. It is proposed to use the developed method of sharpening to further obtain high-quality contour lines of objects. A mathematical model of the method is presented, which is a formula for converting the intensity values of each pixel of a digital image. As a result of this method, the noise component of the image is reduced, and the intensity differences between the background and the objects are increased, and the width of these differences is one pixel. The algorithm of a sequence of processing of ultrasonic images and features of its application have resulted. The results of the developed set of methods are given. The paper presents the results of processing the real image of the hip joint, which visually confirms the quality of the selection of objects on view.*

Keywords: ultrasound images, diagnostics, hip joint, digital processing, sharpening, contour

ZASTOSOWANIE METOD CYFROWEGO PRZETWARZANIA OBRAZU DO UZYSKIWANIA KONTURÓW OBIEKTÓW NA OBRAZACH ULTRASONOGRAFICZNYCH STAWU BIODROWEGO

Streszczenie. *W pracy sformułowano problematykę badań ultradźwiękowych obrazów stawów. Polega ona na tym, że do wczesnej diagnostyki zaburzeń rozwojowych stawów biodrowych konieczne jest częste wykonywanie zdjęć, a najmniej szkodliwe dla zdrowia są badania USG. Jednak jakość uzyskanych obrazów nie jest wystarczająca do tego, by z wysoką jakością przeprowadzić automatyczne pomiary parametrów geometrycznych i diagnostykę odchyleń. Obraz ultrasonograficzny stawu biodrowego ocenia się poprzez ilościowe określenie dokładnych wartości kąta panewki, kąta nachylenia warstwy chrzęstnej i położenia środka głowy kości udowej. Aby uzyskać te parametry geometryczne, musisz konieczne są wyraźne obrazy obiektów. Do obsługi zautomatyzowanych komputerowych systemów pomiarowych konieczne jest stosowanie takich metod przedcyfrowego przetwarzania obrazu, które dawałyby wyraźne kontury obiektów. Znane i dostępne algorytmy przetwarzania obrazu, w szczególności wybór konturu, napotykają problemy w przetwarzaniu określonych obrazów medycznych. Proponuje się wykorzystanie opracowanej metody wyostrzenia w celu dalszego uzyskania wysokiej jakości konturów obiektów. Przedstawiono model matematyczny metody będący formułą do przeliczania wartości natężenia każdego piksela obrazu cyfrowego. W wyniku tej metody zmniejsza się składowa szumu obrazu, a różnice intensywności między tłem a obiektami są zwiększane, a szerokość tych różnic wynosi jeden piksel. Opracowano algorytm kolejności przetwarzania obrazów ultradźwiękowych i cechy jego zastosowania. Podano wyniki opracowanego zestawu metod. W pracy przedstawiono wyniki przetwarzania rzeczywistego obrazu stawu biodrowego, co wizualnie potwierdza jakość doboru obiektów widzenia.*

Słowa kluczowe: zdjęcia ultrasonograficzne, diagnostyka, staw biodrowy, obróbka cyfrowa, ostrzenie, kontur

Introduction

For early diagnosis of dysplasia and study of the development of the hip joints of children, two main areas of examination are used – X-ray and ultrasound. A state of joint development and possible disorders are assessed by measuring the values of the acetabular angle, the angle of inclination of the cartilaginous lip, and the location of femoral head centre. Various computer applications for automated detection and measurement are used to study these values.

X-ray images are obviously clearer, more detailed and do not have such noise in the image as ultrasound images. But their disadvantage is that the process of obtaining X-ray images is harmful to the body and therefore they cannot be done often. If a child shows signs of impaired development of the hip joints, then, in order to monitor dynamics of the joint, an image should be obtained quite often. Therefore, in such cases the ultrasonic technique of reception of joint images is generally applied. It practically does not harm the child's body.

The disadvantages of such images obtained by ultrasound technique are much worse quality, low definition of objects in the image and the presence of noise in the image. Therefore, for higher measurement accuracy and the possibility of using automated computer measurement systems, pre-processing of digital images is quite relevant – it is noise filtering, sharpening and highlighting clear contours of image objects.

1. Formulation of the problem

The main problem is the poor quality of the contours of ultrasound objects after the contours are selected by traditional methods. Specific processing methods should also be used to process specific medical images.

When processing ultrasound images, it is difficult to obtain the contour by the usual gradient method, because the boundaries of the image objects are quite wide, and the values of the gradient increment at the boundary are commensurate to the amplitude of noise in the image. Therefore, the contour line obtained by this method will look continuous, but will be wide and blurred, which makes it almost impossible to automatically select objects of the image and calculate their geometric parameters. When selecting contours by gradient methods, the contour line, although it will be thin and clear, but will have gaps and false fragments of the contour. This, in turn, will distort the shape of the image objects, which will create ambiguity for their automated parameterization. Methods based on differential approaches give for such images a contour consisting of a large number of fragments across the width of the intensity difference. It also makes it almost impossible to automatically determine objects' geometric parameters of ultrasound images [1, 5, 7].

2. Mathematical models

Optoelectronic systems designed to extract geometric features and measure the geometric parameters of objects have a large number of optical transmission links and optical transmission functions.

The central limit theorem states that the law of distribution of the sum of random variables is indefinitely close to the normal law and has an analogue in the theory of linear filtering: the product of several spatial-frequency characteristics of individual parts of the optical system approaches the Gaussian form when their number becomes large. This circumstance attaches special importance to the Gaussian, which makes it possible to accurately describe the spatial-frequency properties of real optoelectronic devices.

The mathematical notation of a normalized Gaussian has the form of general formula (1) shown below:

$$G(x,y)=\exp\left(-\frac{x^2}{2\sigma^2}\right)\cdot\exp\left(-\frac{y^2}{2\sigma^2}\right) \quad (1)$$

where σ – is RMS distribution of the intensity difference at the boundary (Gaussian law constant), *x*, *y* – border coordinates.

If $\sigma > 0$, then a sharp contrast intensity distribution *H(x, y)* turns into a blurred image. The degree of blur is greater the greater the value σ.

Let the object *H(x, y)* have a two-dimensional distribution of the intensity of the sharp edge (Fig. 1).

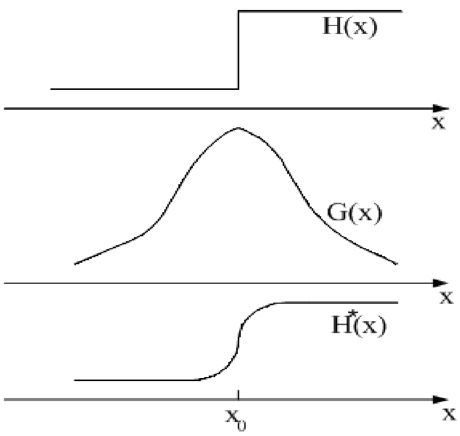


Fig. 1. View of the boundary curve before and after passing through the linear optical system with scattering function G(x)

To restore the shape of the intensity difference at the boundaries of the background and objects to one that is closer to the input, it is proposed to use the developed method of sharpening.

Traditional filters use static masks of weights to pass through the entire image. A feature of the proposed method is the use of a dynamic mask of weights. That is, for each pixel of the image is convoluted with a mask of individual coefficients. The input data for calculating the coefficients is the value of the intensity of the near pixels.

The processing time of the image will obviously increase but having received the image with such form of differences further it is possible to receive accurate continuous contour lines in width in one pixel.

To form a mask of weight coefficients, the following indicators are introduced [5]:

$$k_{wh}=1,\ if\ \ I_{ij}\geq\frac{\sum\limits_{w=i-(n-1)/2}^{i+(n-1)/2}\ \sum\limits_{h=j-(n-1)/2}^{j+(n-1)/2}\left(I_{wh}\right)}{n^2} \quad (2)$$

$$k_{wh}=0,\ if\ \ I_{ij}<\frac{\sum\limits_{w=i-(n-1)/2}^{i+(n-1)/2}\ \sum\limits_{h=j-(n-1)/2}^{j+(n-1)/2}\left(I_{wh}\right)}{n^2} \quad (3)$$

$$z_{wh}=1,\ if\ \ I_{wh}\geq I_{ij} \quad (4)$$

$$z_{wh}=0,\ if\ \ I_{wh}< I_{ij} \quad (5)$$

Using the calculated coefficients, we form a mass of weights. Thus, the general mathematical model of the method [2, 3, 5] described by the system of equations:

$$\begin{cases} I'_{ij}=\dfrac{\sum\limits_{w=i-(n-1)/2}^{i+(n-1)/2}\sum\limits_{h=j-(n-1)/2}^{j+(n-1)/2}\left(I_{wh}\cdot\left(\overline{k_{wh}\oplus z_{wh}}\right)\right)}{\sum\limits_{w=i-(n-1)/2}^{i+(n-1)/2}\sum\limits_{h=j-(n-1)/2}^{j+(n-1)/2}\left(\overline{k_{wh}\oplus z_{wh}}\right)}; \\ k_{wh}=1,\ if\ \ I_{ij}\geq\dfrac{\sum\limits_{w=i-(n-1)/2}^{i+(n-1)/2}\sum\limits_{h=j-(n-1)/2}^{j+(n-1)/2}\left(I_{wh}\right)}{n^2}; \\ k_{wh}=0,\ if\ \ I_{ij}<\dfrac{\sum\limits_{w=i-(n-1)/2}^{i+(n-1)/2}\sum\limits_{h=j-(n-1)/2}^{j+(n-1)/2}\left(I_{wh}\right)}{n^2}; \\ z_{wh}=1,\ if\ \ I_{wh}\geq I_{ij}; \\ z_{wh}=0,\ if\ \ I_{wh}<I_{ij}. \end{cases} \quad (6)$$

where *n* is the size of the mask, *Iij* the input brightness value of the current pixel, *I'ij* the output brightness value of the current pixel.

Figure 2 shows the result of using the proposed method. Here is the shape of the boundary curve of the real ultrasound image, before and after the application of this method.

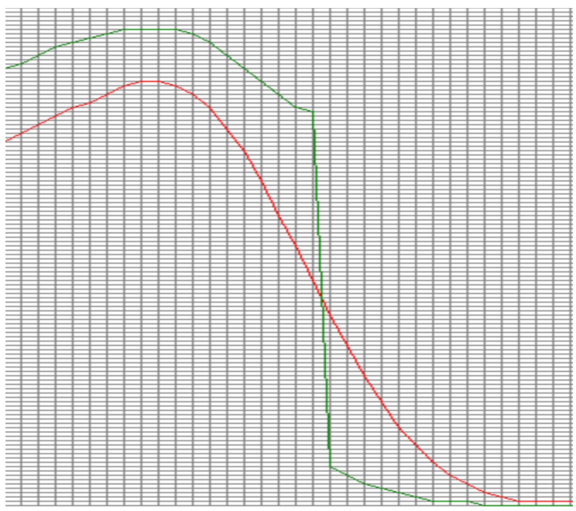


Fig. 2. View of the boundary curve before and after use the proposed method of increasing sharpness

Step of the grid on image is one pixel so as we can see we have one pixel width slope after image processing.

The process of sharpening to a maximum slope at boundary curve involves the following steps [5]:

1) specifying the n×n mask depending on the width of the intensity drops;
2) calculating the average intensity value of pixels in a moving window;
3) comparing the intensity of the central pixel with the average value of the window;
4) comparing intensity values of each pixel in the window with the intensity value of the central pixel in the window;
5) forming the mask of weight coefficients;
6) forming a new mask for the next pixel with weights on the above algorithm.

An edge detection method is proposed basing on finding intersection points [4, 6] of boundary curves of the input image and the image after proposed sharpening.

This will allow you to get contour lines with differentiated intensities depending on the magnitude of the difference, while having a clean background. Therefore, you can easily highlight significant elements of the image and weed out minor ones.

3. Results and discussion

Based on the above, we can conclude that to highlight the contour of the object on the ultrasound image, you must first perform a sharpening. But it is necessary to choose a method of sharpening that would not increase the noise level.

It is proposed to apply the method of sharpening to the maximum slope of the boundary curve [2, 3, 5].

Figure 3 shows part of the ultrasound image of the hip joint (a) and the same image after linear filtration (b).

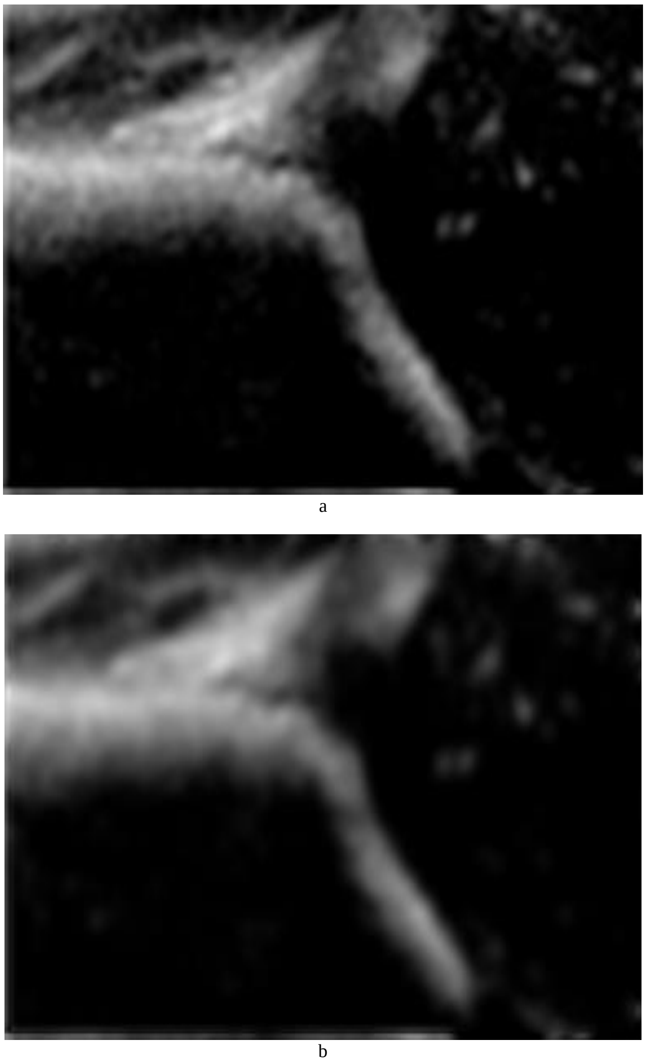


Fig. 3. Part of the ultrasound image of the hip joint: a) the original image, b) after linear filtering

As can be seen from the figures, the original image has a fairly high noise level, so the first stage of processing is linear noise filtering. But after filtering, the already blurred contours become even more blurred.

However, the use of the proposed method [2, 3] of sharpening to the maximum slope of the boundary curve gives good results. Figure 4 (a) shows the result of applying the proposed method of sharpening.

As you sharpen, the inner parts of the image become smoother with less noise, and the background becomes smoother with less noise. Instead, the transition between the background and the object became clear, given the transition of 1 pixel width and the height of the transition, which corresponds to the average difference between the background and the object. Now the contours of objects obtained from such image will look much better visually and will be much more suitable for automated measurement of objects' geometric parameters.

Figure 4b shows the image of the contours obtained by finding intersection points of boundary curves of the images [4, 6].

In this case, filtered and sharpened images (Fig. 3b and Fig. 4a, respectively). Moreover, the pixel intensity at the point of intersection is calculated as the difference between the pixel intensity at the edge of the image difference from high sharpness. This approach makes it possible to obtain thin contour lines and differentiate their intensities.

Figure 4b shows that intensity of the contour lines of studied objects is higher than intensity of side elements contours of the image. This in turn makes it possible to remove unnecessary contour lines if necessary, for the use of automated systems for measuring objects' geometric parameters.

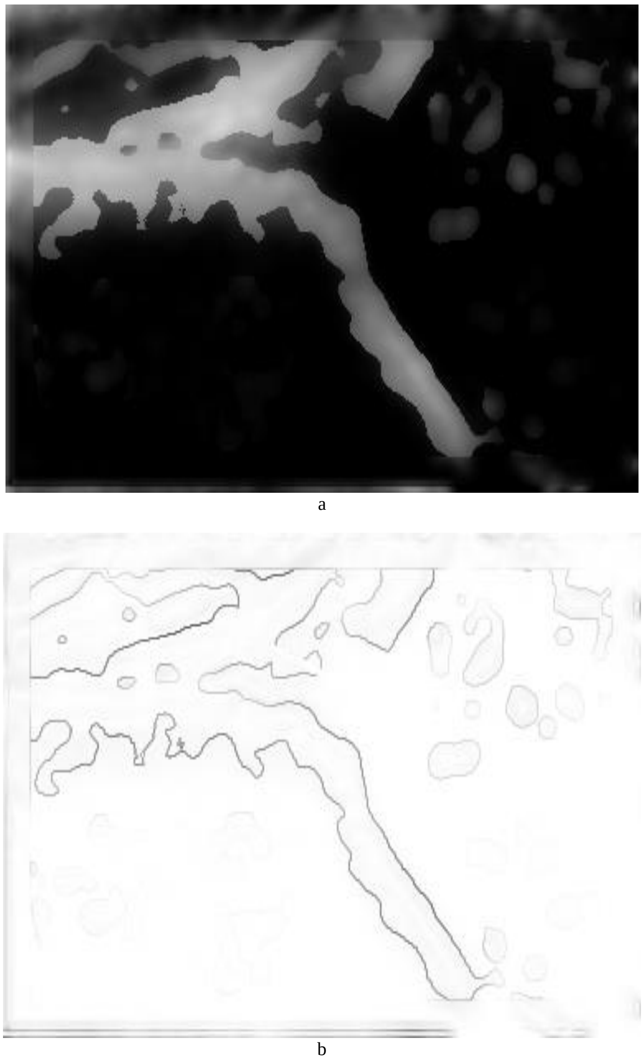


Fig. 4. Part of the ultrasound image of the hip joint: a) after sharpening by the proposed method, b) after the selection of contours by finding the intersection points of the boundary curves

When performing the step of sharpening with the proposed method, a very important aspect is the selection of the mask size for image processing. In the input image, the width of the intensity differences between the background and the objects can be dozens of pixels, depending on the image resolution. The size of the mask should be chosen as the average value of the intensity differences width [8]. Then the processing results will be the best. If you select a mask smaller than required, the original image will be more granular, and if you select a mask larger than required – on the contrary, the curves of the contours will have larger radiuses.

This was confirmed experimentally. For this purpose, the synthesized sample image with clear and sharp differences in intensity was subjected to noise and blur. As a result, the average width of the differences was about 16–20 pixels. It was then reproduced using the proposed sharpening method. The initial and reproduced images were compared using the standard deviation criterion.

A graph of the dependence of the standard deviation on the size of the mask is obtained (Fig. 5).

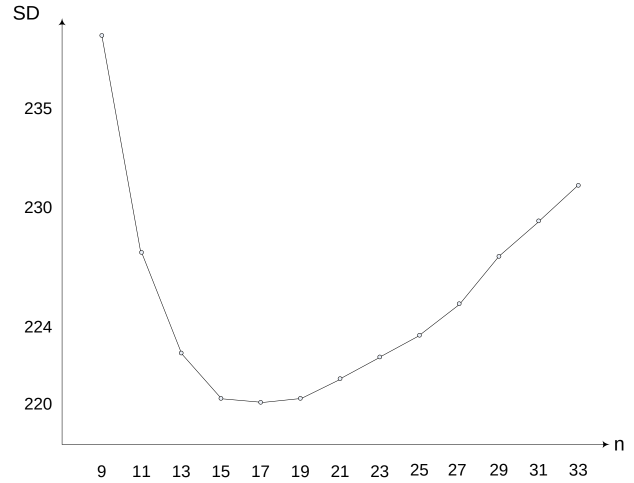


Fig. 5. Dependence of the standard deviation on the size of the mask

The graph shows that the best results are obtained when using a mask with a size close to the average width of the differences [8].

Figure 6 shows the sequence of stages of processing the input image.

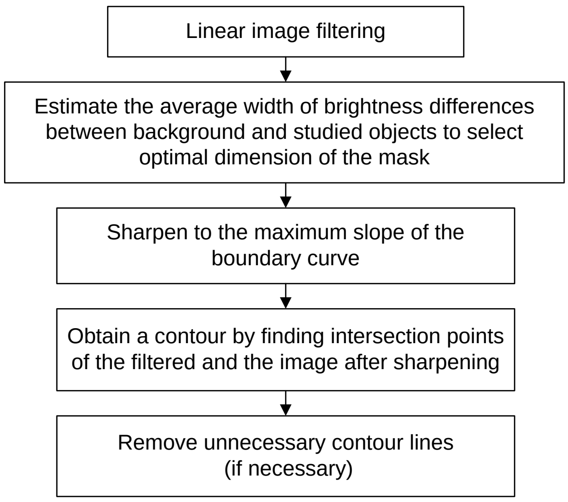


Fig. 6. The sequence of stages of processing the input ultrasound image

4. Summary

Therefore, the paper proposes a way to solve the problem of isolating high-quality contours of objects of the hip joint ultrasound images. The use of a sequence of image processing methods is proposed. A mathematical model of sharpening digital images is presented. The choice of the mask dimension is substantiated. The results of the proposed methods on real ultrasound images are presented.

Consistent use of the proposed methods of ultrasound image processing allows you to clearly identify image objects, get their qualitative contours, clear, narrow and continuous. This, in turn, significantly increases the efficiency of automated computer tools for measuring objects geometric parameters, that speeds up the diagnostic process and allows to more accurately track changes dynamics in the of objects' geometry, in this case – the geometry of the hip joint.

References

[1] Bilynsky Y., Horodetska O., Ratushny P.: Prospects for the use of new methods of digital processing of medical images. 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science TCSET, 2016, 780–783 [http://doi.org/10.1109/TCSET.2016.7452182].

[2] Bilynsky Y., Ratushny P., Klimenko I.: The method of sharpening low-contrast two-dimensional images. Bulletin of the Polytechnic Institute 6/2009, 12–15.

[3] Bilynsky Y., Ratushny P., Melnichuk A.: The method for improving image sharpness. Applicant and patentee Vinnytsa National Technical University 200907326, Pat. 45887, Ukraine, G 06 K 9/36. applic. 13.07.09; publ. 25.11.09, bull. 22.

[4] Bilynsky Y., Yukysh S., Ratushny P.: Edge detection detector based on low-pass filtering. Bulletin of Khmelnytsky National University 1/2009, 230–223.

[5] Bilynsky Y.Y. et al.: Contouring of microcapillary images based on sharpening to one pixel of boundary curves. Proc. SPIE 10445, 2017, 104450Y [http://doi.org/10.1117/12.2281005].

[6] Bilynsky Y.Y. et al.: Controlling geometric dimensions of small-size complex-shaped objects. Proc. SPIE 10445, 2017, 104450I [http://doi.org/10.1117/12.2280899].

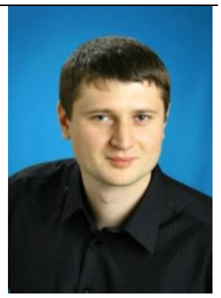
[7] Nikolskyy A. I. et al.: Using LabView for real-time monitoring and tracking of multiple biological objects. Proc. SPIE 10170, 2017, 101703H [http://doi.org/10.1117/12.2261424].

[8] Ratushny P. M. et al.: Research of the mask size for the method of increasing the sharpness to the maximum slope of the boundary curve. Measuring and computing equipment in technological processes. Khmelnytskyi 3/2014, 71–74.

Ph.D. Pavlo Ratushnyi
email: ratushnyj.p.m@vntu.edu.ua

Associate Professor of Electronics and Nanosystems at Vinnytsia National Technical University. In 2011 defended his dissertation on "Methods and system of low-contrast image processing for the evaluation of microcapillaries of human limbs" in the specialty "Biological and medical devices and systems". The main scientific direction is computer processing of biological and medical images for research of geometrical parameters of objects.

http://orcid.org/0000-0002-3365-8699



Ph.D. Yosyp Bilynskyi
e-mail: yosyp.bilynsky@gmail.com

Associate Professor of Electronics and Nanosystems at Vinnytsia National Technical University. He defended his doctoral dissertation in 2009. He has more than 250 scientific works, including 65 in professional publications, 65 patents, 4 monographs, 10 educational and methodological.

http://orcid.org/0000-0002-9659-7221

M.Sc. Stepan Zhyvotivskyi
e-mail: zhyvotivskyi.s@gmail.com

Postgraduate student of Vinnytsia National Technical University, Faculty of Information Technology and Computer Engineering, majoring in metrology.

http://orcid.org/0000-0002-2379-0645

otrzymano/received: 28.04.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2655

OVERVIEW OF THE USE OF X-RAY EQUIPMENT IN ELECTRONICS QUALITY TESTS

Magdalena Michalska
Lublin University of Technology, Department of Electronics and Information Technology, Lublin, Poland

Abstract. *Surface-mount technology is now widely used in the production of many components. The development of the miniaturised electronics industry forces the development of increasingly accurate inspection methods. X-ray and computed tomography are methods to accurately assess the quality of a circuit board. The article discusses the basics of image formation of the tested electronics, the development of the design of the devices used and examples of x-ray, computed tomography applications.*

Keywords: defect inspection, electronic components, industrial radioscopic systems, X-ray, computed tomography

PRZEGLĄD WYKORZYSTANIA URZĄDZEŃ RENTGENOWSKICH W BADANIACH JAKOŚCI ELEKTRONIKI

Streszczenie. *Obecnie technologia montażu powierzchniowego jest szeroko stosowana w produkcji wielu podzespołów. Rozwój zminiaturyzowanego przemysłu elektronicznego wymusza rozwój coraz to dokładniejszych metod inspekcji. Metodami pozwalającymi w dokładny sposób ocenić jakość płytki drukowanej jest wykorzystanie promieniowania rentgenowskiego i tomografii komputerowej. W artykule omówiono podstawy powstawania obrazów badanej elektroniki, rozwój konstrukcji wykorzystywanych urządzeń, przykłady zastosowań RTG i tomografii komputerowej.*

Słowa kluczowe: kontrola defektów, komponenty elektroniczne, przemysłowe systemy radioskopowe, promieniowanie rentgenowskie, tomografia komputerowa

Introduction

The emergence of surface assembly and miniaturization of electronics made it possible to develop smaller mobile devices [3]. Assessing the quality of manufactured components and electronics assemblies in the electronics industry encounters many difficulties. There are many problems with the quality of the components in the components, the production process and the selection of the [14].

The following components are assessed on the surface of the circuit boards: their soldering method, the size of the mounting holes, the presence of paths, mechanical damage [1]. The presence of many solders on the surface of the tiles requires automation of the entire evaluation process. Automated devices are created for this purpose. These include automatic optical inspection [4, 5], laser head measurements, X-ray tubes and computed tomography. The technical requirements set at different stages of production challenge research laboratories [9]. During non-destructive testing, electronics are exposed to extremely high or low temperatures, humidity, vibrations, impacts.

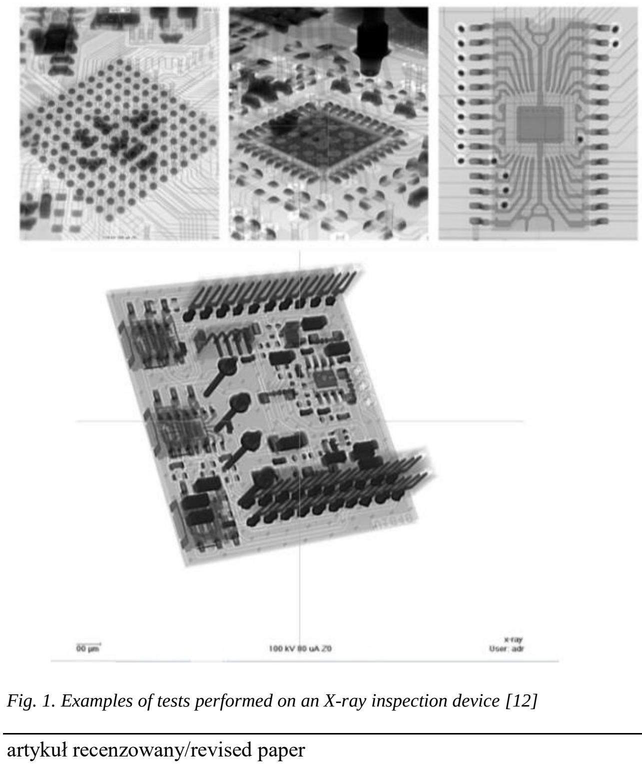


Fig. 1. Examples of tests performed on an X-ray inspection device [12]

The most common defects on the printed surface of the board include excess soldering, displacement of the geometry of small elements, tilt of the elements relative to the board, soldering porosity, excess soldering, unwanted splashes. Fig. 1 shows examples of images of SMD tiles tested. Increasingly, applications are using 3D visualization so that opera-track can better assess the quality of the manufactured product.

1. Basics of radiology inspection

X-rays are a type of electromagnetic radiation with a wavelength from a few pm to 10 nm [6]. Electromagnetic radiation is distinguished by an oscillating electrical and magnetic field in the system, shifted relative to each other at an angle of 90 degrees (Fig. 2). X radiation is characterized by a certain wavelength and frequency.

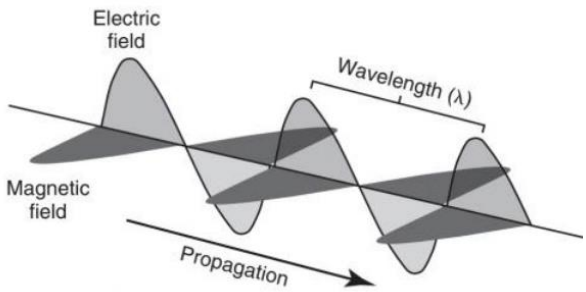


Fig. 2. X-rays propagation [28]

X-ray radiation is obtained by an X-ray tube and electrons quickly hitting a metal element. It is built of glass vacuum banks, in which they are connected to a high-voltage source of cathode and anode. The vacuum produced prevents electrons from interacting with air molecules. So they move very quickly [6]. The cathode, usually made of tungsten fiber, emits electrons. Electrons are accelerated in the electric field and emit X-rays when they collide with the anode. X-ray tube operation requires anode cooling [11, 26].

X-ray check-up scans allow accurate monitoring of several platelet quality criteria (PCB). Unlike X-ray inspection scans, CT scans provide complete 3D information [23, 29]. X-ray inspection is possible even without sufficient lighting. The resulting tomography sections allow the reconstruction of images in two and three dimensions. As part of ct software development, automatic 3D reconstruction is becoming faster and more efficient [8, 7, 20].

X radiation can penetrate various materials. Their density determines the color in the resulting image and allows you to distinguish the materials from each other. [17, 18]. Fig. 3 shows the differences in absorption, photon transmission for different materials. The higher the density of the material, the lower its apenetration. High dose absorption gives a smaller number of transmitted photons. The opposite occurs for materials with low absorption efficiency.

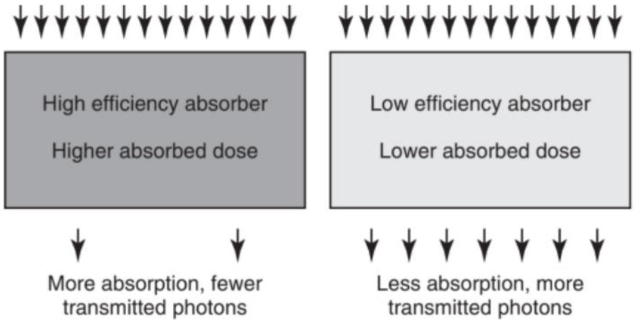


Fig. 3. Comparison of two materials with different properties very significant for X-ray imaging [28]

2. Solutions development for industrial radioscopic systems

In [21] x-ray inspection is used for technological processes in 2 and 3 dimensions, due to the use of projections from multiple projections of the test element. An important aspect in X-ray examinations is the amount of radiation dose that the tested electronics take. To increase the measurement accuracy of many manufacturers increases the amount of radiation energy. There are elements sensitive to it.

In [10] results obtained through X-ray and CT inspections. The authors paid particular attention to the difference in the resulting image received. Radioscopy allows inspection in two dimensions, computed tomography in addition to measuring objects in 2D gives the possibility of its full visualization in 3D. Also important is the process of reconstructing images, where a lot of time is devoted to developing the best algorithms to prepare a model of the projected element in 3D. Real-time 3D computed tomography imaging is possible with pulsed multipixel sources. The use of small CNT FE cathode and fast switching gives new measurement capabilities [24].

3. Industry systems based on X-ray

Automating the process makes it a reliable method to monitor the prototyping process, implement serial production and quality control at different stages. The emergence of many systems based on X-rays allowed for inspection of many components of electronics. During the production of electronics, many defects related to solder connections are created. X-rays are used to determine their quality.

The use of image processing algorithms obtained through X-rays is becoming an important part-time. The difficulty lies in segmenting the solder connections and the spaces in them. To this end, a weave network used to classify images after segmentation was created in [30].

In order for X-ray inspection to be an even more perfect way of imaging, further work is needed to improve it [21]. X-ray sensors were improved to provide better contrast and resolution to the resulting images. It is also important to continuously improve the algorithms responsible for transforming the images received. Machine learning methods, process speed and efficiency, hierarchical control are taken care of.

In [31] investigated how empty spatial solder connections affect its quality. The size, place and frequency of defects were taken into account. The SMT X-ray radiography control process is proposed in [25]. Table 1 provides a comparison of inspections of the main types of defects on SMT boards concerning defects in brazed connectors. In [21] defects such as: solder bridges, open connects, lifted lead, aignment between pad and lead, solder thickness and others were observed using visual inspection and X-rays.

Table 1. Comparison of defects occurring on brazed connectors. Markings used: ++ VERY RELIABLE , + RELIABLE, o FAIRLY RELIABLE [21]

Features	X-Ray Inspection	Visual Inspection
Solder bridges	++	+
Open connects/lifted lead	+	+
Alignment between pad and lead	+	o
Solder thickness/volume	++	–
Shape of heel fillet	+	–
Shape of toe/center fillet	+	o
Wetting angles	+	o
Cold or disturbed joint	–	+
Dull solder	–	+
Porosity and voids	++	–
Solder balls	+	o

In [16] X-ray images (Fig 4a) were analysed. To this end, an algorithm based on contour ball joints. classicization was created. Scientists have developed a radar greyhound. If the result of the analysis falls outside a specified circle (Fig. 4b), it is considered bad. Figure 4c was considered to be the correct level of 10, the higher results were considered to be reshaped BGA. In Fig. 3a in the yellow circle marked objects that did not pass the quality control process.

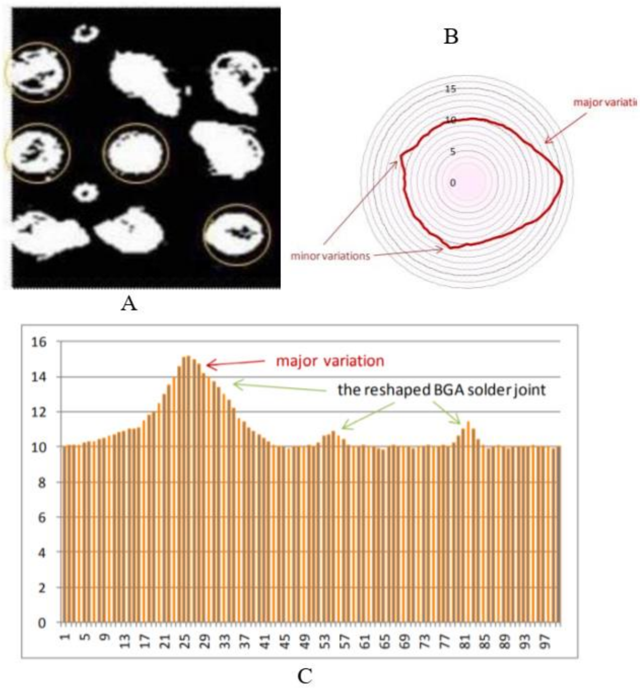


Fig. 4. Image classification x-ray ball joints: A – original images, in yellow circles element classified as vicious balls; B – radar chart; C – contour plot [16]

In [32] the quality of chips, chip washers was checked for soldering process. The development of x-ray lamp and detector technology enables work on nanofocus and multifocus X-ray systems. Developed software, image processing algorithms, increased stability, position constancy allows the use of X-rays for electronics control [19].

4. Industry systems based on CT

The use of computed tomography in the electrical industry has greatly accelerated the way quality control and individual tiles. Many methods based on X-ray systems have been used in non-destructive test (NDT) inspection and testing processes.

In [13] uses a quantitative, three-dimensional method of analysis of soldering properties and defects by CT, the use of which gives a number of new methodological possibilities In a precise and efficient way, a quality control process for PCB components can be ensured. The process allows for the catching of defective components and an increase in the quality of manufactured assemblies or components ingested in electrical equipment and systems [15].

In [27] oblique CT allows you to examine the quality of the shape of the solder tumor. Due to the limitations of the mechanical structure and the speed of performing scans in [27] a different approach to oblique CT and 3D reconstruction is presented. The use of an open X-ray generator and the FBP 3D method (Filtered Reverse Projection) yielded very good results. Projection from different directions is possible using a rotary flat panel detector transfer. Solutions are also being created based on Planar Tomography Computer (PCT) [21]. It enables much faster reconstruction based on the linear movements of the test object.

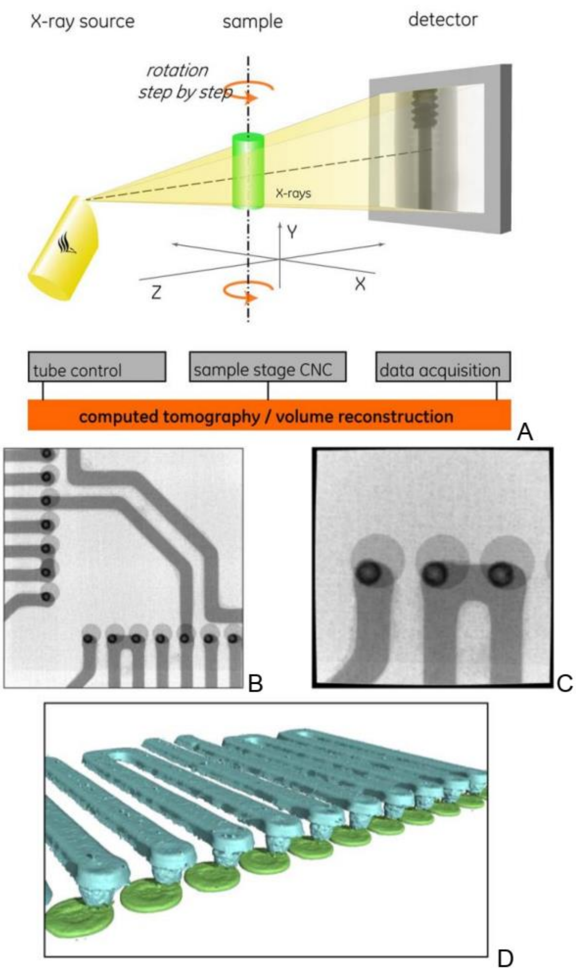


Fig. 5. A – functional and structural diagram CT; B – X-ray images (1 µm/pixel); C – X-ray image (0.4 µm/pixel); D – Visualization of 3D CT, voxel 1.2 µm, through diameter about 50 µm [26]

Researchers are working on different focal length x-ray lamps. Higher resolution of the resulting image can be achieved by modifying the focal length of the beam. Among these modifications, microfocus should be calculated [23, 29]. Nanofocus and microfocus CT are increasingly used in the study of miniature electronics components. In [26] nanofocus images have been obtained using X-ray microscope and CT. The system in figure 5 provides CT reconstruction software and region-of-interest-CT (ROI-CT). The use of x-ray image enhancers in 2D allows you to achieve much better image quality. For CT, however, image quality has been improved with low noise and a DXR detector.

In [2] compared images obtained from full CT and off-lin PCT. Both tools allow for 3D rendering, full CT is better suited for Z and X, Y 2D planar view image quality. In PCT 2D images are done at an angle in advance, the data set is much more limited, reconstruction does not give the full result. In full µCT we have a different shape. Figure 6B shows a view from the software window, which can measure the thickness of the soldering.

Another good solution is the FE source, which guarantees control over the position of the beam emitter at the nanofocus [26]. Individually addressed electron beams allow you to control the shape of the X-ray beam. Multipixel sources [8, 29] used in CT have a number of advantages. These include: higher spatial resolution and subsequent formation of the radiation beam. The images shown in Figure 7 were created by CT based on an electron control system and x-ray beam shape.

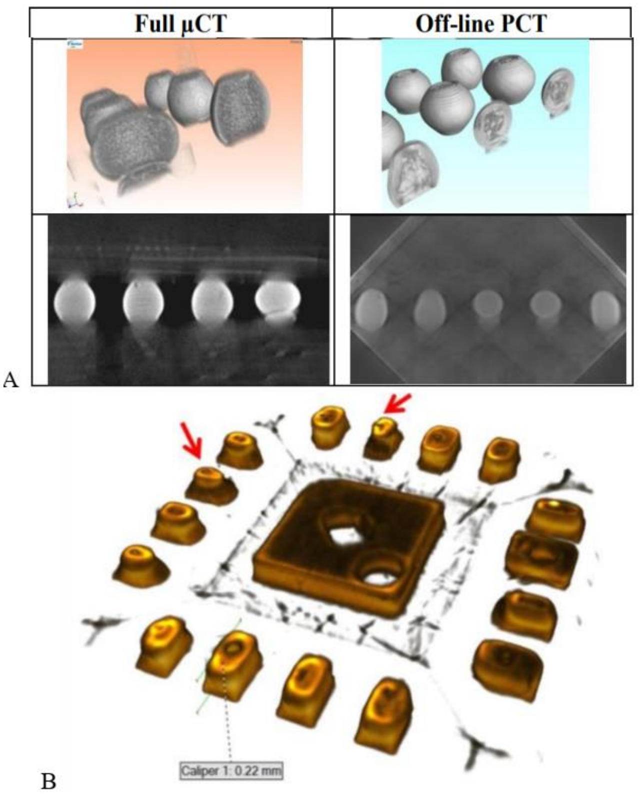


Fig. 6. A – Comparison of images obtained using full µCT and offline PCT methods; B – PCT 3D model with marked missing connections (red arrows) [14]

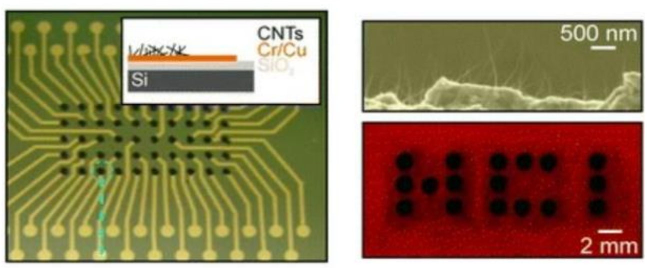


Fig. 7. Comparison of two materials with different properties very significant for X-ray imaging [23, 29]

5. Summary

The use of surface mounted devices (SMD) creates the development of imaging technologies during the production process. The methods of imaging the soldering joint of the substrate, assessed in many works, showed that the proposed systems allowed satisfactorily obtaining information on which components are good and which are defective. Of course, the results of this work were very diverse. The measurement systems developed by scientists have found wide application in implementations in factories around the world.

References

[1] Ahi K., Asadizanjani N., Shahbazmohamadi S., Tehranipoor M, Anwar M.: Terahertz characterization of electronic components and comparison of terahertz imaging with x-ray imaging techniques. Proc. SPIE 9483, 94830K [https://doi.org/10.1117/12.2183128].

[2] Bernard D., Golubovic D., Krastev E.: 3D board level x-ray inspection via limited angle computer tomography. Proceedings of the SMTA International Conference, Orlando, Florida, 2012.

[3] Celik T., Tjahjadi T.: Contextual and variational contrast enhancement. IEEE Transactions on Image Processing 20(12), 2011, 3431–3441.

[4] Chang K. H.: Development of optical inspection system for surface mount device light emitting diodes. Master thesis, National Sun Yat-sen University, Taiwan, 2012.

[5] Chang W., Su C., Guo D.: Automated optical inspection for the runout tolerance of circular saw blades. Int. J. Adv. Manuf. Technol. 66, 2013, 565–582.

[6] Doyle J. F.: Wave propagation in structures. Springer Verlag, New York Inc. 1997.

[7] Gao B., Yue G. Z., Qiu Q., Cheng Y., Shimoda H., Fleming L, Zhou O.: Fabrication and electron field emission properties of carbon nanotube films by electrophoretic deposition. Adv. Mater. 13, 2001, 1770–1773.

[8] Guzilov I. A., Kuzmich K. V., Maslennikov O. Y., Smirnova E. V., Minakov P. V., Poroykov A. Y., Rakhimov A. T., Seleznev B., Sen V. V.: Multi beam X-ray tube with the field emitter on the base of nanocrystalline graphite for computer tomography. IEEE International Vacuum Electronics Conference IVEC '09, 2009, 289–291.

[9] Hanke R., Fuchs T., Salamon M., Zabler S.: X-ray microtomography for materials characterization. Editors: G. Huebschen, I. Altpeter, R. Tschuncky, H.-G. Herrmann: Materials Characterization Using Nondestructive Evaluation (NDE) Methods. Woodhead Publishing 2016, 45–79

[10] Hanke R., Fuchs T., Uhlmann N.: X-ray based methods for non-destructive testing and material characterization, Nuclear Instruments and Methods. Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 591(1), 2008, 14–18.

[11] Hendricks R.: On-line inspection enables 6 sigma quality. Circuits Assembly 12, 1990, 23–27.

[12] https://www.eae-elektronik.pl/inspekcja-rtg/ (20.04.2021).

[13] Huang R., Sorini A., McNulty J.: Quantitative solder inspection with computed tomography. 2014 IEEE Symposium on Product Compliance Engineering (ISPCE), San Jose, USA, 2014, 82–85, [http://doi.org/10.1109/ISPCE.2014.6842006].

[14] Juha M.: Micro-focus x-ray imaging. Editors: J. S. Heyman: Electronics Reliability and Measurement Technology. William Andrew Publishing, 1988, 60–67.

[15] Kruse R. J., Bossi R. H.: X-Ray Tomographic Techniques for Inspection of Electronic Components. Editors: D. O. Thompson, D. E. Chimenti: Review of Progress in Quantitative Nondestructive Evaluation. Springer, Boston, MA, 1991 [https://doi.org/10.1007/978-1-4615-3742-7_5].

[16] Laghari M. S., Memon Q. A.: Identification of Faulty BGA Solder Joints in X-Ray Images. International Journal of Future Computer and Communication 4(2), 2015, 122–125.

[17] Liu Z, Yang G, Lee Y. Z., Bordelon D., Lu J., Zhou O.: Carbon nanotube based microfocus field emission x-ray source for microcomputed tomography. Appl. Phys. Lett. 89, 2006, 1–3.

[18] Madsen J.: Focal spot size measurements for microfocus X-ray sets. NDT Int 22, 1989, 292–296.

[19] Maur F.: X-ray inspection for electronic packaging latest developments. Fifth International Conference on Electronic Packaging Technology Proceedings ICEPT2003, Shanghai, China, 2003, 235–239 [http://doi.org/10.1109/EPTC.2003.1298731].

[20] Neubauer C., Hanke R.: Improving X-ray inspection of printed circuit boards by integration of neural network classifiers. Proc. Int. Electron. Manufact. Technol. Symp. 1993, 14–18.

[21] Neubauer C., Schropfer S., Hanke R.: X-ray inspection of solder joints by planar computer tomography (PCT). Proceedings of 16th IEEE/CPMT International Electronic Manufacturing Technology Symposium, La Jolla, CA, USA, 1, 1994, 60–64 [http://doi.org/10.1109/IEMT.1994.404691].

[22] Neubauer C.: Intelligent X-ray inspection for quality control of solder joints. IEEE Transactions on Components, Packaging, and Manufacturing Technology: C 20(2), 1997, 111–120 [http://doi.org/10.1109/3476.622881].

[23] Parmee R. J., Collins C. M., Milne W. I.: X-ray generation using carbon nanotubes. Nano Convergence 2(1), 2015 [https://doi.org/10.1186/s40580-014-0034-2].

[24] Qian X., Tucker A., Gidcumb E., Shan J., Yang G., Calderon-Colon X., Sultana S., Lu J., Zhou O., Spronk D., Sprenger F., Zhang F., Kennedy D., Farbizio T., Jing Z.: High resolution stationary digital breast tomosynthesis using distributed carbon nanotube x-ray source array. Med. Phys. 39, 2012, 2090–2099.

[25] Rooks S., Sack T.: Xray Inspection of Flip Chip Attach Using Digital Tomosynthesis. Circuit World 21(3), 1995, 51–55 [https://doi.org/10.1108/eb044036].

[26] Roth H., Neubrand T., Mayer T.: Improved inspection of miniaturised interconnections by digital X-ray inspection and computed tomography. 12th Electronics Packaging Technology Conference, Singapore 2010, 441–444 [http://doi.org/10.1109/EPTC.2010.5702680].

[27] Teramoto A., Yamada M., Murakoshi T., Tsuzaka M., Fujita H.: High Speed Oblique CT System for Solder Bump Inspection. IECON 2007 – 33rd Annual Conference of the IEEE Industrial Electronics Society, Taipei, Taiwan, 2007, 2689–2693 [http://doi.org/10.1109/IECON.2007.4460065].

[28] Thrall D. E.: Textbook of Veterinary Diagnostic Radiology. Elsevier Health Sciences, 2017.

[29] Wang S., Calderon X., Peng R., Schreiber E. C., Zhou O., Chang S.: A carbon nanotube field emission multipixel X-ray array source for microradiotherapy application. Appl. Phys. Lett. 98, 2011, 213701–213703.

[30] Wankerl H., Stern M. L., Altieri-Weimar P., Al-Baddai S., Kurt-Jürgen L., Roider F., Lang E. W.: Fully convolutional networks for void segmentation in X-ray images of solder joints. Journal of Manufacturing Processes 57, 2020, 762–767.

[31] Yunus M., Srihari K., Pitarresi J. M., Primavera A.: Effect of voids on the reliability of BGA/CSP solder joints. Microelectronics Reliability 43(12), 2003, 2077–2086

[32] Zhang S., De Baets J., Van Calster A.: A new approach to flip chip on board technology using SMT compatible processes. Microelectronics International 16(3), 1999, 39–42.

M.Sc. Magdalena Michalska
e-mail: magdalena.michalska@pollub.edu.pl

Ph.D. student at Department of Electronics and Information Technology, Lublin University of Technology. Recent graduate Warsaw University of Technology. Her research field covers medical image processing, 3D modelling, optoelectronics, spectrophotometry. Author of more than 10 publications.

http://orcid.org/0000-0002-0874-3285

otrzymano/received: 6.05.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2656

SIMULATION AND EXPERIMENTAL RESEARCH OF CLAW POLE MACHINE WITH A HYBRID EXCITATION AND LAMINATED ROTOR CORE

Marcin Wardach, Paweł Prajzendanc, Kamil Cierzniewski, Michał Cichowicz, Szymon Pacholski, Mikołaj Wiszniewski, Krzysztof Baradziej, Szymon Osipowicz
West Pomeranian University of Technology in Szczecin, Faculty of Electrical Engineering, Department of Electrical Machines and Drives, Szczecin, Poland

Abstract. *This paper presents the design and research results of a claw pole machine with hybrid excitation. This machine is excited by permanent magnets and an electromagnetic coil. Both excitation sources are located in the rotor of the machine. Additionally, the rotor is made of a laminated core. This approach facilitates the process of its construction and enables the implementation of even very complicated structure of the rotor, which would be difficult in case of making the rotor from a one piece of material. This paper presents the construction as well as the results of simulation and experimental tests of the machine prototype. The tests showed that the proposed machine has the ability to adjust the voltage in a wide range. Such as a feature could be used, for example, to increase the speed of motor operation in case of an electric vehicle application, but also to regulate the voltage in wind turbines which generators operate at varying rotor speeds resulting from changing wind speed.*

Keywords: wind energy, generators, permanent magnet machines, finite element analysis

BADANIA SYMULACYJNE I EKSPERYMENTALNE MASZYNY KŁOWEJ ZE WZBUDZENIEM HYBRYDOWYM I PAKIETOWANYM RDZENIEM WIRNIKA

Streszczenie. *W pracy przedstawiono konstrukcję oraz wyniki badań maszyny kłowej ze wzbudzeniem hybrydowym. Maszyna ta jest wzbudzana od magnesów trwałych oraz cewki elektromagnetycznej. Obydwa źródła wzbudzenia znajdują się w wirniku. Ponadto wirnik ten zbudowany jest z pakietowanego rdzenia. Takie podejście ułatwia proces jego budowy i umożliwia implementację nawet bardzo skomplikowanej struktury wirnika, co byłoby utrudnione w przypadku jego wykonania z jednej bryły materiału. W niniejszym artykule przedstawiono budowę oraz wyniki badań symulacyjnych i eksperymentalnych wykonanego prototypu maszyny kłowej ze wzbudzeniem hybrydowym. Badania pokazały, że zaproponowana konstrukcja posiada możliwość regulacji napięcia w szerokim zakresie. Taką cechę można wykorzystać zarówno do zwiększenia zakresu prędkości obrotowej przy pracy silnikowej w przypadku zastosowania w pojeździe elektrycznym, jak i do regulacji napięcia w turbinach wiatrowych, których generatory pracują przy zmiennych prędkościach obrotowych wirnika, wynikających ze zmieniającej się prędkości wiatru.*

Słowa kluczowe: energia wiatrowa, generatory, maszyny z magnesami trwałymi, metoda elementów skończonych

Introduction

One of the best known machines with a claw design is a three-phase generator, i.e. an alternator. The alternator is mainly responsible for providing the right amount of electric energy, which is used to power the electric devices in cars. The alternator is mostly driven from the engine crankshaft by means of a belt transmission, therefore it must be properly adapted to work in a wide range of shaft revolutions. Such as an adaptation effectively ensures adequate charging of the battery, but also enables the use of all energy receivers located in the vehicle. And in today's vehicles, such as devices have a high demand for electricity and, at the same time, are not always resistant to any fluctuations in the supply voltage.

It is also necessary that alternator has high efficiency, thanks to which all of the components in the car will be able to work properly. In order for the alternator to work properly with the automotive electrical system, the generated voltage must be properly rectified. The alternator itself should be designed to be durable and also maintenance-free. Any slightest defect could result in a power failure and unwanted immobilization of the vehicle.

The paper presents research on a claw pole machine, which can also be used as an alternator in cars, but besides, it could be a generator of a low-power wind turbine or an electric vehicle drive motor.

1. Machines with hybrid excitations

Hybrid excited motors and generators are machines that combine the features of two machines in their design. It means a machine with permanent magnets and machines with excitation windings. Due to this design, these machines have much better properties. Synchronous machines with hybrid excitation are characterized by having two sources of excitation [5, 6, 9, 10]. The first is the excitation that comes from permanent magnets. The second one comes from the use of additional windings powered by direct current.

Considering the machine in the context of operation in the motor regime, using the pulse width modulation (PWM) method, it is possible to lower the average value of the motor supply voltage, and thus reduce the rotational speed of its rotor. However, this method cannot increase the motor speed above the idle speed. The maximum rotational speed is determined by the voltage induced in the armature, which is proportional to the product of the flux and the angular velocity. Being able to weaken the field it will be possible to influence on the flux, which also decreases. Furthermore, the value of induced voltage in armature decreases also to the state before increase. When the flux changes, the torque changes also. The changed torque is equal to the product of the associated flux and the armature current. During flux changes, the product of torque and angular velocity is the same all the time and does not change. Because of that we have constant power on the shaft.

Excitation with strong permanent magnets causes difficulties in obtaining a rotational speed higher than the rated speed. Obtaining a speed above, the rated speed is possible with increasing the supply voltage or weakening the field. The second method is used in motors with a sinusoidal field distribution with magnets embedded in the rotor. The field weakening is obtained at high armature currents and causes a large reduction in the efficiency of the machine. This loses one of the main advantages of these machines. Weakening the field in this way will create additional potential problems. In the event of a disturbance, in the operation of the inverter, large currents flow, which may cause a serious failure of the drive system.

Of course there are also other methods of adjusting the rotational speed. Alternatively, the supply voltage can be increased to increase the speed or apply the *d*-axis current strategy [2]. However, such methods will make the power electronic system powering the motor more complicated.

2. Topologies of claw pole machines

Many constructions with claw pole topology are known. With some of them, there may be some limitations; the lack of a controlled excitation flux or requiring special areas where the leakage flux is limited. Structures of this type can be used, for example, in electric car drives and in wind farms as generators.

In machines with a claw construction, the excitation winding is located between the centers of the rotor. In this machine we have a standard excitation coming from the rotor windings. Additionally, the permanent magnets are also used. With this permanent magnet structure, there are two main ways to arrange the permanent magnets. The first method (Fig. 1) gives the possibility of placing permanent magnets on the claw poles [11], another possibility allows to place the magnets between the rotor poles (Fig. 2) [7]. Such as a solution necessitates the use of additional solutions, magnetic barriers in the area of the rotor, which will reduce the flux leakage, because a part of the magnetic flux in this structure could close inside the rotor itself, without passing into the air gap.

Another solution (Fig. 3) of the claw pole machine may be the concept where the claws are placed on the stator, while the permanent magnets are mounted on the rotor surface [3]. However, such as a structure does not have flux excitation regulation. In addition, it requires the use of the technique of sintering magnetic composites.

Another possible solution (Fig. 4) is to place the magnets on the outer rotor and the claw poles on the inner stator [4].

There are also constructions of a claw pole machine with hybrid excitation in which the permanent magnets are located in a toroidal form and are located on separate cores (Fig. 5) [8]. One of the more serious disadvantages of this construction of machines is the complex structure which makes it difficult to easily and quickly build such a machine. The advantage of this design is the absence of slip rings and brushes as the excitation coils are stationary. Unfortunately, such a solution forces the excitation flux to pass from the coils through two air gaps, which can limit the control possibilities.

The paper [1] presents the concept of a claw pole machine with an external stator and a stationary excitation control system (Fig. 6). However, the rotor is made of claw poles and permanent magnets attached in their vicinity. In this case, the lack of brushes and rings is advantageous, but it is also a structurally very complicated solution. Moreover, the magnetic flux from the excitation control system must also cross two air gaps, which increases the flux leakage from the stationary excitation coil.

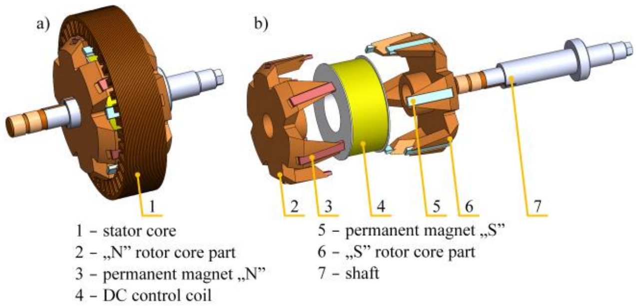


Fig. 1. Hybrid excited claw pole machine with permanent magnets mounted on claws [11]: a) stator and rotor, b) rotor in exploded view

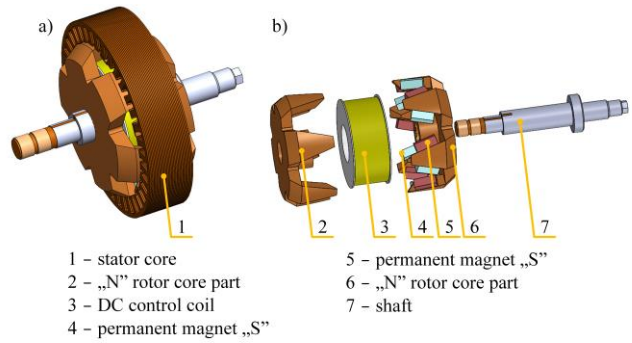


Fig. 2. Hybrid excited claw pole machine with permanent magnets mounted between claws [7]: a) stator and rotor, b) rotor in exploded view

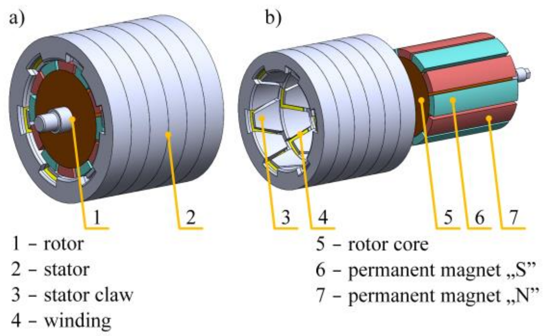


Fig. 3. Machine with claws on stator [3]: a) stator and rotor, b) exploded view

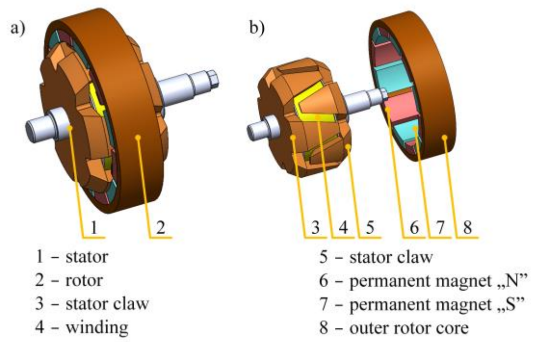


Fig. 4. Claw pole machine with outer rotor [4]: a) stator and rotor, b) exploded view

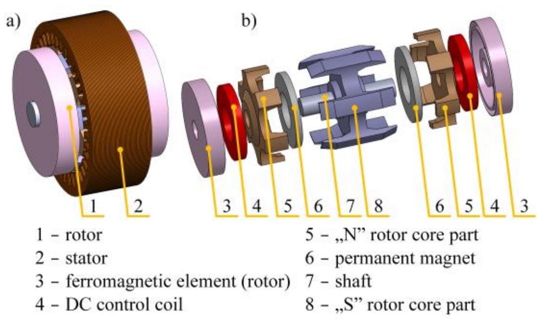


Fig. 5. Hybrid excited claw pole machine with stationary DC control coils [8]: a) stator and rotor, b) rotor in exploded view

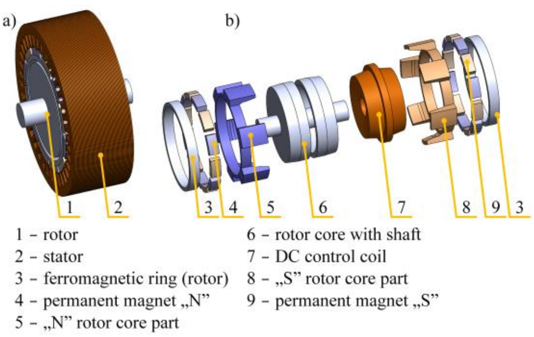


Fig. 6. Hybrid excited claw pole machine with stationary DC control coil [1]: a) stator and rotor, b) rotor in exploded view

3. Research model of a claw pole machine with hybrid excitation

Initial simulation studies and visualization of the concept of a hybrid excited claw pole machine with a laminated rotor were presented in the previous publication. This paper [12] presents the relationship between the volume of used magnetomotive force sources in the form of permanent magnets and the parameters of the machine itself. In line with the conclusions, it was decided to build an experimental model with 15 mm wide magnets, but an additional change was made to [12], i.e. all coils of each phase were connected parallelly to reduce the induced voltage. This model is the subject of the research presented in this paper. Figure 7 shows a visualization of the constructed machine. The basic parameters of the machine are summarized in Table 1.

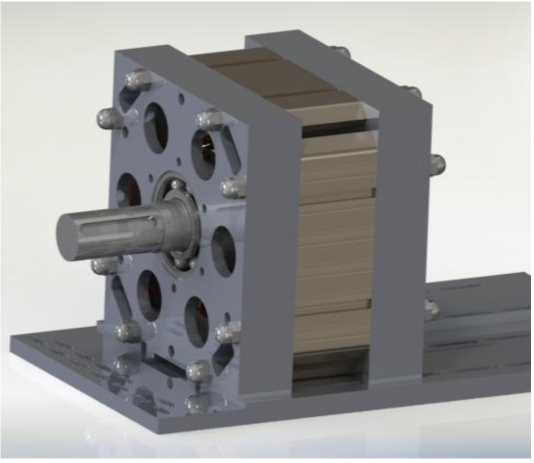


Fig. 7. Visualization of the prototype of a claw pole machine

Table 1. Basic parameters of the machine

name	symbol	value
number of poles	p	12
external dimensions of the stator plates	a x a	326 x 326 mm
inner diameter of a stator	i_{ds}	164 mm
number of stator slots	s	36
outer diameter of a rotor	o_{dr}	162 mm
PM width	w_{PM}	15 mm
PM thickness	t_{PM}	3 mm
machine length	l	100 mm
air gap	g	1 mm
number of turns of DC control coil	z_{DC}	1800

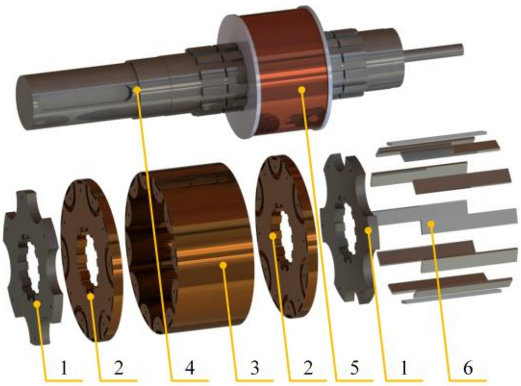
The machine has 6 pairs of poles. The rotor claws were obtained by appropriate arrangement of the three types of rotor plates (Fig. 8). The sheets of the first type ("1" on Fig. 8) with a thickness of 10 mm are made of carbon steel. They are a path only for the flux of one type of poles, which means that the initial sheet conducts the magnetic flux from the "*N*" pole, and the end – from the "*S*" pole. Therefore, the sheets "1" are rotated 180 electrical degrees with respect to each other.

The sheets of the second type ("2" on Fig. 8) are made of M400-50A electrical steel with a thickness of 0.5 mm. These are packed of 20 sheets with a total thickness of 10 mm and are a path for a single type of flux (similar to the first type of sheets), but there are opposite pole ends between the poles. The two packages are also rotated 180 electrical degrees to each other.

Sheets of the third type ("3" on Fig. 8) are also made of M400-50A steel, they are the continuations of the poles of both types ("*N*" and "*S*"). The sheets of the first type protrude beyond the stator area, while the second and third types have magnetic bridges 1 mm thick and are coplanar with the stator plates. Bridges perform a mechanical function and are so narrow that they are completely saturated.

Inside the rotor's core (in the area of the third type of metal sheets) there is a coil that regulates the excitation flux. The whole is put on the steel shaft, which is also a path for the magnetic flux from the magnets and the rotor coil.

The three-phase stator has 36 slots in which 36 coils are placed in a two-layer arrangement. All coils of a given phase are connected parallelly, so the phase windings have 12 parallel branches.

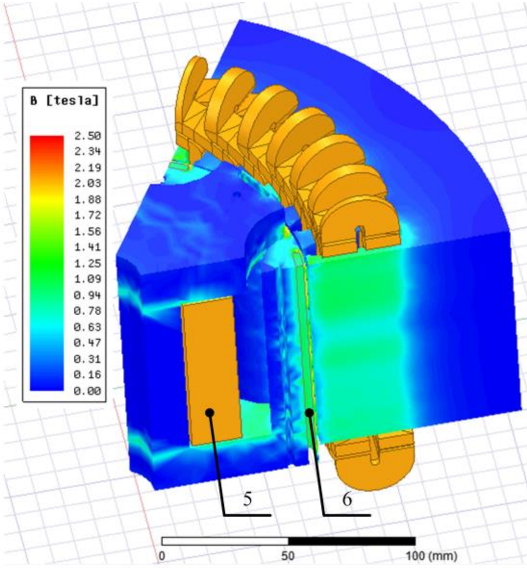


1 – 1st type steel plate (10 mm thick)
2 – 2nd type steel plates (0.5 mm thick)
3 – 3rd type steel plates (0.5 mm thick)
4 – shaft
5 – DC control coil
6 – permanent magets (15 x 3 x 100 mm)

Fig. 8. The rotor of the claw pole machine prototype in an exploded view

4. Results of simulation tests

Figure 9 shows the simulation model of the tested claw pole machine.



5 – DC control coil
6 – permanent maget

Fig. 9. FEA model of the claw pole machine

During the simulation tests, the induced voltage waveforms, the cogging torque and the electromagnetic torque of the machine were determined in various power conditions of the rotor excitation coil. Figure 10 shows the waveforms of the voltage induced in the selected machine phase for currents in the rotor coil from -3 A to 3 A.

The next figure (Fig. 11) shows the waveforms of the back-emf induced between two phases, assuming the association of phases into a star.

The result of the cogging torque as a function of the excitation current is shown in figure 12.

The electromagnetic static torque as a function of the stator current and the current in the DC rotor control coil was also tested. Examples of waveforms (for stator current I_s = 20 A) are shown in the next figure (Fig. 13).

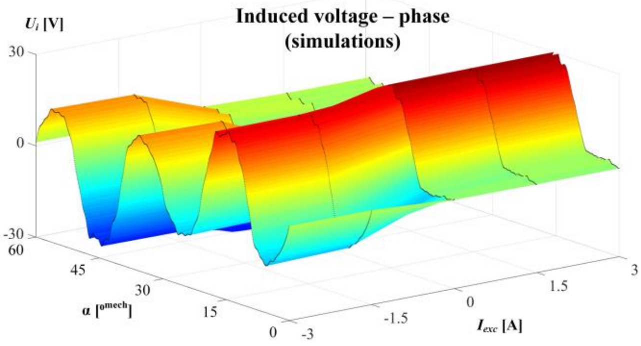


Fig. 10. Induced phase voltage – simulations

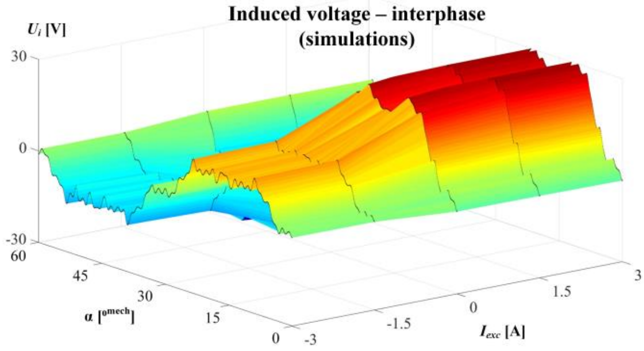


Fig. 11. Induced phase-to-phase voltage – simulations

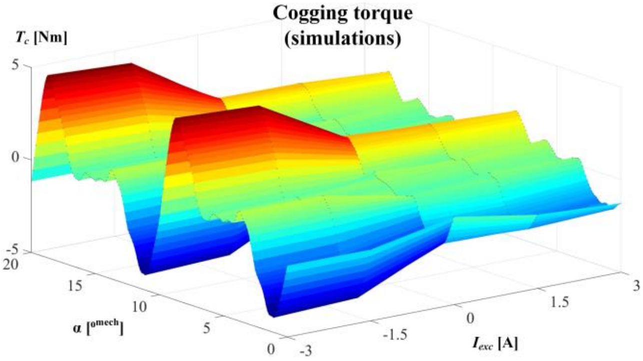


Fig. 12. Cogging torque – simulations

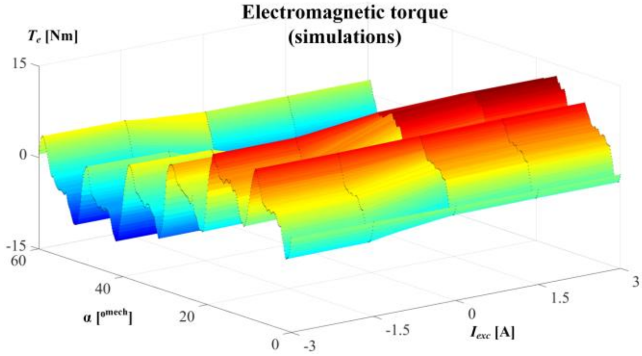


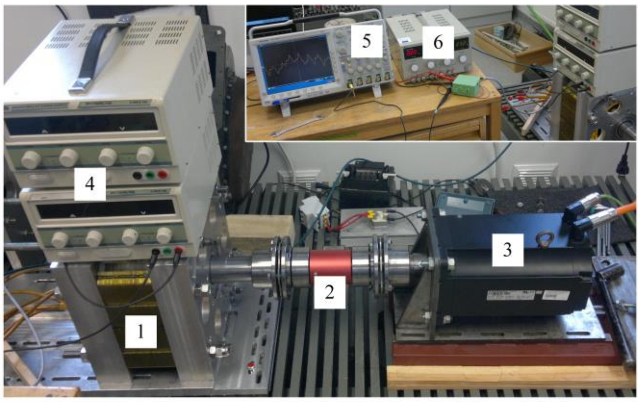
Fig. 13. Electromagnetic torque – simulations

5. Results of an experimental research

Based on the simulation tests performed, an experimental model of a claw machine with hybrid excitation and a laminated rotor core was built. Figure 14 shows the experimental stand to test the prototype.

The waveforms of back-emf phase and phase-to-phase voltages are presented in the next figures 15 and 16.

Figure 17 shows the cogging torque as a function of the current in the rotor coil I_{exc}, while Figure 18 shows the waveforms of the electromagnetic torque measured in a static state. During this study, the machine was star connected. One of the phases is connected to the positive pole of the DC power supply and the negative pole to the other two phases connected parallely. These tests are analogous to those carried out in the simulation.



1 – claw pole machine prototype
2 – torque meter
3 – driving machine
4 – DC power suppliers for stator windings
5 – oscilloscope
6 – DC power supplier for rotor control coil

Fig. 14. Experimental stand to test the claw pole machine prototype

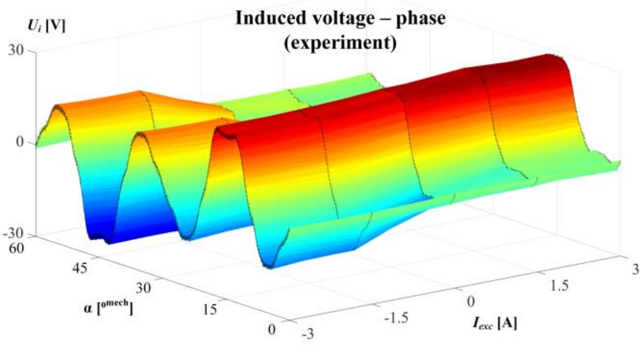


Fig. 15. Induced phase voltage – experiment

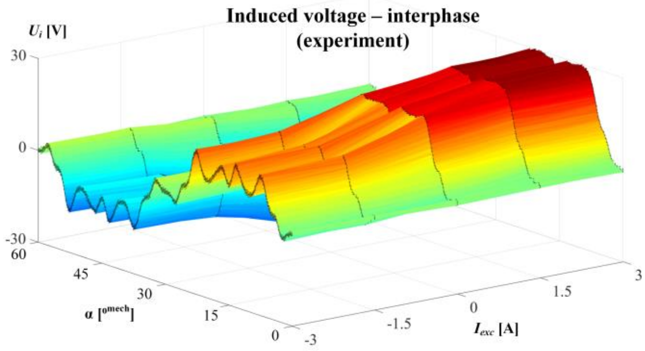


Fig. 16. Induced phase-to-phase voltage – experiment

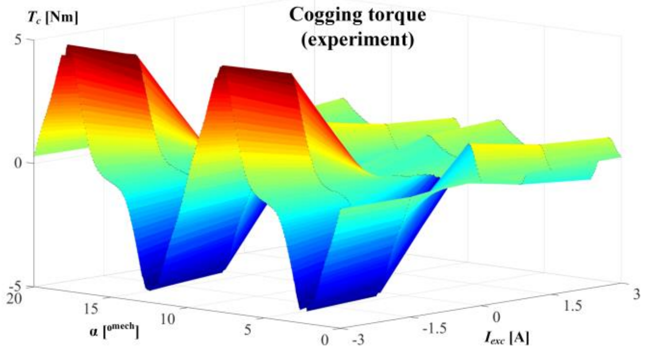


Fig. 17. Cogging torque – experiment

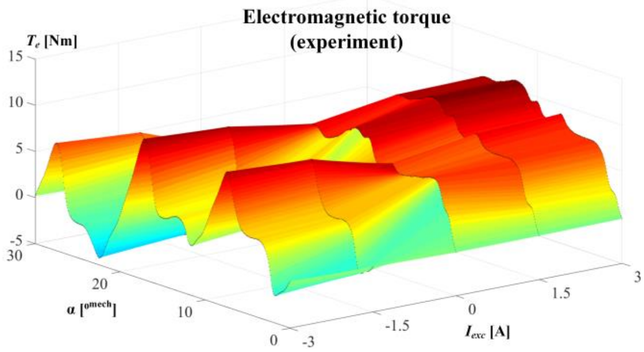


Fig. 18. Electromagnetic torque – experiment

6. Comparison of research results

Comparing the results of simulation and experimental studies, it is possible to come to very interesting conclusions. Figure 19 shows the change in the rms value of the induced voltage depending on the excitation current for phase (red color) and phase-to-phase (blue color) voltages obtained during simulation and experimental tests.

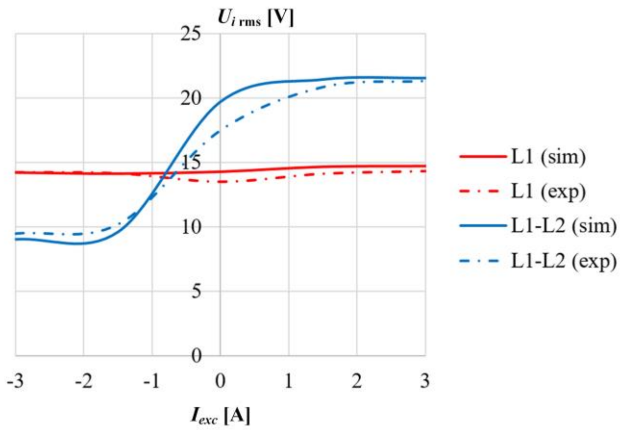


Fig. 19. Rms value of back-emf vs. excitation current I_{exc}

A similar comparison for the maximum voltage values is shown in the next figure 20.

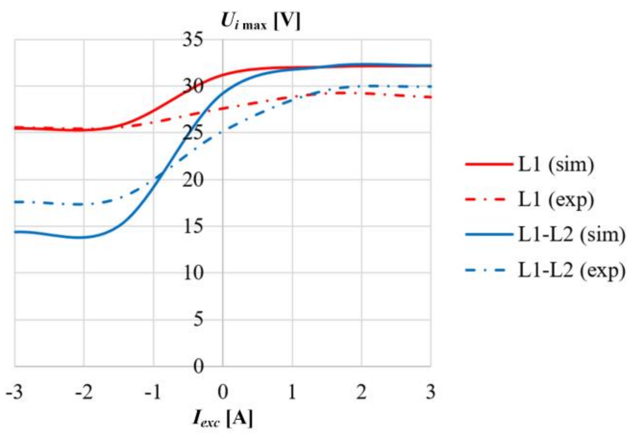


Fig. 20. Maximal value of back-emf vs. excitation current I_{exc}

Moreover, the results of the research in terms of the cogging torque, presented in figure 21, were compared.

Comparing the obtained results of simulation and experimental tests, it can be concluded that the numerical model correctly reflects the real prototype. Analyzing the results, there can be seen that the parameters of the machine do not obviously change depending on the value of the current in the excitation coil I_{exc}. Figure 19 shows that the root mean square value of the induced voltage for phase waveforms changes insignificantly, while the rms value of phase-to-phase voltages changes in the range from approx. 9 V to approx. 21 V, i.e. FCR (field control range) is as high as 2.33. The maximum voltage value changes slightly differently, ranging from 25 V to 32 V for phase voltages and 14 V to 32 V for phase-to-phase voltages (see Fig. 20).

In case of the cogging torque, increasing the excitation does not cause significant changes in the cogging torque (see Fig. 21), but in the field of decreasing, this torque increases from about 2 Nm to about 5 Nm. The increase in pulsation is also clearly visible in case of the electromagnetic torque, which is also shown in figure 13 (simulations) and figure 18 (experiment).

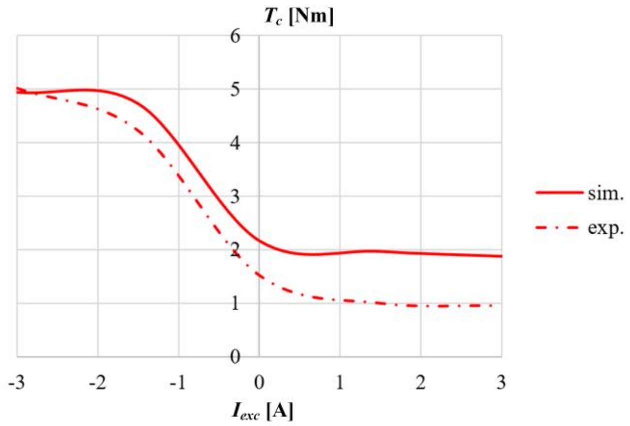


Fig. 21. Maximum value of the cogging torque

7. Conclusions

As a part of the research, simulation and experimental analysis of the parameters of a claw pole machine with hybrid excitation having a laminated rotor core were carried out. The paper presents the numerical and experimental model of the machine as well as the analogous test results. The influence of the current in the excitation coil on the induced phase and phase-to-phase voltage as well as on the cogging and electromagnetic torque was investigated.

The research shows that despite the fact that the rms value of the induced phase voltage changes slightly, when the machine is star-connected, a wide range of interfacial voltage regulation is visible, for which the FCR coefficient is 2.33. According to the authors, such a dependence results from the complicated shape of the induced voltage waveform, in which there are maxima and minima in appropriate time ranges. The resultant phase-to-phase voltage, which results from the difference between the instantaneous values of the voltages in the two phases, results in a wide adjustment range.

In case of torque tests, the conclusion is as follows: flux straightening does not cause a visible increase in the pulsation of the cogging and electromagnetic torque, while the higher the current in the field of weakening, the greater the cogging torque and pulsations of the electromagnetic torque.

Acknowledgments

This work has been supported with the grant of the National Science Centre, Poland 2018/02/X/ST8/01112.

References

[1] Burkhardt Y., Schleicher K., Klöpzig M.: A novel hybrid excited synchronous machine for (H)EV applications, IEEE Xplore, 2014.

[2] Di Barba P., Mognaschi M.E., Bonislawski M., Palka R., Paplicki P., Piotuch R., Wardach M.: Hybrid excited synchronous machine with flux control possibility. International Journal of Applied Electromagnetics and Mechanics 52/2016, 1615–1622 [http://doi.org/10.3233/JAE-162190].

[3] Guo Y., Zhu J., Dorrell D., Lu H. Y., Wang Y.: Development of a Claw Pole Permanent Magnet Motor with a Molded Low-Density Soft Magnetic Composite Stator Core, Proceedings of IEEE Energy Conversion Conference & Expo 2009, 294–301.

[4] Guo Y., Zhu J. G., Zhong J. J., Wu W.: Core Losses in Claw Pole Permanent Magnet Machines With Soft Magnetic Composite Stators. IEEE Transactions on Magnetics 39(5)/2003, 3199–3201.

[5] Hua H., Zhu Z.Q., Zhan, H.: Novel Consequent-Pole Hybrid Excited Machine with Separated Excitation Stator. IEEE Transactions on Industrial Electronics 63/2016, 4718–4728.

[6] Jahns T. M.: Flux-weakening regime operation of an interior permanent-magnet synchronous motor drive. IEEE Transactions on Industrial Application 23/1987, 681–689.

[7] Leroy V., Foveau V.: Claw rotor equipped with an insulator for an excitation coil and magnets, and rotary electrical machine equipped with a claw rotor, patent US 20130009504 A1, 2013.

[8] Melcescu L., Cistelecan M. V., Craiu O., Popescu M.: Numerical Analysis of Claw Pole Synchronous Machine with Hybrid Contactless Excitation. Electrical Review 7b/2012, 106–109.

[9] Wang Y., Deng Z.: Hybrid Excitation Topologies and Control Strategies of Stator Permanent Magnet Machines for DC Power System. IEEE Transactions on Industrial Electronics 59/2012, 4601–4616.

[10] Wardach M., Paplicki P., Palka R.: Hybrid Excited Machine with Flux Barriers and Magnetic Bridges. Energies 11/2018, 676 [http://doi.org/10.3390/en11030676].

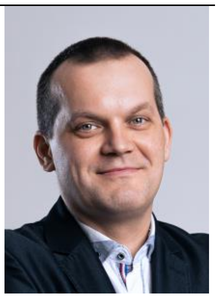
[11] Wardach M.: Hybrid excited claw pole generator with skewed and non-skewed permanent magnets. Open Physics 15/2017, 902–906 [http://doi.org/10.1515/phys-2017-0108].

[12] Wardach M.: The Influence of Permanent Magnet Amount on No-load Parameters of Hybrid Excited Claw Pole Machine with Laminated Rotor. Selected Issues of Electrical Engineering and Electronics (WZEE'2018), Szczecin, Poland, 2018.

D.Sc. Ph.D. Eng. Marcin Wardach
e-mail: marcin.wardach@zut.edu.pl

Marcin Wardach was born in Barlinek, Poland in 1980. He graduated and received a Ph.D. degree from the Electrical Department, Szczecin University of Technology, Szczecin, Poland, in 2006 and 2011, respectively. From 2020 until now, he has been an Associate Professor with the Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin. His research interests include the design of electrical machines and drives.

http://orcid.org/0000-0002-1017-9054



MSc. Eng. Paweł Prajzendanc
e-mail: pawel.prajzendanc@zut.edu.pl

Paweł Prajzendanc was born in Pyrzyce, Poland in 1990. Received a B.S. degree in electrical engineering from West Pomeranian University of Technology, Szczecin, Poland in 2015 and an M.S. degree in 2017. He is currently pursuing a Ph.D. degree in electrical engineering at the Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin, Poland. His awards and honors include a scholarship of Prime Minister of Poland. His research interests include the design of electrical machines and drives.

http://orcid.org/0000-0002-1662-4390

Eng. Kamil Cierzniewski
e-mail: cierzniewski_kamil@zut.edu.pl

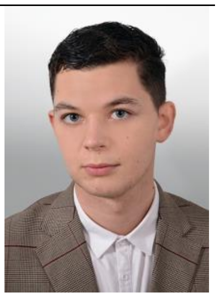
Kamil Cierzniewski was born in Koszalin, Polan in 1997. Received a B.S. degree in electrical engineering for West Pomeranian University of Technology, Szczecin, Poland in 2020. He is currently pursuing M.Sc. degree in electrical engineering at the Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin, Poland. He is a president of Student's Science Club of Association of Polish Electrical Engineers. His research interests include the design of electrical machines and drives.

http://orcid.org/0000-0003-3453-5233

Eng. Michał Cichowicz
e-mail: michal_cichowicz@zut.edu.pl

Michał Cichowicz was born in Gubin, Poland in 1998. Received a BEng degree in automatic control and robotics from West Pomeranian University of Technology Szczecin, Poland in 2021. He is currently pursuing two MEng degrees: electrical engineering and mechanical engineering. He is also a president of Student's Science Club of Practical Robotics "SKORP". His research interests include: the design of electrical machines and drives; robotics; mathematics.

http://orcid.org/0000-0002-8258-0330



Szymon Pacholski
e-mail: szymon_pacholski@zut.edu.pl

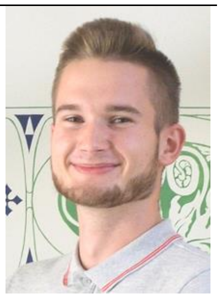
Szymon Pacholski was born in Goleniów, Poland in 1997. He graduated from Zespół Szkół Elektryczno-Elektronicznych w Szczecine as electrotechnician. In 2018 he started studying at West Pomeranian University of Technology. His research interest include design of vehicle drives, electrical machines and mechanical issues.

http://orcid.org/0000-0002-7588-4168

Mikołaj Wiszniewski
e-mail: wm46928@zut.edu.pl

Mikołaj Wiszniewski was born in Szczecin, Poland in 1999. He is graduate of technical school on mechatronics science profile. In 2019 he became student at West Pomeranian University of Technology on automation and robotics profile. His research interests include the design of electrical machines, PLC drivers and programming manipulators.

http://orcid.org/0000-0002-8258-8922



Krzysztof Baradziej
e-mail: bk42039@zut.edu.pl

Krzysztof Baradziej was born in Szczecin, Poland in 1997. He is graduate of technical school on electronics science profile. In 2017 he became student at West Pomeranian University of Technology on electrotechnical profile. His research interests include microcontrolers and li-ion cells.

http://orcid.org/0000-0003-1293-3409

Szymon Osipowicz
e-mail: os44085@zut.edu.pl

Szymon Osipowicz was born in Pyrzyce, Poland in 1999. He is graduate of high school on mathematics and physics profile. In 2018 he became student at West Pomeranian University of Technology on automation and robotics profile. His research interests include programming manipulators, PLC drivers, programming in C# language and the design of electrical machines.

http://orcid.org/0000-0003-3938-6388

otrzymano/received: 7.05.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2654

BATTERY SWAPPING STATIONS FOR ELECTRIC VEHICLES

Aleksander Chudy
Lublin University of Technology, Department of Electrical Engineering and Electrotechnologies, Lublin, Poland

Abstract. *Battery swapping is a promising technology when compared with the traditional electric vehicle charging stations. The time spent at a battery swapping station might be similar to the time spent at a filling station. The article presents information on attempts to implement this solution, methods of battery swapping, infrastructure and operation of battery swapping stations, as well as the benefits and key challenges of the battery swapping technology.*

Keywords: battery swapping, battery swapping stations, electromobility, electric vehicles

STACJE WYMIANY AKUMULATORÓW SAMOCHODÓW ELEKTRYCZNYCH

Streszczenie. *Wymiana akumulatorów samochodów elektrycznych to obiecująca technologia w stosunku do tradycyjnego podejścia do stacji ładowania samochodów elektrycznych, ponieważ czas spędzony na stacji wymiany akumulatorów może być podobny do czasu spędzonego na stacji benzynowej. W artykule przedstawiono informacje na temat prób implementacji tego rozwiązania, sposobów wymiany akumulatorów, infrastruktury i działania stacji wymiany akumulatorów oraz korzyści i kluczowych wyzwań tego rozwiązania..*

Słowa kluczowe: wymiana akumulatorów; stacje wymiany akumulatorów; elektromobilność; samochody elektryczne

Introduction

The development of electric vehicles (EVs) is widely popularised by their owners, EVs manufacturers, global governments and investors mainly due to the ability to reduce carbon dioxide emissions and thus reduce the environmental impact of the transport sector. The increasing penetration of electric vehicles has the potential to reduce pollutant emissions and reduce the consumption of conventional fuels. Moreover, EVs can also contribute to increasing the use of renewable energy sources such as wind and solar energy sources as the batteries in electric cars are capable of storing electricity [6].

Charging EVs' batteries plays a key role in the adaptability of electromobility. The current charging process mainly involves connecting an EV to a household electrical outlet or charging station and leaving it on for a few hours. This method requires much more time than refueling a vehicle with an internal combustion engine and is a serious barrier to the popularisation of EVs. The exception is super-fast high-power charging stations (up to 350 kW) that allow you to charge an electric car in about 20 minutes depending on the battery capacity. However, fast charging of many electric vehicles may have a negative impact on the parameters of energy quality [8, 9].

One solution to overcome obstacles related to charging EVs is to replace discharged batteries with fully charged ones at a battery swapping station (BSS). Unlike charging electric vehicles with a wired or wireless method, replacing the battery could only take a few minutes. Thanks to BSSs, they could extend their journeys and reduce the time needed for further driving [23].

1. Attempts to implement battery swapping in practice

In the early 1900s, manual battery swapping services were first proposed to overcome the limited range of EVs. This solution was popular until the mid-1920s [15].

In 2007, the company Better Place was founded. It provided the first commercially implemented automatic battery swapping service, but it was only available for electric cars – it was not implemented for heavy vehicles (buses and trucks). However, the company declared bankruptcy in Israel in May 2013, Mainly due to a flawed strategy of building large battery swapping stations in selected geographic locations. Users were often forced to travel long distances to swap batteries, which resultied in their dissatisfaction. Although the Better Place company went bankrupt, it demonstrated that automation of battery swapping is feasible in practice. Battery swapping technology for electric buses was first commercially deployed by China in 2008. During the Summer Olympics, batteries were replaced in about 50 buses running on various routes. Since then, battery swapping in such vehicles has been widely implemented in China, Japan, and South Korea [17, 26].

In 2014, the Chinese company NIO undertook the dissemination of battery swapping technology (NIO Power Grid technology). It offers a combination of multiple energy charging solutions focused on improving the user experience. The core of this strategy is based on NIO's chargeable, replaceable and upgradable smart energy service system. The company announced that by October 5, 2020, the total number of electric vehicle battery swapping services in China reached 1 million. This is twice as many as on May 25, 2020. This shows how fast the demand for this service is growing. Since spring 2020, NIO has also increased the number of battery swap stations from 131 to 155 [17, 26].

2. Battery swapping methods

The following are the various swapping methods, which are classified according to the location of the battery in the vehicle and the robotic arm's application point [1, 2]:

- bottom swapping is for vehicles with a battery that is located at the bottom of the vehicle. The swapping station is designed such that the vehicle is parked on a raised platform and the batteries are switched from the bottom using a robotic arm and other accessories that are usually located below ground level,
- top swapping: This is more widely seen with electric buses, where the batteries are mounted on top and the rooftop opens as the bus arrives, allowing the swapping to be done by a robotic arm,
- sideways swapping is commonly seen in buses and other vehicles where the sideways configuration is the most practical,
- rear swapping is seen in vehicles where the battery is mounted backwards. Typically in the case of vehicles with a big trunk.

3. The principle of operation of battery swapping stations

In order to establish and successfully implement battery swapping technology for electric vehicles such as cars, vans, and buses, extensive planning must be carried out, covering all necessary requirements, from the availability of batteries and chargers to the storage and management of data via the cloud and communication between components so as to ensure interoperability. BSSs can only operate successfully if there is continuous communication between the different components of the system (a vehicle, the BSS and information system). The information system is used to communicate with both the vehicle and the station [2].

The vehicle communicates with the information system using the WAVE communication system (integration of several communication networks into one), while the station uses local Internet access. When the battery charge of the vehicle is low, the information system receives a notification from the vehicle requesting a battery swapping service. The information system informs the station of the location of the vehicle and the estimated time of arrival so as to prepare an available battery until the vehicle arrives at the station. When the vehicle arrives at the station, the driver swipes one's membership card and the information system verifies all relevant data contained therein. This data includes information about the vehicle, battery, swapping history, completed transactions and other relevant information. All data must then be archived in the cloud and be accessible to the station owner and customer to ensure full transparency of procedures. Once the swapping is completed, the discharged battery is monitored for state of charge, degradation level, battery age, or number of charge and discharge cycles completed [1, 2].

The main components of the charging station include:

- a control room (controlling and monitoring the overall operation of the BSS),
- charging racks coupled with battery racks,
- a swapping track (the zone where the batteries are located during the swap),
- a swapping lane (the road where the vehicle is during the swap),
- a swapping robot,
- battery and charger service room,
- a service room for other BSS components.

In the BSS, the driver who arrives first should have priority compared to those who arrive later as is the case at conventional filling stations. In order to meet the demand of all customers approaching the BSS, it is necessary to recharge the batteries immediately after the swap so that they are available for the next customers. An integrated BSS should maximise the range with minimal waiting time for battery swapping.

In the literature, there are models presented where discharged batteries are transported to a battery charging station and after recharging, they are transported to the BSS again. The idea of this approach is presented in Fig. 1. Transporting a large number of centrally charged batteries usually requires the use of a logistics system. The battery charging strategies and schedules, location of BSSs as well as their construction have received much attention from researchers [3, 10, 13, 20–22, 24].

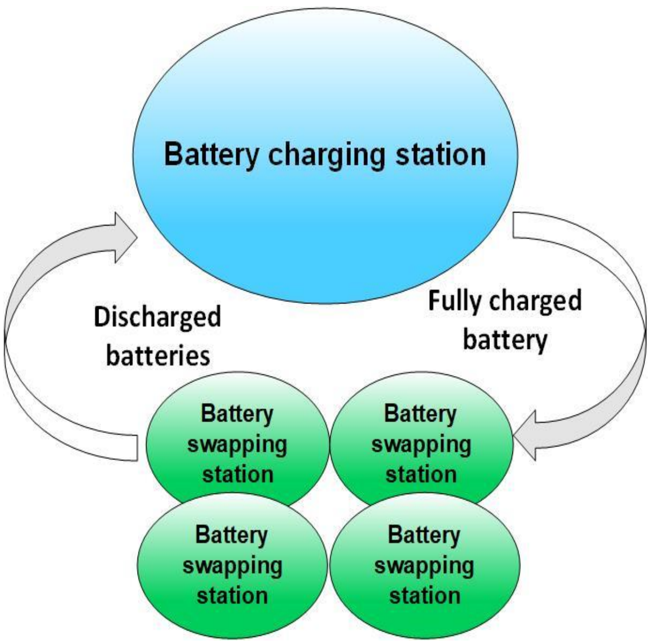


Fig. 1. An approach of battery swapping with the use of an independend battery charging station

Due to constraints such as a geographical location, a limited availability of BSSs and overcrowding at stations, there is a need to design a more flexible and efficient EV battery swap architecture. Two modes of battery swapping can be distinguished – passive mode and active mode. In passive mode, the goal of having enough energy to drive an EV anytime and anywhere is not realized. Drivers must travel to the BSS to replace the battery pack. Due to the limitations of the location of the station and the number of existing batteries, there may still be a long waiting time for an available battery at that time [2].

In the active mode, battery swapping is performed using an additional vehicle that pulls up to a vehicle with a battery that is near or discharged. Recently, a new high-speed EV battery swapping device has been developed [19]. Such a device can be mounted on a vehicle that is converted into a mobile BSS. The battery removal and installation operations take place at the same time. Thus, the whole swapping process is very fast and takes only a few minutes (approximately 3 minutes in the experimental environment). Using a van as the BSS, the exchange can be carried out at any time and any place.

4. Advantages of battery swapping stations

As shown in Fig. 1. battery swapping requires mutual interactions between a power system, an EV owner, and the battery swapping station.

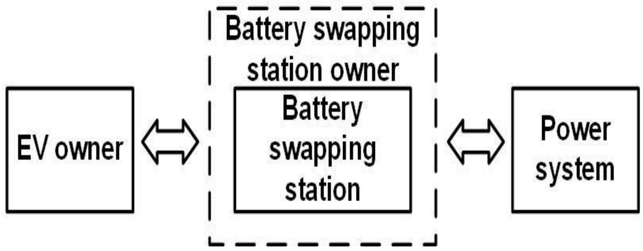


Fig. 1. Diagram of the interdependence of the electric vehicle owner, the battery swapping station owner, and the power system

The EV owner would like to replace his discharged battery with a fully charged one in the shortest time possible and move on while the BSS owner considers the most favorable electricity price for him to recharge the discharged battery and minimize the associated costs. Also important is the inherent interdependence between the BSS and the power system.

The behavior of EVs owners is unpredictable, and there are limited options for coordinated charging and discharging of grid-connected EV batteries at this time [7]. Uncontrolled charging of plug-in EVs might have a significant impact on the power system as they can contribute to load growth during peak demand. A battery swap solution offers a controlled charging strategy in terms of scheduling battery charging time without plugging in and immobilizing an entire EV for more than 20 minutes. Charging stations for the batteries themselves or battery swap stations that are also charging stations are able to defer charging to off-peak demand hours, which can solve the grid overload problem [4, 25].

From the power system's point of view, BSSs are a large flexible load. The energy storage capability of EV batteries provides an excellent opportunity for the owner of the BSS to offer grid services. By controlling the charging and discharging times of the batteries, the potential peak demand or overload caused by the increasing penetration of electric vehicles can be flattened. This can be achieved by setting an intelligent charging schedule without having to upgrade the current grid infrastructure. EV batteries can also help regulate frequency of a grid.

With a rapidly growing market for EVs, the need for fast chargers could become a major concern for consumers. To keep up with technological advances, infrastructure will need to be upgraded on the household side as well as at public charging stations. In this respect, battery swap stations are superior to traditional charging infrastructure because upgrades are only needed at the swap station location.

From the EV owner's perspective, battery swapping technology has several key advantages:

- lowering the price of the electric vehicle (the exchange stations own the batteries),
- reduced concern about battery life (use of advanced charging strategies),
- reduced cost of infrastructure upgrades.

5. Challenges of battery swapping stations

Speed is the biggest advantage of battery swapping. The whole operation can take a few minutes, which is about the same time as it takes to fill up a vehicle with a combustion engine. Another advantage of this solution is that the owner does not have to leave the vehicle to change a battery. He does not have to touch publicly accessible wires, buttons or display, which greatly facilitates social distance. However, the issues of battery swapping are more complex and have many challenges.

- Infrastructure

The infrastructure required to implement this solution is much more complex and costly than charging batteries at conventional charging stations. The difference between the BSS and a charging station is that the demand of the BSS can be continuously controlled. Furthermore, to meet the daily demand of any station, the number of charged batteries should always exceed a certain percentage of the daily demand. One possible strategy is to have one battery in the vehicle and another at the swap station, i.e. two batteries for each car [16].

- Feasibility

The design of batteries as well EVs themselves is currently one of the biggest obstacles to the battery swapping technology. During the battery swapping process, sparks can occur and contacts can degrade. The design of the battery pack should consider the reliability and safety of the vehicle owner when removing the battery pack and placing it back into the vehicle [5, 14].

- Interchangeability

The compatibility of the batteries as well as the battery swapping technology itself can ensure the success of this solution and make it dominant in the market. The main requirement is the availability of similar interchangeable battery packs from different manufacturers, which of course requires their common agreement. This might cause limitation of innovation and battery development will state as cell manufacturers, at the end of the product chain, will have to design standardized cells. It can be argued that there could be different power segments for a set of batteries, but this would create a supply and demand problem, as well as possible compatibility issues with vehicles that cannot use higher or lower power batteries.

- Battery degradation

The range of an electric car decreases as the battery deteriorates. Customers will mostly demand the newest batteries possible because they will provide more range compared to batteries that have been around longer. For this reason, the use of a particular battery in battery exchange stations can be quite short despite favourable charging conditions. However, such batteries can later be used as energy storage from renewable energy sources [18].

- Battery ownership

In the current model, the vehicle owner owns the battery. If BSSs were to become more popular, vehicle owners would have to buy an additional battery or rent one. It also seems interesting to be able to buy an electric car without owning the battery. This has several advantages. Since the vehicle owner would not own the battery then the electric car would be less expensive. The owner could pay the cost of leasing the battery along with the cost of charging and replacing the battery at the station. This could be charged each time the battery is replaced or could be, for example, a monthly fee. A form of long-term leasing could be expensive because the costs include a minimum of two sets of batteries and the service fee of the swapping station (conventional station charging requires only one battery, purchased with the car) [2].

6. Battery swapping stations for public transport

One of the most important and widely used means of public transportation is city buses. Due to their economic feasibility and their ability to serve almost all possible urban routes with little maintenance, they are very popular. The characteristic features of buses are as follows [12]:

- they drive predetermined routes with a specific schedule,
- they run almost exclusively during the hours of transport services,
- they stop for a short break between routes, which means that after the end of one route the bus is immediately prepared for the next route with a minimal stopping time. This is to maximize the utilisation of the bus fleet without keeping them at the depots for long periods of time.

These features can be interpreted in two ways and can be arguments for, but also against, electrification of public transport. A serious issue that needs special attention is the fastest possible recharging of the bus batteries. A battery swapping may then be necessary to avoid schedule delays and to ensure the highest possible usability of the buses. The advantages of this solution are [11]:

- speed (compared to the various vehicle charging strategies available) which allows the consumer to continue driving without major delays,
- shortening the waiting time for the bus battery to be charged, as well as extending the life of the battery (it does not need to be quickly charged at 50 kW and above),
- bidirectionality of energy flow, and the ability to provide ancillary services. Batteries can mitigate issues related to the variability and uncertainty of renewable energy production.

Using a battery swapping technology can be cost-effective for bus owners, as well as for BSS owners, due to the mutually beneficial cost reductions resulting from charging the batteries during low load periods and the ability to return excess energy to the grid.

An electric bus with a replaceable battery is being developed under a Polish research project. The project is implemented by the bus manufacturer Autosan, the National Centre for Research and Development, and the Łukasiewicz PIMOT consortium [27].

7. Conclusions

The swapping of EV batteries is a promising alternative to traditional battery charging stations. The implementation of such a technology could positively affect the electric power system due to the possibility of implementing auxiliary services (spinning reserve, active power regulation, load balancing, reactive power compensation, support for renewable energy sources).

This paper presents aspects of the implementation of battery substations, the principle of the substation, how batteries can be replaced and the benefits of BSSs for drivers, substation owners and the power system. Furthermore, the key challenges of BSSs are discussed.

To fully use the potential of the BSS, EV batteries should be easily replaceable and accessible. One important requirement would be the establishment of consistent standards – how to replace batteries as well as battery standards. According to the author, the best strategy for using BSSs is a model where the battery is owned by the company that owns the BSS. The most important advantage, besides the reduced charging time, is that the purchase price of an EV could then be lower, which could contribute to the development of electromobility.

References

[1] Adegbohun F., Jouanne A. von, Lee K.: Autonomous Battery Swapping System and Methodologies of Electric Vehicles. Energies 12(4), 2019, 667 [http://doi.org/10.3390/en12040667].

[2] Ahmad F., Saad Alam M., Saad Alsaidan I., Shariff S. M.: Battery swapping station for electric vehicles: oppor-tunities and challenges. IET Smart Grid 3(3), 2020, 280–286 [http://doi.org/10.1049/iet-stg.2019.0059].

[3] Armstrong M., El Hajj Moussa C., Adnot J., Galli A., Riviere P.: Optimal recharging strategy for battery-switch stations for electric vehicles in France. Energy Policy 60, 2013, 569–582 [http://doi.org/10.1016/j.enpol.2013.05.089].

[4] Banerjee A., Murali V., Venkoparao V. G.: Measuring Performance Impact of Battery Swapping on Mobility Behavior. 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), 2018, 1–5.

[5] Barré A., Suard F., Gérard M., Montaru M., Riu D.: Statistical analysis for understanding and predicting battery degradations in real-life electric vehicle use. Journal of Power Sources 245, 2014, 846–856 [http://doi.org/10.1016/j.jpowsour.2013.07.052].

[6] Chudy A.: The review of selected electrical energy storage techniques. IAPGOS 9(1), 2019, 23–28 [http://doi.org/10.5604/01.3001.0013.0890].

[7] Chudy A.: Zarządzanie flotą pojazdów elektrycznych w sieciach inteligentnych – sterowanie ładowaniem oraz zagadnienia optymalizacyjne. Interdyscyplinarność w badaniach naukowych prace doktorantów Politechniki Lubelskiej. Wydawnictwo Politechniki Lubelskiej, Lublin 2020.

[8] Chudy A., Mazurek P.: Electromobility – the Importance of Power Quality and Environmental Sustainability. J. Ecol. Eng. 20(10), 2019, 15–23 [http://doi.org/10.12911/22998993/112713].

[9] Chudy A., Stryczewska H. D.: Electromagnetic compatibility testing of electric vehicles and their chargers. IAPGOS 10(3), 2020, 70–73 [http://doi.org/10.35784/iapgos.1687].

[10] Feng J., Hou S., Yu L., Dimov N., Zheng P., Wang C.: Optimization of photovoltaic battery swapping station based on weather/traffic forecasts and speed variable charging. Applied Energy 264, 2020, 114708 [http://doi.org/10.1016/j.apenergy.2020.114708].

[11] Kim J., Song I., Choi W.: An Electric Bus with a Battery Exchange System. Energies 8(7), 2015, 6806–6819 [http://doi.org/10.3390/en8076806].

[12] Li W., Li Y., Deng H., Bao L.: Planning of Electric Public Transport System under Battery Swap Mode. Sustainability 10(7), 2018, 2528 [http://doi.org/10.3390/su10072528].

[13] Liang Y., Cai H., Zou G.: Configuration and system operation for battery swapping stations in Beijing. Energy 214, 2021, 118883 [http://doi.org/10.1016/j.energy.2020.118883].

[14] Mahmoudzadeh Andwari A., Pesiridis A., Rajoo S., Martinez-Botas R., Esfahanian V.: A review of Battery Electric Vehicle technology and readiness levels. Renewable and Sustainable Energy Reviews 78, 2017, 414–430. [http://doi.org/10.1016/j.rser.2017.03.138].

[15] Mahoor M., Hosseini Z. S., Khodaei A., Kushner D.: Electric Vehicle Battery Swapping Station. 2017 [https://arxiv.org/abs/1710.06895].

[16] Mak H.-Y., Rong Y., Shen Z.-J. M.: Infrastructure Plan-ning for Electric Vehicles with Battery Swapping. Management Science 59(7), 2013, 1557–1575 [http://doi.org/10.1287/mnsc.1120.1672].

[17] Qi W., Zhang Y., Zhang N.: Scaling Up Battery Swapping Services in Cities. SSRN, 2020 [http://doi.org/10.2139/ssrn.3631796].

[18] Quinard H., Redondo-Iglesias E., Pelissier S., Venet P.: Fast Electrical Characterizations of High-Energy Sec-ond Life Lithium-Ion Batteries for Embedded and Stationary Applications. Batteries 5(1), 2019, 33 [http://doi.org/10.3390/batteries5010033].

[19] Shao S., Guo S., Qiu X.: A Mobile Battery Swapping Service for Electric Vehicles Based on a Battery Swap-ping Van. Energies 10(10), 2017, 1667 [http://doi.org/10.3390/en10101667].

[20] Sun B., Sun X., Tsang D. H.K., Whitt W.: Optimal battery purchasing and charging strategy at electric vehicle battery swap stations. European Journal of Operational Research 279(2), 2019, 524–539. [http://doi.org/10.1016/j.ejor.2019.06.019].

[21] Tan X., Sun B., Wu Y., Tsang D. H.K.: Asymptotic performance evaluation of battery swapping and charging station for electric vehicles. Performance Evaluation 119, 2018, 43–57 [http://doi.org/10.1016/j.peva.2017.12.004].

[22] Verma A.: Electric vehicle routing problem with time windows, recharging stations and battery swapping stations. EURO Journal on Transportation and Logistics 7(4), 2018, 415–451 [http://doi.org/10.1007/s13676-018-0136-9].

[23] Wu C., Lin X., Sui Q., Wang Z., Feng Z., Li Z.: Two-stage self-scheduling of battery swapping station in day-ahead energy and frequency regulation markets. Applied Energy 2020, 116285 [http://doi.org/10.1016/j.apenergy.2020.116285].

[24] Yang J., Sun H.: Battery swap station location-routing problem with capacitated electric vehicles. Computers & Operations Research 55, 2015, 217–232 [http://doi.org/10.1016/j.cor.2014.07.003].

[25] Zeng B., Luo Y., Zhang C., Liu Y.: Assessing the Impact of an EV Battery Swapping Station on the Reliability of Distribution Systems. Applied Sciences 10(22), 2020, 8023 [http://doi.org/10.3390/app10228023].

[26] nio.com/blog/brief-history-battery-swapping (18.03.2021).

[27] transinfo.pl/infobus/autosan-zbuduje-elektrobus-z-wymiennymi-bateriami-_more_122722 (16.03.2021).

M.Sc. Eng. Aleksander Chudy
e-mail: a.chudy@pollub.pl

He is graduated from Lublin University of Technology in 2018. In the same year he started his Ph.D. studies in the field of Electrical Engineering. Currently he works as an assistant in the Department of Electrcal Engineering and Electrotechnologies of Lublin University of Technology. His research is focused on the impact of electromobility on power quality parameters and electromagnetic compatibility.

http://orcid.org/0000-0002-3183-8450

otrzymano/received: 2.05.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2659

OVERVOLTAGE PROTECTION OF PV MICROINSTALLATIONS – REGULATORY REQUIREMENTS AND SIMULATION MODEL

Klara Janiga
Lublin University of Technology, Faculty of Electrical Engineering and Computer Science, Lublin, Poland

Abstract. *In low-voltage power networks with a large share of distributed energy sources, the phenomenon of overvoltage is increasingly observed. Although it may be desirable to raise the voltage value downstream of the network, in some cases the upper allowable voltage limit is exceeded. The method of eliminating voltage rises commonly used in the Polish power system is the installation of overvoltage protections, disconnecting the source from the grid. Such action reduces the profitability of prosumer installations, discouraging future potential investors. It turns out, however, that this is not the only disadvantage of such a solution. Sudden and uncoordinated disconnections and reconnections of more energy sources cause abrupt voltage changes that negatively affect the voltage conditions in the network. The aim of the paper is to present the operating algorithms of a standard overvoltage relay used in inverters of photovoltaic microinstallations. These algorithms – described in standards and national regulations – were tested in a typical inverter used in public low-voltage networks and implemented in the created simulation model of the relay. The described tests will be used for further work to demonstrate the need to coordinate the operation of overvoltage protections or replace them with other measures to improve voltage conditions in the grid with high share of photovoltaic sources.*

Keywords: overvoltage protection, distributed power generation, photovoltaic systems, power distribution lines

ZABEZPIECZENIE NADNAPIĘCIOWE MIKROINSTALACJI PV – WYMAGANIA I MODEL SYMULACYJNY

Streszczenie. *W sieciach elektroenergetycznych niskiego napięcia z dużym udziałem rozproszonych źródeł energii coraz częściej obserwuje się zjawisko podskoku napięcia. Choć samo podniesienie wartości napięcia w głębi sieci może być pożądane, to jednak w niektórych przypadkach dochodzi do przekroczenia górnej dopuszczalnej granicy napięciowej. Powszechnie stosowanym w polskim systemie elektroenergetycznym sposobem eliminacji podskoków napięcia jest instalowanie zabezpieczeń nadnapięciowych, wyłączających źródło z sieci. Działanie takie obniża opłacalność instalacji prosumenckich, zniechęcając przyszłych potencjalnych inwestorów. Okazuje się jednak, że nie jest to jedyna wada takiego rozwiązania. Nagłe i nieskoordynowane wyłączenia i ponowne załączenia większej liczby źródeł energii powodują skokowe zmiany napięcia, które negatywnie wpływają na warunki napięciowe w sieci. Celem artykułu jest zaprezentowanie algorytmów działania standardowego przekaźnika nadnapięciowego stosowanego w falownikach mikroinstalacji fotowoltaicznych. Algorytmy te – opisane w normach i krajowych przepisach – przetestowano w typowym falowniku stosowanym w publicznych sieciach niskiego napięcia oraz zaimplementowano w utworzonym modelu symulacyjnym przekaźnika. Opisane badania posłużą do dalszych prac, mających wykazać konieczność koordynacji działania zabezpieczeń nadnapięciowych lub zastąpienia ich innymi środkami poprawy warunków napięciowych w sieci z dużym nasyceniem fotowoltaiką.*

Słowa kluczowe: zabezpieczenie nadnapięciowe, rozproszona generacja, systemy fotowoltaiczne, elektroenergetyczne sieci niskiego napięcia

Introduction

The appearance of distributed energy sources in the low-voltage (LV) network brings many benefits to both the installation owner and the distribution system operator (DSO). Prosumer installations – which in Polish conditions are mostly photovoltaic (PV) microinstallations – generate additional power in the grid, which allows to cover the growing demand for electricity. Thus, it reduces the need to build new units in conventional power plants, replacing them with emission-free sources. The production of electricity close to the place of its consumption allows to reduce power losses in the network and also – in some situations – to improve the quality parameters of energy, for example by reducing voltage drops.

Although the increase in the number of PV microinstallations connected to the LV grid is a desirable phenomenon, it also has some negative effects. One of the most significant effect is the increase in the voltage value (so-called “overvoltage” or “voltage boosting”) [1, 2, 4]. This phenomenon occurs when the power generated exceeds the demand in the given network. Such a situation usually takes place only for a part of the day, in networks with a large share of PV installations, when high insolation is accompanied by low local power demand. A measure commonly used in the national power system to prevent overvoltage is the use of overvoltage protections implemented in PV installation inverters. These protections disconnect the generator when the voltage at the connection point of the PV source is too high, which, although it causes the intended voltage reduction, is also associated with a decrease in the profitability of the installation for the owner due to a break in energy production. You can also expect other negative effects related to the uncoordinated operation of many protections of this type, causing sudden voltage changes. The research described in this paper was intended to explore the topic of algorithms for the operation of standard overvoltage relays used in PV inverters. These studies are the starting point for further work to show the negative impact of uncoordinated operation of overvoltage protections on the stability of voltage conditions in the network.

The paper is structured as follows: section 1 provides an overview of the standards and national regulations for the operation and setting of overvoltage relays in low-voltage grid-connected PV microinstallations. Section 2 presents the results of tests of the inverter in terms of the operation of the overvoltage protection. Section 3 contains a description of the created simulation model of overvoltage protection. Section 4 describes the most important conclusions and further research directions.

1. Overvoltage protection in PV installations – requirements of standards and regulations

A photovoltaic microinstallation can be defined – according to the Renewable Energy Sources Act [5] – as an installation with a total electric power of up to 50 kW connected to the power grid with a rated voltage lower than 110 kV. These installations are included in the group of type A power generating modules (PGM) according to the division introduced in Commission Regulation (EU) 2016/631 of April 14, 2016 establishing the EU network code (so-called the NC RfG – Network Codes Requirements for Generators) [6]. The provisions of this regulation as well as other supplementary national regulations are compiled in detail in the documents issued by individual distribution network operators, such as "The Instruction of Distribution Network Operation and Maintenance (IRiESD)" (e.g. [7]) or the "Set of requirements for type A power generating modules, including microinstallations ”(e.g. [12]). Although each of the DSOs published their own version of the above documents, the provisions contained therein are consistent throughout the country.

As regards equipping microinstallations with protection against voltage increase, the domestic DSOs specified the requirements contained in table 1.

Table 1. Required settings of protection against voltage increase in microinstallations connected to the LV grid [12]

Protection type	Trip value		Maximum disconnection time	Minimum operate time
Over-voltage – stage 1*	1.1 U_n	253.0 V	3.0 s	-
Over-voltage – stage 2	1.15 U_n	264.5 V	0.2 s	0.1 s
* 10-minute mean value, according to EN 50160. Detailed requirements for the measurement of the mean value are included in the Polish standards.				

The overvoltage protection should be integrated with the inverter and operate in two stages. The second stage of protection is based on the determination of the RMS voltage value, and its almost instantaneous activation takes place when the value of 1.15 U_n is exceeded. The first stage of protection should respond to a 10-minute mean value, determined in accordance with EN 50160:2010 [8] and national standards. The PN-EN 50549-1:2019-02 [9] and the earlier PN-EN 50438:2014-02 [10] standards specify the method of determining the criteria values of protection. According to the above standards, RMS values should be set as "true r.m.s." or as fundamental component-values. Determining the 10-minute mean value should meet the following requirements:

- comply with 10 min aggregation of EN 61000-4-30, class S;
- be based on the calculation of the square root of the arithmetic mean of the squared input values over 10 min;
- a moving window shall be used (in deviation from EN 61000-4-30);
- the calculation of a new 10-min value at least 3 s is sufficient, which is then be compared with the trip value.

The 61000-4-30 standard [11] defines the method of aggregation in the 10-minute interval for class S. The basic time for measuring the voltage value should be a 10-period interval (for a system with a frequency of 50 Hz). Determining the 10-minute mean value should therefore be based on the RMS values determined in the intervals of 0.2 s (time of 10 periods).

Automatic reconnection (after tripping of the interface protection) can take place when the voltage at the connection point is kept within the limits of 0.85 $U_n \leq U \leq 1.1\ U_n$ for the minimum observation time equal to 60 seconds.

2. Tests of the operation of overvoltage protection in the PV inverter

To test the overvoltage protection, an inverter of a three-phase PV microinstallation of one of the buildings of Lublin University of Technology (LUT) was selected. This installation consists of 60 photovoltaic panels installed on the roof of the Centre for Innovation and Advanced Technologies of LUT. The total power of the microinstallations is 15 kW. The SMA Sunny Tripower STP 150000TL-10 inverter used in the installation is a typical device that allows the connection of photovoltaic panels to the public low-voltage network. The inverter is equipped by the manufacturer with the required set of protections, including protection against voltage increase, the settings of which are presented in the table 2. These settings result from the selected country standard setting – compliance with EN50438 has been selected.

The test stand for testing the operation of overvoltage protection has been equipped with a three-phase autotransformer enabling smooth voltage regulation. The change of voltage in the autotransformer allows to reproduce the variability of voltage conditions in the network and lead to the situation of exceeding the set voltage thresholds, which should result in activation of the protection. Electrical parameters, such as voltages at the inverter connection point and generated power, were recorded on a HIOKI 3196 power quality analyzer with a recording time resolution of 15 seconds.

Table 2. Overvoltage protection settings in the tested SMA inverter

Function	Setting	Time delay
Voltage increase protection	253.00 V	-
Lower maximum threshold	264.50 V	200 ms

To illustrate the operation of the 1st stage of overvoltage protection (delayed operation – based on the measurement of a 10-minute value), two test scenarios were selected. In the first scenario, the rated voltage was maintained for more than 10 minutes, and then the voltage was increased by means of an autotransformer to a value exceeding the protection activation threshold (approx. 258 V). After activation of the protection, the voltage was lowered again to the rated value so that the source could reconnect. The waveform of voltage and power generated in a microinstallation is shown in figures 1 and 2.

From the data recorded by the analyzer, one can read the protection tripping time and the generator reconnection time. The tripping time, measured from the moment of exceeding the threshold of 1.1 U_n (253 V) to the impulse for switching off the generation, was 7:45 minutes, while the time to full disconnection of the source (generated power equal to zero) was 8:15 minutes. The reconnection time, measured from the voltage drop below the 1.1 U_n threshold to the moment when the source generates power, was approx. 1:30-2:00 minutes.

The second test scenario is similar, but the value maintained before the protection is triggered is higher – instead of 230 V, the value was set to 237 V. After the time exceeding 10 minutes, this value was increased again to the value of approx. 258 V. If the protection algorithm complies with the standard and by the 10-minute average value, this protection should operate in a shorter time than in scenario 1. Recorded waveforms and designated times are shown in figure 3.

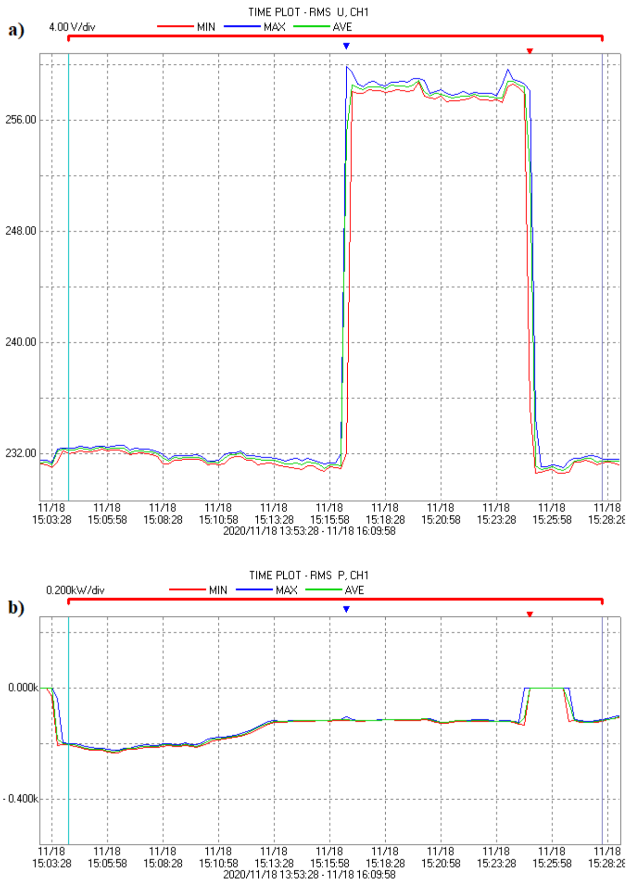


Fig. 1. Waveforms recorded in the analyzer for test scenario 1: a) measured voltage, b) generated active power (negative values mean power fed back to the network)

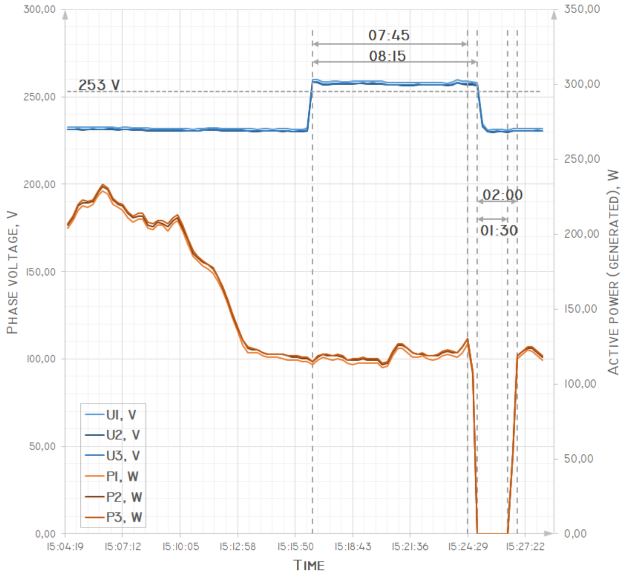


Fig. 2. Voltage and power waveforms for scenario 1 with marked test results – protection tripping time and source reconnection time

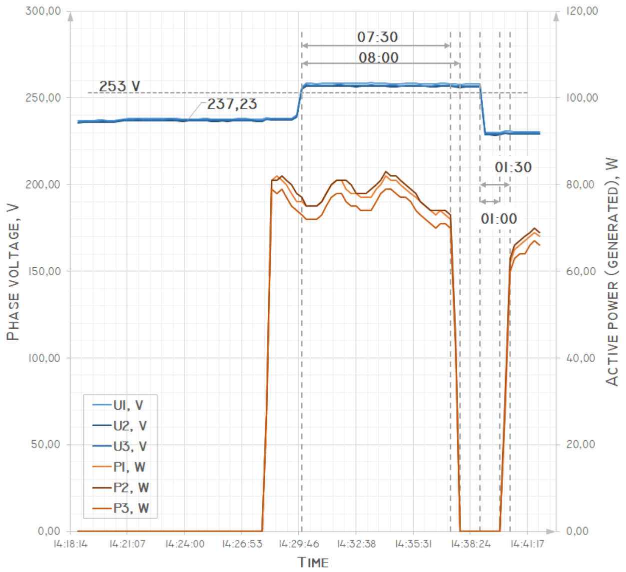


Fig. 3. Voltage and power waveforms for scenario 2 with marked test results – protection tripping time and source reconnection time

In scenario 2, as expected, the protection tripping times were shorter. The tripping time (from the moment of exceeding the set value of 1.1 U_n to the impulse for switching off the generation) was 7:30 minutes, while the time until the complete disconnection of the source was 8:00 minutes. The reconnection time, measured from the voltage drop below the 1.1 U_n threshold to the moment when the source generates power, was approx. 1:00–1:30 minutes. The results are burdened with a relatively large measurement error due to the recording time resolution of 15 seconds. The conducted tests – described above in the two scenarios and the others, not described in this paper – allowed to determine that the speed of the protection operation is influenced not only by the voltage value after exceeding the threshold of 1.1 U_n, but also by the preceding values, which may indicate that the protection responds to 10-minute mean value in a moving window, i.e. in accordance with the standard. Also, the second stage of protection turned out to be compliant with the requirements of the standard – after setting the autotransformer to a value exceeding 1.15 U_n, the protection immediately turned off the generator.

3. Simulation model of overvoltage protection of PV microinstallations

The creation of a fully functional protection model for the purposes of simulation is key to show various scenarios of operation, but above all to show the interaction of many such protection systems installed in the network with a large number of prosumer installations. The overvoltage relay model was made in the DIgSILENT PowerFactory software, using the DSL language (DIgSILENT Simulation Language) enabling modeling time continuous controls and processes.

The operation of the relay model is based on the measurement of instantaneous voltage values at the connection point (POC) of the PV source. These values are sampled at the specified sampling rate. From the thus obtained discrete signal, the RMS value is determined at 10-period intervals (0.2 seconds). This value is then aggregated for 10 minutes according to the standard [11] and the 10-minute average value is determined by the moving window method with a width of 600 seconds. The mean value is calculated as the square root of the arithmetic mean of the squared input values over 10 min. The correctness of the model was verified by trying to reproduce the conditions of the laboratory tests described in section 2. Therefore, the simulations considered a simple system consisting of an MV/LV transformer enabling the change of the voltage value in the network, and the low-voltage line supplied from it. A three-phase PV source with a rated power of 50 kW (cos φ = 1) with an overvoltage relay and a load of 50 kW was connected at the end of the line.

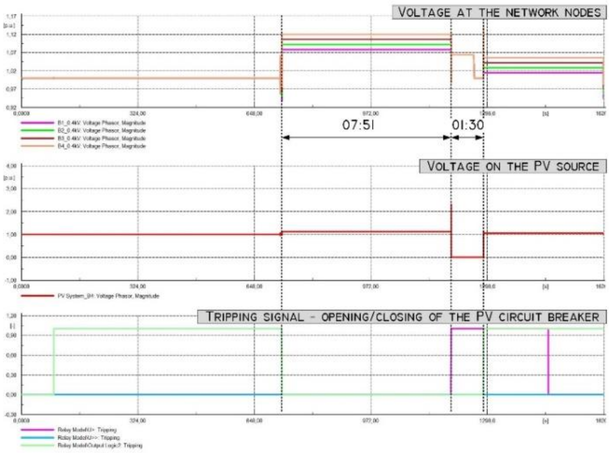


Fig. 4. Simulation results (scenario 1) – overview

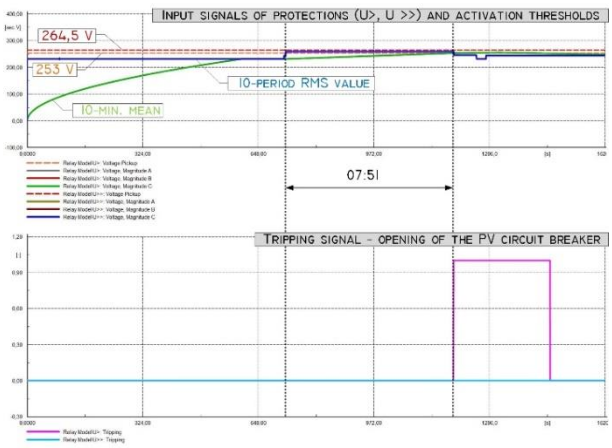


Fig. 5. Simulation results (scenario 1) – operation of overvoltage protection

The aim of the simulations was to recreate the laboratory test scenarios presented in section 2. Voltage changes at the point of connection were obtained by changing the transformer tap, changing the value of the generated power and the load power. The schedule of the two simulation scenarios is presented in the form of table (table 3).

Table 3. Simulation scenarios for tests of 1st stage overvoltage protection (U>)

Simulation time	Scenario 1	Scenario 2
0–12:00 min	The voltage at the POC is kept at 230 V	The voltage at the POC is kept at 237 V
12:00–12:05	Gradual increase in voltage to 258 V	Gradual increase in voltage to 258 V
Unknown	Protection stage 1 (U>) triggering	Protection stage 1 (U>) triggering
21:00 min	Lowering the voltage to the rated value	Lowering the voltage to the rated value
Unknown	Reconnection of microinstallations	Reconnection of microinstallations

The recorded waveforms for the simulation scenario 1 are shown in figures 4 and 5. The protection stage 1 was triggered at 19:56 minutes. Therefore, the tripping time, measured from exceeding the threshold of 1.1 U_n (at 12:05 min) to the occurrence of the signal for tripping the circuit breaker, was 7:51 minutes.

For scenario 2 of the simulation, the waveforms presented in figures 6 and 7 were recorded. The protection stage 1 was activated at 19:44 minutes. The tripping time, measured from exceeding the threshold of 1.1 U_n (at 12:05 min) to the occurrence of the signal for tripping the circuit breaker, was therefore 7:39 minutes. The reconnection time in the simulation resulted from the introduced delay setting (90 seconds) and was the same for both scenarios.

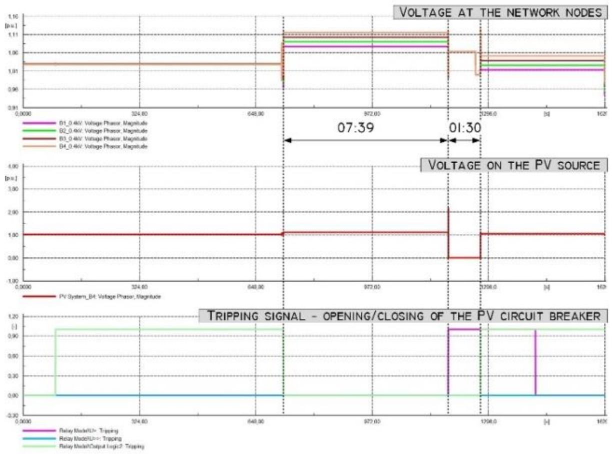


Fig. 6. Simulation results (scenario 2) – overview

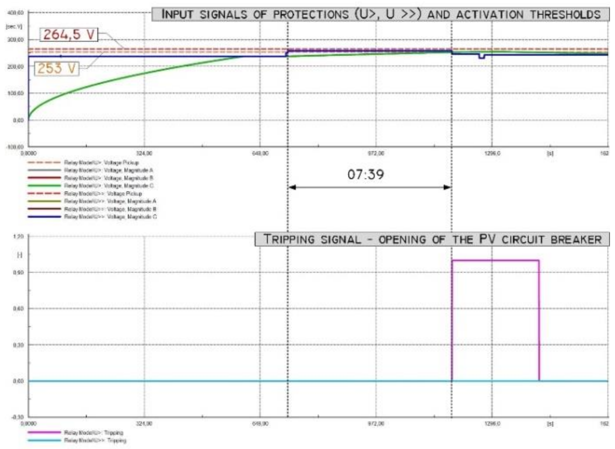


Fig. 7. Simulation results (scenario 2) – operation of overvoltage protection

Table 4. Comparison of the results of laboratory tests of the SMA inverter protection and simulation results on the protection model in PowerFactory

	Tripping time (U>) in the SMA inverter	Tripping time (U>) in the PowerFactory model
Scenario 1	7:45 min	7:51 min
Scenario 2	7:30 min	7:39 min

The simulations did not take into account additional delays in the operation of the protection occurring in the actual system. The comparison of the obtained tripping times of the protection – in the laboratory test and in the simulation – is included in table 4. The results confirm the compliance of the simulation model with the requirements of standards and regulations, as well as with the algorithm of protection operation in a typical inverter.

4. Summary

The paper presents the requirements of national regulations and standards regarding the operation of overvoltage protections integrated with the inverter of photovoltaic microinstallations. The functionality of this type of protection has been tested in laboratory tests, on the example of a typical inverter used in microinstallations connected to public low-voltage distribution networks. A simulation model was also built in the PowerFactory software, and its correctness was verified in the simulations. The results obtained in the simulations are consistent with the results obtained from the tests of the SMA inverter, therefore it can be concluded that the model complies with both the regulations and the actual operation of the overvoltage relay.

The simulation model will be used in further research aimed at indicating the potential negative effects of the operation of a greater number of such protections in networks with a large share of PV microinstallations. Lack of coordination of overvoltage protections (limiting the power delivered to the network to zero) may result in negative dynamic phenomena that threaten the stability of voltage conditions in the network. The introduction of a uniform system enabling the coordination of protections operations could prevent such phenomena. In addition, the introduction of such a system will certainly be facilitated by the provisions of the NC RfG code [6, 12], which impose on the manufacturers of inverters the need to ensure the possibility of remote disconnection of microinstallations above 10 kW. In order to prevent the necessity to disconnect microinstallations, it is advisable to first implement voltage regulation methods so as not to exceed the permissible limits [3].

References

[1] Appen J. et al.: Time in the sun: The challenge of high PV penetration in the German electric grid. IEEE Power and Energy Magazine 11(2), 2013, 55–64 [http://doi.org/10.1109/MPE.2012.2234407].
[2] Hashemi S., Østergaard J.: Methods and strategies for overvoltage prevention in low voltage distribution systems with PV. IET Renewable Power Generation 11(2), 2017, 55–64 [http://doi.org/10.1049/IET-RPG.2016.0277].
[3] Janiga K.: A review of voltage control strategies for low-voltage networks with high penetration of distributed generation. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska – IAPGOS 3, 2020, 60–65.
[4] Mahmud N., Zahedi A.: Review of control strategies for voltage regulation of the smart distribution network with high penetration of renewable distributed generation. Renewable and Sustainable Energy Reviews 64/2016, 582–595.
[5] Renewable Energy Sources Act of February 20/2015, Dz. U. 2015 poz. 478
[6] Commission Regulation (EU) 2016/631 of April 14, 2016 establishing the EU network code.
[7] Instruction of Distribution Network Operation and Maintenance (IRiESD), PGE Dystrybucja S.A.
[8] PN-EN 50160:2010 Parametry napięcia zasilającego w publicznych sieciach elektroenergetycznych.
[9] PN-EN 50549-1:2019-02 Wymagania dla instalacji wytwórczych przeznaczonych do równoległego przyłączania do publicznych sieci dystrybucyjnych.
[10] PN-EN 50438:2014-02 Wymagania dla instalacji mikrogeneracyjnych przeznaczonych do równoległego przyłączania do publicznych sieci dystrybucyjnych niskiego napięcia.
[11] PN-EN 61000-4-30: 2015-05 Kompatybilność elektromagnetyczna (EMC) - Część 4-30: Metody badań i pomiarów - Metody pomiaru jakości energii.
[12] Set of requirements for type A power generating modules, including microinstallations, PGE Dystrybucja S.A.

M.Sc. Klara Janiga
e-mail: k.janiga@pollub.pl

Klara Janiga received M.Sc. degree in electrical engineering (specializing in power engineering) from the Lublin University of Technology (LUT), Poland. She is currently an assistant in the Department of Power Engineering and Ph.D. student at LUT. Her research interest is concentrated on distributed generation, in particular on photovoltaic systems.

http://orcid.org/0000-0002-1798-0434

otrzymano/received: 8.05.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2657

DETERMINATION OF HYDRODYNAMIC PARAMETERS OF THE SEALED PRESSURE EXTRACTOR

Nataliaya Kosulina[1], Stanislav Kosulin[2], Kostiantyn Korshunov[1], Tetyana Nosova[3], Yana Nosova[3]

[1]Kharkiv Petro Vasylenko National Technical University of Agriculture, Department of Biomedical Engineering and Theoretical Electrical Engineering, Kharkiv, Ukraine, [2]Kharkiv Medical Academy of Postgraduate Education, Department of Oncological Surgery Ray Therapy and Palliative Help, Kharkiv, Ukraine, [3]Kharkiv National University of Radio Electronics, Biomedical Engineering Department, Kharkiv, Ukraine

Abstract. *The subject matter of the article: Sealed extractor with pressure. The goal of the work: determination of hydrodynamic parameters of the sealed extractor with pressure. As a result of the analysis of the technological process and equipment used in the factories for primary processing of wool, shortcomings and problems are identified and means for their elimination are proposed. It is proposed to use small-sized equipment to work on waste-free technology based on a hydrodynamic pressure extractor. Extraction as an efficient mass transfer process for removing organic components from aqueous solutions has the advantages of low operating temperatures and efficiency. The design features of the sealed pressure extractor are as follows: high angular velocities, the moment of inertia of rotating details, powerful pressure, the presence of nodes that provide a supply and discharge of liquids, tightness. The kinematic and geometric parameters of the rotor affect the sealed extractors' performance. In sealed extractors, the heavy fraction in the field of centrifugal forces will accumulate on a large radius of the inner side of the rotor and for its movement it is necessary to create an excess pressure at the extractor inlet.*

Keywords: water extractor, primary processing of wool, fluid flow hydrodynamics, rotor of extractor, angular velocity of rotation

OKREŚLANIE PARAMETRÓW HYDRODYNAMICZNYCH USZCZELNIONEGO EKSTRAKTORA

Streszczenie. *Przedmiotem badań przedstawionych w artykule jest ekstraktor ciśnieniowy. Celem pracy jest określenie jego parametrów hydrodynamicznych. W wyniku analizy procesu technologicznego i urządzeń stosowanych w fabrykach wstępnego przerobu wełny zidentyfikowano problemy i zaproponowano sposoby ich eliminacji. Proponuje się wykorzystanie małogabarytowych urządzeń do pracy w technologii bezodpadowej opartej na hydrodynamicznym ekstraktorze ciśnieniowym. Ekstrakcja jako skuteczny proces przenoszenia masy do usuwania składników organicznych z roztworów wodnych wykazuje zalety w postaci niskich temperatur roboczych i efektywności. Cechy konstrukcyjne ciśnieniowego ekstraktora ciśnieniowego to duże prędkości kątowe, moment bezwładności obracających się części, wysokie ciśnienia, obecność węzłów zapewniających dopływ i odprowadzanie cieczy oraz szczelność. W ekstraktorach ciśnieniowych na wydajność wpływają parametry kinematyczne i geometryczne wirnika. W zamkniętych ekstraktorach frakcja ciężka w polu sił odśrodkowych będzie się gromadzić na dużym promieniu wewnętrznej płaszczyzny wirnika i do jej przemieszczenia konieczne jest wytworzenie nadciśnienia na wlocie do ekstraktora.*

Słowa kluczowe: ekstraktor wody, pierwotna obróbka wełny, hydrodynamika przepływu płynów, wirnik ekstraktora, kątowa prędkość obrotowa

Introduction

Analysis of the work of the textile industry in recent years gives grounds to conclude that the decline in production continues, and the economic conditions of the operation of enterprises do not allow stopping the decline in production [17].

The technical level of the equipment installed at the factories for the primary processing of wool does not meet modern requirements, and the engineering industry of Ukraine does not produce machines and equipment for this industry [16].

Analysis of the technological process and equipment used for the primary processing of wool revealed the following significant disadvantages [6]:

- negative impact on the environment;
- energy intensity of production;
- material consumption;
- large overall and mass characteristics;
- poor fiber quality.

To eliminate these shortcomings, ways are outlined that, in our opinion, allow solving the listed problems, and to do this it is necessary:

- to develop a new waste-free technology for the primary processing of wool;
- to develop small-sized equipment for work on a waste-free technology based on a hydrodynamic pressure extractor with automated control from a computer/ACS TP

1. Analysis of the latest achievements and publications

The large number of various contaminants in industrial wastewater determines the numerous methods, methods, techniques and technological schemes used in their purification [5, 7, 18, 19].

The existing cleaning methods can be divided into five groups: mechanical group which involves averaging, aeration, settling, sedimentation in hydrocyclones, filtering, filtration; chemical (reagent) group which contains neutralization and filtration through active loading, coagulation, chemical precipitation and coprecipitation of harmful substances, oxidation and reduction; physical-chemical group which consists of extraction, evaporation, adsorption, ion exchange, electrocoagulation and electrolysis, flotation, crystallization; biological group which involves aerobic oxidation in bio-filters and in aeration tanks, anaerobic digestion in digesters; thermal group which involves vaporization and evaporation, waste water incineration, sludge drying incineration.

The considered methods, in turn, are divided into regenerative and destructive ones. Regenerative methods include most of the physicochemical methods, such as extraction, evaporation, adsorption, ion exchange, electrolysis, crystallization. The rest of the methods are essentially destructive, however, in some specific cases, they can be used to carry out regenerative treatment, that is, to extract useful substances from wastewater. A number of methods can be classified as widely used in the treatment of industrial wastewater. These are, first of all, settling, filtering, averaging, neutralization, chemical precipitation, biological treatment [6, 8, 19].

The practice of recent years has included the method of flotation for the purification of waste- water from the primary processing of wool [2, 12, 20]. The treatment of wastewater by the flotation method has a number of advantages, which predetermines the prospects of this method and the possibility of its use for the purification of both industrial and domestic wastewater [1].

The disadvantages of the reagent-floatation technology include the need to implement the flotation process at low pH values up to 4.2 ... 4.5, it requires the use of acid-resistant equipment not only in the treatment facilities block, but also in the sludge treatment block [11].

Analysis of the literature and the results of our own research showed that the most acceptable method of wastewater treatment with their return to the circulating system can be the use of centrifugation and electro-flotation [10, 14].

The main stages of the electro-flotocoagulation process are:
1) Generation of an electrolytic coagulant;
2) Electrolytic gas generation;
3) Coagulation of impurities;
4) Fixation of electrolytic gas on the surface of coagulated particles (formation of flotation complexes) [13, 15].

A significant disadvantage of electro-flotocoagulation is the processes of passivation of the surface of soluble electrodes and the adsorption of impurities on them.

Therefore, maintaining the optimal current density is a prerequisite for the operation of the electro-coagulator [9].

The performed analysis shows that the degree of purification depends on many factors, the main of which are: current density, processing time, height of the liquid layer, temperature of the liquid, and the degree of dissociation of water into ions, estimated by the pH value.

The design features of the sealed pressure extractor are high angular velocities, the moment of inertia of rotating parts, high pressures, the presence of nodes that ensure the supply and removal of liquids, and tightness. Automated control is assumed by means of a computer/ACS TP.

2. Main research material

The purpose of the work is to determination of hydrodynamic parameters of the sealed pressure extractor.

The main elements of equipment for the primary processing of wool (PPW) are: a unit for washing and pressing wool, equipment for water regeneration and fat release, a drying unit, an automated control system for complex washing of wool. Fig. 1 shows the block diagram of the primary processing of wool.

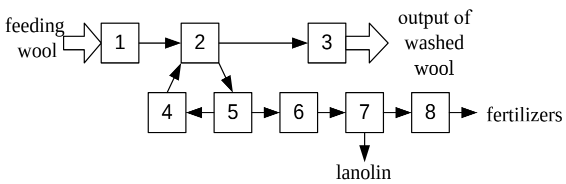


Fig. 1. Block diagram of the primary processing of wool: 1 – dispenser; 2 – washing machine; 3 – drying machine; 4 – water preparation tank; 5 – waste water regenerator; 6 – fat accumulator; 7 – separator; 8 – fertilizer drying chamber

The process of wool processing is automated by the automated process control system. To regulate the pressure of the fractions, select the rotor speed, take into account design features and obtain the best performance, it is necessary to understand the processes taking place directly in the sealed pressure extractor.

The hydrodynamics of the flow of two immiscible liquids through the extractor rotor equipped with perforated cylinders is not fully studied. Experimental work made it possible to establish only a qualitative picture of the hydrodynamic process and the influence of individual factors on the nature of its course.

When the fluid rotates together with the rotor, the pressure at any point located inside the working volume is determined by the angular velocity of rotation of ω rotor, the volumetric weight of the liquid γ and the radius of the location of this point R_i from the axis of rotation [6, 8]. The calculation formula for determining the pressure will be as follows:

$$P_i = \frac{\gamma\omega^2 R_i^2}{2g} \qquad (1)$$

If the inner cavity of a rotating rotor is limited by cylinders with radii R_n and R_k, then the calculation is performed according to formula 2:

$$P_i = \frac{\gamma\omega^2 \left(R_k^2 - R_n^2\right)}{2g} \qquad (2)$$

To determine the hydrodynamics during the rotation of the rotor, a program (GIDRAVL-1) was developed. Figures 2 and 3 present graphs showing the relationship between rotor diameter, rotor speed and developed pressure.

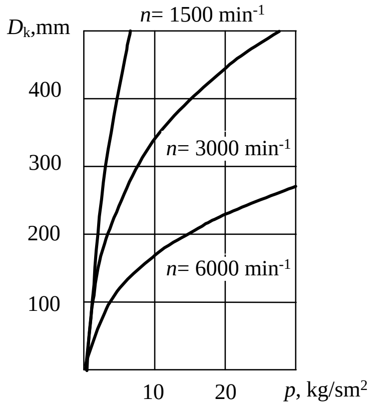


Fig. 2. Relationship between rotor diameter and developed pressure

The rotor of the extractor is cylindrical. With continuous feeding of the extractor with heavy and light fractions, counter axial flows are created in the rotor, closed in the annular space between the rotor cylinders.

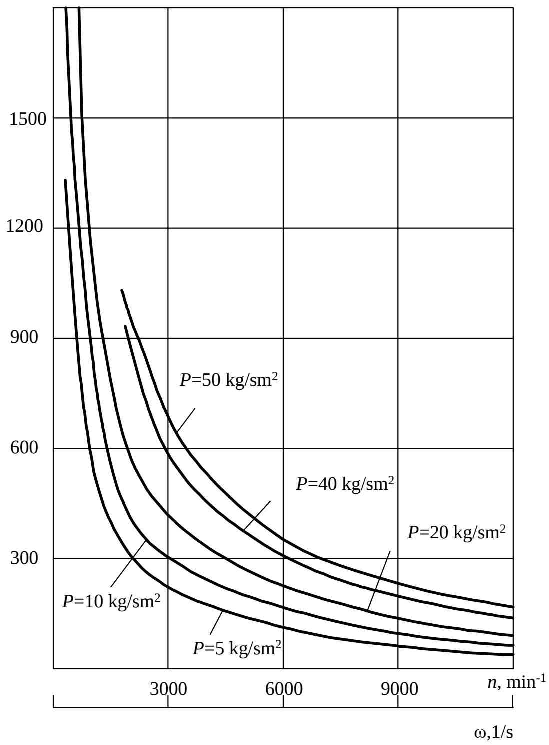


Fig. 3. The relationship between the diameter of the rotor, the frequency of its rotation and the developed pressure

Let us apply the Bernoulli equation for a cylindrical flow in the field of centrifugal forces. Its form is for the case when the liquid is only under the influence of centrifugal forces:

$$\frac{\omega^2\left(R^2 - r^2\right)}{2} + \frac{p}{\gamma} + \frac{V^2}{2} = \text{const} \qquad (3)$$

Consider the movement of fluid from one section to another. Fig. 4 shows a section of flow 1 – 2 in a centrifugal field.

The equation for the real fluid flow along the rotor cylinders will be:

$$\frac{\omega^2\left(R^2-r_1^2\right)}{2}+\frac{p_1}{\lambda}+\frac{\alpha_1V_1^2}{2}=\frac{\omega^2\left(R^2-r_2^2\right)}{2}+\frac{p_2}{\lambda}+\frac{\alpha_2V_2^2}{2}+h, \quad (4)$$

where h is the loss of specific energy of the flow between the considered sections; α are Coriolis coefficients in the considered sections.

Based on the above equations, we obtain the formula for the liquid outflow from a rotating cylinder.

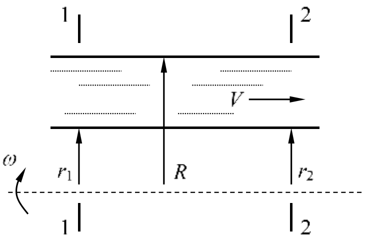


Fig. 4. Section of flow 1 – 2 in a centrifugal field

Let a liquid flow out of a rotating cylinder through a hole in the wall under a constant pressure.

To determine the rate of fluid outflow, we use equation (4), composing it for cylindrical surfaces of radii R (outer) and r_0 (inner) fluid layer. Then, neglecting the magnitude of the velocity on the free surface of the liquid and taking the Coriolis coefficient equal to unity, and also assuming that the pressure in the sections is equal to atmospheric, we obtain:

$$\frac{\omega^2\left(R^2-r_0^2\right)}{2}=\frac{V^2}{2} \quad (5)$$

where from

$$V=\omega\sqrt{R^2-r_0^2} \quad (6)$$

If the rotating cylinder is completely filled with liquid, then:

$$V=\omega R=U_0 \quad (7)$$

Thus, when the cylinder is filled, the rate of fluid outflow through the hole is equal to the circumferential speed of rotation of the cylinder.

We accept the following assumptions: p_2 is equal to zero, p_1 is not equal to zero, $\alpha = 1$,

$$\psi=\frac{h}{p} \quad (8)$$

To determine the flow rate through the hole in the cylinder, we use the formula:

$$W=\varphi\varepsilon S\sqrt{\frac{2p(1-\psi)}{\gamma}+\omega^2\left(R^2-r_0^2\right)} \quad (9)$$

where S is the cross-sectional area of the hole, φ is the velocity coefficient, ε is the compression coefficient of the jet.

The average value of the potential and kinetic energy per unit mass of the liquid flowing through the flow section, referred to the conditional surface, is called the specific energy. The specific energy in a rotating cylinder is expressed by the equation:

$$E_p=\frac{1}{2}\omega^2\left(R^2-r_0^2\right)+\frac{p}{\gamma}+\frac{\alpha V^2}{2} \quad (10)$$

Neglecting atmospheric pressure, we obtain the following expression for the specific energy referred to the inner surface of a rotating cylindrical rotor:

$$E_0=\frac{1}{2}\omega^2\left(R^2-r_0^2\right)+\frac{\alpha V^2}{2} \quad (11)$$

This is the specific energy of the section.

The volumetric velocity through the entire cross-section of the annular flow:

$$V=\frac{\pi\left(p_1-p_2\right)}{\mu L}\int_r^R\left(\frac{1}{2}\left(R^2-r^2\right)+r_0^2\ln\frac{r}{R}\right)rdr \quad (12)$$

where μ is the dynamic coefficient of viscosity (nsec/m^2).

After integration, transformations and reductions, we find the volumetric velocity (m^3/sec):

$$V=\frac{\pi\left(p_1-p_2\right)}{\mu L}\left(\frac{R^4}{4}+\frac{3}{4}r_0^4-R^2r_0^2-r_0^4\ln\frac{r_0}{R}\right) \quad (13)$$

3. Conclusions

The above calculation formulas describe the processes occurring in sealed extractors. The performance in these units is influenced by the kinematic and geometric parameters of the rotor. In sealed extractors, the heavy fraction in the field of centrifugal forces will accumulate on a large radius of the inner plane of the rotor and for its movement it is necessary to create excess pressure at the inlet of the extractor.

The results obtained make it possible to accurately compose algorithms for the operation of a computer/ACS TP that control the process of cleaning wool and automatic control systems for the sealed pressure extractors. As well as supervisory control and data acquisition (SCADA) systems, emergency protection systems (ESD) and other smaller control systems (for example, systems based on programmable logic controllers (PLC)) [3, 4]. The automated process control system of the wool cleaning process has a unified operator control system for the technological process in the form of several control panels, means for processing and archiving information about the process, typical automation elements: sensors, control devices, and actuators [21, 22]. Industrial networks are used for information communication of all subsystems.

In the ACS TP system of the water purification process after wool washing, there was a consistent increase in the complexity of the tasks facing the control systems from the control of individual sealed pressure extractors and parameters, to the automation of the process as a whole, and the automation of control systems. This makes it possible not only to carry out effectively management and control in the production area, but also to eliminate partially the influence of the human factor in management, which helps to avoid mistakes. Currently, the issues of increasing the autonomy of the ACS TP of the water purification process after wool washing, the redistribution of functions in the direction of increasing the load in decision-making at the ACS are relevant. In this case, the issues of developing the intellectual component of the automated process control system in the direction of creating algorithms for responding in real time to emerging critical situations during water purification in the process of washing wool are relevant.

References

[1] Aksenko A.A. et al.: A. c. 914506 (USSR). Apparatus for electrochemical purification of contaminated liquid. B.I. 11/1982.

[2] Arterchuk A. G., Burchenko G. M., Zhurkov V. S.: Purification of oil-emulsion waste water from machine-building enterprises by the method of electrochemical coagulation. Inform. leaflet 102, 1974.

[3] Avrunin O. et al.: Development of up-to-date laboratory base for microprocessor systems investigation. 2009 19th International Crimean Conference Microwave & Telecommunication Technology, Sevastopol 2009, 301–302.

[4] Avrunin O. G. et al.: Experience of Developing a Laboratory Base for the Study of Modern Microprocessor Systems. Proceedings of I International Scientific and Practical Conference "Theoretical and Applied Aspects of Device Development on Microcontrollers and FPGAs" MC&FPGA-2019, Kharkiv, Ukraine, 2019, 6–8.

[5] Bateup B. O.: Clean green wool. Top-Tech '96: a CSIRO conference, Geelong 1996, 93–95.

[6] Demidov A. V.: Development and research of a roller device for increasing the efficiency of wool spinning after washing. Autoreferral of the doctoral thesis. Ivanovo 2005.

[7] Lupikov V. S.: Obrupting the method of measuring electrical penetration for drying with an electro-magnetic field. Bulletin of the National Technical University "Kharkiv Polytechnic Institute" 4, 2011, 131–134.
[8] Muntyan V.A.: Analysis of technological processes and devices for primary processing of wool. Energy saving, energetics, energy audit 1(71), 2010, 62–65.
[9] Mykhaylova L. M. et al.: Acoustic vibrations hydrodynamic emitter parameters determination. Telecommunications and Radio Engineering 79(3), 2020, 231–248 [https://doi.org/10.1615/TelecomRadEng.v79.i3.50].
[10] Nazaryan M. M. et al.: A. c. 710988 (USSR). Apparatus for electrochemical cleaning of contaminated liquid. B.I. 3/1980.
[11] Nazaryan M. M. et al.: A. c. 644738 (USSR). Apparatus for electrochemical cleaning of contaminated liquid. B.I. 4/1979.
[12] Nazaryan M. M., Efimov V. T., Mataev A. R. et al.: A. c. 555056 (USSR). Method of automatic regulation of electrochemical wastewater treatment processes. B.I. 15/1977.
[13] Nazaryan M. M. et al.: A. c. 688445 (USSR). Apparatus for electrochemical cleaning of contaminated liquids. B.I. 36/1979.
[14] Nazaryan M. M. et al.: A. c. 827408 (USSR). Apparatus for electrochemical cleaning of contaminated liquid. B.I. 17/1981.
[15] Nazaryan M. M. et al.: A. c. 899488 (USSR). Method of automatic regulation of processes of electrochemical purification of contaminated liquid. B.I. 3/1982.
[16] Potapskiy P. V., Mikhailova L. N.: Analysis of the mathematical model of heating wool in bales by electromagnetic energy. Eastern European Journal of Advanced Technologies 86/2009, 115–119.
[17] Rogachev N. V., Vasilyeva L. G., Timoshenko N. K. et al.: Wool. Primary processing and market. Editors: N. K. Timoshenko. VNIIMPRASKHN, Moscow 2000.
[18] Rogachev N. V.: New in the technology of primary processing of wool. Sat. Technologies for primary processing of wool. Nevinnomyssk 1990, 17–27.
[19] Rogachev N. V.: Waste water treatment and wool modernization. Textile industry 21, 1976.
[20] Semenets V. et al.: Determination of parameters of plane hydrodynamic radiator of acoustic vibrations. Radiotekhnika: All-Ukr. Sci. Interdep. Mag. 196, 2019, 167–179.
[21] Semenets V. et al.: Trends in Training Modern Technicians. First International Scientific and Practical Conference "Theoretical and Applied Aspects of Device Development on Microcontrollers and FPGAs" MC&FPGA-2019, Kharkiv, 2019 [https://doi.org/10.35598/mcfpga.2019.013].
[22] Semenets V.: Technical aspects for development laboratory base for learning FPGA and microcontroller systems. 10th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics, Lviv-Polyana, Ukraine, 2009.

D.Sc. Nataliaya Kosulina
e-mail: kosnatgen@ukr.net

Doctor of Technical Sciences, Professor, Department of Biomedical Engineering and Theoretical Electrical Engineering, Kharkiv Petro Vasylenko National Technical University of Agriculture, Head of Department, Ukraine. Research interests: theoretical fundamentals of electrical engineering, electromagnetic field theory, methodology of scientific research, biomedical engineering, information electrical technologies in agrarian and industrial complex.

http://orcid.org/0000-0003-4055-8087

M.Sc. Stanislav Kosulin
e-mail: kosulinmd@gmail.com

Assistant department of Oncological surgery ray therapy and palliative help, KhMAPE (Kharkiv Medical Academy of Postgraduate Education), Ukraine. Research interests: electromagnetic field theory, biomedical engineering, information electrical technologies in agrarian and industrial complex.

http://orcid.org/0000-0003-0791-0034

M.Sc. Kostiantyn Korshunov
e-mail: hkostn@ukr.net

Postgraduate student, Department of Biomedical Engineering and Theoretical Electrical Engineering, Kharkiv Petro Vasylenko National Technical University of Agriculture, Ukraine. Research interests: electromagnetic field theory, biomedical engineering, information electrical technologies in agrarian and industrial complex.

http://orcid.org/0000-0002-4993-3800

Ph.D. Tetyana Nosova
e-mail: gryshkov@imp.uni-hannover.de

Ph,D,, associate professor, Kharkiv National University of Radio Electronics, Biomedical Engineering Department, Kharkiv, Ukraine. Research Interests: biomedical signal processing and images, functional diagnostics of musculoskeletal system.

http://orcid.org/0000-0003-4442-8001

Ph.D. Yana Nosova
e-mail: yana.nosova@nure.ua

PhD, Senior Lecturer of the Department of Biomedical Engineering, Kharkiv National University of Radio Electronics, Ukraine. As part of the team of authors of the patent "Device for the testing of respiratory disorders of smell" was awarded the diploma of the winner of the All-Ukrainian contest "Invention of the Year – 2016" in the nomination "Best invention in Kharkiv region". Research Interests: biomedical signal processing and images.

http://orcid.org/0000-0003-4310-5833

otrzymano/received: 8.05.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2658

DEVELOPMENT OF A DEVICE FOR MEASURING AND ANALYZING VIBRATIONS

Anzhelika Stakhova, Volodymyr Kvasnikov
National Aviation University, Department of Computerized Electrical Systems and Technologies, Kiev, Ukraine

Abstract. *It is known that vibration is generated to one degree or another by all moving parts of machines. Vibration processes arising in the process of functioning of machines and equipment are highly informative, reflecting the technical condition of many parts and assemblies quite fully. Therefore, the use of vibration diagnostics systems makes it possible to determine a defect even at the stage of inception, thereby excluding emergency stops of machinery. An urgent task is to create a vibration analysis device, which consists of a sensor for measuring vibration, the principle of operation of which is based on the use of a direct piezoelectric effect. The output signal of the sensor will go to a personal computer or any other device with the Windows operating system, where a spectral analysis of the measured vibration signal is carried out using a Matlab-based diagnostic program.*

Keywords: up vibrations, defect, vibroacoustic parameters, vibration measurement

OPRACOWANIE URZĄDZENIA DO POMIARU I ANALIZY DRGAŃ

Streszczenie. *Wiadomo, że wibracje w mniejszym lub większym stopniu są generowane przez wszystkie ruchome części maszyn. Procesy wibracyjne powstające w trakcie pracy maszyn i urządzeń mają bardzo bogaty charakter informacyjny, w pełni odzwierciedlając stan techniczny wielu części i zespołów. Dlatego zastosowanie systemów diagnostyki drganiowej pozwala na określenie usterki już na etapie jej powstania, wykluczając tym samym awaryjne zatrzymania maszyn. Pilnym zadaniem jest stworzenie urządzenia do analizy drgań, które składa się z czujnika do pomiaru drgań, którego zasada działania opiera się na wykorzystaniu bezpośredniego efektu piezoelektrycznego. Sygnał wyjściowy czujnika trafia do komputera osobistego lub innego urządzenia z systemem operacyjnym Windows, gdzie analiza widmowa mierzonego sygnału drgań jest przeprowadzana za pomocą programu diagnostycznego opartego na Matlabie.*

Słowa kluczowe: drgania, wady, parametry wibroakustyczne, pomiar drgań

Introduction

Modern development of technological equipment has led to the fact that machines and other devices of heavy construction have become an integral part of human everyday life. Continuous improvement of technical equipment accelerates the pace of work, as well as the load at which modern devices work. As a result, the parts used in these machines wear out faster. And failure of one part of the device can lead to breakage of other part or and, even, all mechanism.

Often such breakdowns are observed, after which the equipment becomes unsuitable for repair. It becomes necessary to replace equipment parts, and in some cases, even have to replace the entire mechanism. But modern equipment is equipped with the latest technological solutions, has a very high cost and expensive maintenance. Therefore, the task of ensuring the reliability of the equipment used, as well as to attract the latest technologies for diagnosing equipment becomes urgent.

Therefore, the equipment of each enterprise requires periodic inspection. This will prevent premature failure of the equipment, to predict the location of the defect before it is formed. Replace a part of a certain part of the machine or take the necessary measures to stop the development of the defect. It should be noted that the reason for the failure of the machine is the use of parts in which there was a defect during manufacture.

There are many methods of diagnosing equipment and one of the methods of modern diagnostics is vibration diagnostics. Vibration diagnostics is now widely used [2, 3, 8, 9], since vibration processes sufficiently fully reflect the technical condition of many parts and assemblies. Vibration diagnostic methods are aimed at detecting and identifying such malfunctions of the unit that affect its vibration.

It uses vibroacoustic processes in her work, which have a very large amount of information about the operation of the machine, about the condition of its parts. Thus at processing of the received signals the modern technological apparatus which allows to carry out inspection during work of the equipment, without its dismantling is used. Methods such as temperature control, lubricant analysis and others, with the right approach, are practically not required - they are replaced by vibration analysis. In addition, modern methods of information processing are used. These diagnostic methods are the most sensitive to changes in the vibration properties of the object of diagnosis under the influence of defects.

The introduction of modern computer technology into production makes it possible to process large volumes of measurement data, which significantly contributes to the development of modern methods of equipment maintenance with the search for new informative parameters, including maintenance based on the technical (actual) state [12]. It remains an urgent task to create a device for measuring and analyzing vibrations, with the search for a diagnostic parameter, which will be sensitive to the point of becoming more sensitive to the point of view of the reliability of control with a further elaborate diagnostic program based on Matlab.

1. Diagnostic parameters

Information about the status of the details of the object carries a certain signal. It can be of different physical nature, the main thing is that it contains information about the state of the object under study and is easy to read by a certain device. On the way from the formation of the defect to its receipt in a certain diagnostic apparatus, the signal undergoes many numerical transformations. As a result, its shape and physical nature change, but the information transmitted by this signal must remain unchanged. Otherwise, we will receive a signal that will carry false information about the state of the object. Such a signal will be of no use to us.

The defect can be detected and predicted in the early stages of its development. The spectrum and shape of the vibration signal stores information about the presence of characteristic defects of the rolling bearings. This information has specific features depending on the type of defect and the place where it was found. Such information is manifested by some features in the signal. One of such signs is the presence of asynchronous peaks in the spectrum. In addition, with developed defects, harmonics of such asynchronous peaks can be observed. The spectrum can contain both discrete peaks and blurred "humps" in which vibration energy is concentrated.

Let's move on to consider the signal itself, its parameters and the principles of its processing.

The signal that comes to a particular diagnostic apparatus is the oscillation of the object being monitored. Vibrodiagnostic is based on the measurement, analysis and processing of vibrations that occur during operation. Rolling and plain bearings can be a source of vibration in all machines.

Vibration is not always the same. It can be deterministic or random. Accordingly, deterministic vibration is divided into periodic and non-periodic, and random - into stationary and non-stationary. Periodic deterministic vibrations are harmonic and polyharmonic [9].

Periodic vibration is characterized by the period of oscillations, frequency, frequency range and angular frequency.

Harmonic vibrations can be recorded by the sine or cosine function. Their argument is the phase of harmonic oscillations, which depends on the initial phase, time and frequency. Depending on the magnitude of the phase shift, harmonic oscillations are divided into in-phase and out-of-phase.

Polyharmonic vibrations occur when the same system is affected by different independent harmonic oscillations in the general case with different amplitudes and initial phase angles. The shape of polyharmonic oscillations is greatly influenced by the shift of the initial phases between harmonics. Beats can occur with such oscillations.

Harmonic analysis of periodic (polyharmonic) vibration is its representation in the form of the sum of harmonic oscillations, in which periodic oscillations are represented by a Fourier series and non-periodic – by the Fourier integral.

Values that characterize the intensity of periodic, including harmonic oscillations:

- peak values (largest and smallest values of the oscillating value);
- amplitude of oscillations (sum of absolute values of the largest and smallest peak values of the oscillating value);
- mean values: arithmetic mean and rms.

Vibroacoustic signals are stochastic oscillations, so when building their mathematical models using the methods of the theory of random processes. Currently, the most common model of vibroacoustic signals are stationary random processes with a discrete spectrum, which in the General case can be represented in a complex form in the following form:

$$\xi(t) = \sum_{k=-\infty}^{\infty} \gamma_k e^{i2\pi fk} \tag{1}$$

where γ_k - independent equally distributed complex random variables in which the mathematical expectation $M[\gamma_k]=0$, and the variance $D[\gamma_k]=\sigma_k^2$.

The correlation function of process (1) is an almost periodic function and is equal

$$R(\tau) = \sum_{k=-\infty}^{\infty} \sigma_k^2 e^{i2\pi fk\tau} . \tag{2}$$

From formula (2) it is seen that the variance of process (1) is equal

$$D[\xi(t)] = R(0) = \sum_{k=-\infty}^{\infty} \sigma_k^2 . \tag{3}$$

The variance σ_k^2 of random amplitudes γ_k of harmonic components with frequencies f_k is called the spectrum of a random process (1).

The main statistical parameters for determining the state of the bearing are: mathematical expectation $M\xi(\tau)$, variance $D\xi(\tau)$, correlation coefficient ρ, asymmetry coefficient γ_3 and excess coefficient γ_4.

Mathematical expectation, average value – one of the main numerical characteristics of each random variable. It is a generalized concept of the average value of a set of numbers in the case when the elements of the set of values of this set have different "weight", price, importance, priority, which is characteristic of the values of a random variable [10].

The variance is a measure of the deviation of the values of a random variable from the center of distribution. Larger values of the variance indicate larger deviations of the values of the random variable from the center of distribution.

Correlation coefficient is an indicator used to measure the density of the relationship between performance and factor characteristics in the correlation-regression model with a linear relationship.

The asymmetry coefficient is a numerical characteristic of the probability distribution of a real random variable:

$$k = \mu_3 / \sigma^3 \tag{4}$$

The asymmetry k of the theoretical distribution of probabilities of a random variable is the ratio of the central moment of the third order μ_3 to the cube of the standard deviation σ^3.

The coefficient of excess is a numerical characteristic of the probability distribution of a real random variable. The coefficient of excess characterizes the "steepness", i.e., the rate of increase of the distribution curve compared to the normal curve. The excess γ of the theoretical distribution is called the characteristic calculated by the following formula:

$$\gamma_4 = \frac{\mu_4}{\sigma^4} - 3 \tag{5}$$

where μ_4 – the central moment of the fourth order, σ^2 – the variance [6].

An important task is to choose the diagnostic parameter that would be most sensitive to changes in the state of the object and determine the high reliability of control. In the problems of vibration diagnostics of bearing units of electric machines, such a parameter is the excess coefficient of diagnostic signals, the use of which allows to determine such defects as lack of lubrication, skew, defect of the inner ring.

The excess method can also be used to study the condition of bearings. Excess is a diagnostic parameter that characterizes the deviation of the probability density of instantaneous vibration values from the normal distribution (Gaussian distribution). Excess is characterized by the coefficient of excess (5).

When micro- and macro-shocks appear in the bearing, the probability density curve takes on a more acute nature and the value of the excess coefficient can be used to judge the degree of development of the defect:

- $\gamma_k < 3$ - corresponds to a serviceable condition of the bearing;
- $\gamma_k > 3$ - admissible operation of the bearing, but with its fast replacement;
- $\gamma_k < 5$ - inadmissible operation of the bearing.

The method of excess is not sensitive to the speed of rotation and load of the bearing, and does not require knowledge of its geometric parameters. Also, the excess coefficient changes its value not only when shock pulses appear in the bearing, but also increases with the deterioration of the oil. The last fact allows to use this method and for sliding bearings. [9].

In probability theory and statistics, Student's t-distribution or t-distribution is a type of probability distribution that arises in the problem of estimating the expected value of a normally distributed population when the sample size is small. This distribution is the basis of the popular Student's t-test of statistical significance the difference between the mathematical expectations of the two samples, and the confidence interval of the difference between the expected values of the two samples.

The distribution density of the Student's t-distribution has a distribution density function given by the formula

$$p(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} . \tag{6}$$

where ν is the number of degrees of freedom, $\Gamma(x)$ is the gamma function [7].

2. Device structure

The main content of vibrometry is the measurement of vibration parameters. Vibration of bodies is one of the most complex oscillatory processes, if we bear in mind all the degrees of freedom of the body and a wide range of oscillations, which can be non-periodic. These include steady-state, random, impulsive oscillations, etc. Vibration parameters: amplitude, frequency, peak-to-peak and a number of others - are measured using special off-the-shelf measuring devices and systems.

The general diagram of a device for measuring and analyzing vibration contains two main blocks (Figure 1): a vibration transducer and an electronic processing unit. The functional purpose of the first unit is to convert mechanical vibrations into an electrical signal. The principle of operation of piezoelectric transducers is based on the use of direct or reverse piezoelectric effects. The direct piezoelectric effect is the ability of some materials to form electric charges on the surface when a mechanical load is applied, the opposite is a change in the mechanical stress or geometric dimensions of the material sample under the influence of an electric field. As piezoelectric materials, natural materials are usually used – quartz and tourmaline, as well as artificially polarized ceramics based on barium titanate ($BaTiO_3$), lead titanate ($PbTiO_3$) and lead zirconate ($PbZrO_3$).

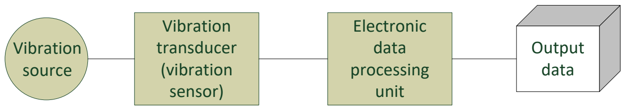


Fig. 1. Block diagram of the device

The second unit – the electronic processing unit – serves to "decode" the received signal. As a rule, an analog-to-digital converter is installed at the input of such blocks, and the main part of operations on the signal is performed already in digital form, which expands the functionality of the post-processing process, improves noise immunity and allows information to be output via an external interface.

When used in production, stationary vibrometers can be part of control systems as feedback sensors; for these purposes, some models of vibrometers have an analog output signal (usually voltage).

To obtain a comprehensive characteristic of the vibration process, a spectrum analyzer can be added to the measuring system. If the spectrum analyzer is multichannel, it can serve as the basis for a distributed vibration diagnostics system containing more than one vibration sensor.

The sensor is made on the basis of a piezoelectric sensing element [11] and is an absolute vibration transducer, that is, vibration is converted directly into electrical vibrations. The piezoelectric element is characterized by its resistance to overloads and a high natural frequency (from 30 to 50 kHz), therefore sensors with piezoelectric transducers are especially suitable for high-frequency processes in the range up to 20 kHz. The lower limiting frequency of such converters lies in the range of 0.5–10 Hz. The sensor output signal is proportional to vibration acceleration, the maximum value of which is from 1 to 10 m/s.

The standard version of the piezoelectric element consists of a shielded metal disk on which a ceramic membrane is located. To connect to measuring equipment, leads are used that are connected to the thrust bearing. The principle of operation of the elements is based on measuring the capacity of the membrane. Signal transmission is carried out through the thrust bearing. The difference in discharges is recorded in the measuring instruments. Data can be processed and stored via terminals on the sensors.

The general view of the sensor is shown in Figure 2. The sensor is made in a sealed plastic case of increased strength. The cable is connected to the vibration transducer, on the one hand, through the terminals connected to the thrust bearing. On the other hand, the cable is connected to the sound card of a personal computer via a 3.5 mini jack connector or a USB connector.

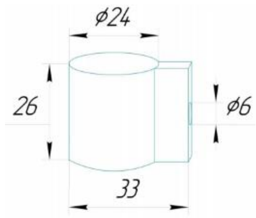


Fig. 2. General view of the sensor with linear dimensions

The sensor is installed on the measured object using an iron cobalt magnet, which is more resistant to mechanical stress, demagnetization and high temperatures than ceramic and neodymium magnets. The location of the magnet inside the insulated housing excludes its separation from the surface of the sensor, which greatly facilitates the process of measuring the vibration parameters.

The sensor is installed using a magnetic mount in the following sequence:

a) the sensor with a magnetic mount is installed on the object measurements;
b) the sensor is connected to the sound card of the personal computer using a cable.

During installation, make sure that the sensor is firmly attached (does not swing) on the measurement object. Otherwise, the measurement results may be inaccurate. Since the cable wire is shielded, it is allowed to cross the cable with other cables and conductors. It is not recommended to remove the sensor installed on the holding magnet by pulling the cable.

When solving the overwhelming majority of vibration metering problems, the measurement errors of vibration parameters can be 10 ... 15%. The main share of the total measurement error in the measurement information conversion path falls on the sensor. The sensor, in comparison with other equipment, is in the most difficult operating conditions. In addition to vibration, as a rule, a number of other so-called influencing quantities act on it, which include: temperature, pressure, humidity, electromagnetic and electrostatic fields, radiation. Although measures are being taken to reduce the sensitivity of the sensors to influencing quantities, it is not possible to make it equal to zero.

3. Results and discussion

With the help of the Matlab programming environment, monitoring and analysis of vibration parameters, storage of the results of vibration studies are carried out [5]. The Matlab environment allows the use of spectral analysis of the measured vibration signal. This analysis is used to identify components in the frequency domain. An adequate area of its use is the study of various kinds of stationary signals consisting of the sum of components with a constant period. This method is based on the direct Fourier transform [6].

The specificity of signal analysis based on the Fourier transform is traced in the analysis of signals with a constant period (for example, harmonic). For this, a real signal:

$$x(t) = A_1 \cos(\omega t + \varphi_1) = a_1 \cos(\omega_1 t) + b_1 \sin(\omega_1 t), \tag{7}$$

with a period T are laid out in a Fourier series in multiple frequencies. This series can be presented in trigonometric form:

$$\begin{aligned} x(t) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} A_n \cos(n\omega_1 t + \omega_n) = \\ &= \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega_1 t) + b_n \sin(\omega_1 t) \end{aligned}, \tag{8}$$

$$\frac{a_0}{2} = \frac{1}{T} \int_{-T/2}^{T/2} x(t)\,dt, \tag{9}$$

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos(n\omega_1 t)dt\ ,\ b_n = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos(n\omega_1 t)dt\ ,$$

where a_n, b_n are the coefficients.

In complex form:

$$x(t) = \frac{1}{2} + \sum_{n=-\infty}^{\infty} A_n^* e^{jn\omega_1 t}\ , \quad (10)$$

$$A_n^* = \frac{2}{T}\int_{-T/2}^{T/2} x(t)e^{jn\omega_1 t}dt\ , \quad (11)$$

where A_n^* is a complex coefficient.

For a non-periodic signal, the Fourier series is replaced by the Fourier integral:

$$x(t) = \frac{1}{2\pi} + \int_{-\infty}^{\infty} S^*(\omega)e^{j\omega t}d\omega\ , \quad (12)$$

$$S^*(\omega) = \left|\int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt\right|, \quad (13)$$

Expression (12) is called the direct integral Fourier transform, expression (13) is called the inverse transform. In these transformations, instead of the amplitude, the concept of the spectral component $S^*(\omega)$ of the signal is used.

Currently, various discrete algorithms that implement the Fourier transform are widely used. After discretizing $x(t)$ and performing the appropriate substitutions ($\omega = 2\pi n/N$, where N is the number of samples in the investigated time interval (0; T)), expressions (12) and (13) can be written in the following form:

$$S_n^*(\omega) = \frac{1}{N}\sum_{i=0}^{N-1} x_i e^{-j\frac{2\pi}{N}n_i}\ , \quad (14)$$

$$x_i = \frac{1}{N}\sum_{n=1}^{N} S_n^* e^{j\frac{2\pi}{N}n_i}\ , \quad (15)$$

The above system of equations requires N^2 computational operations. Due to the fact that the coefficients in expression (15) are conjugate, it becomes possible to partition the original matrix into matrices with the number of elements not exceeding $\log_2 N$, each of which contains 2 nonzero numbers. This transformation is called the Fast Fourier Transform. Such a mathematical apparatus makes it possible to conduct deeper studies of equipment vibration, to localize the signal in the frequency domain, which makes it possible to represent the signal in the form of separate additive components. These components have characteristic features for each individual defect, which allows not only to integrally assess the technical condition of the equipment, but also to identify the type of defect with a certain degree of reliability.

4. Conclusion

Vibration diagnostics is now widely used, since vibration processes sufficiently fully reflect the technical condition of many parts and assemblies. An important task is to choose the diagnostic parameter that would be most sensitive to changes in the state of the object and determine the high reliability of control. Therefore modern methods of vibroacoustic diagnostics were investigated, diagnostic signs of characteristics of signals are presented, their basic properties, ways of their use at research of a vibro signal are opened. The principles on which they work are described and the characteristics of the signal they use are indicated.

The paper proposes a diagram of a vibration analysis device, which consists of a sensor for measuring vibration, the principle of operation of which is based on the use of a direct piezoelectric effect. The output signal of the sensor goes to a personal computer or any other device with the Windows operating system, where a spectral analysis of the measured vibration signal is carried out using a Matlab-based diagnostic program.

References

[1] Babak V., Beregun V., Krasilnikov A.: Methods and means of vibrodiagnostics of units of cogeneration installations. Scientific Proceedings of STUME 1(216), 2017, 141–145.

[2] Barkov A. V., Barkova N. A.: Vibration diagnostics of machines and equipment. Vibration analysis. SMTU, Sankt-Petersburg 2004.

[3] Barkova N., Barkov A., Grishchenko D.: Vibration diagnostics of equipment units with gas turbine engines. Vibroengineering Procedia 25, 2019, 89–94.

[4] Beregun V. S., Krasilnikov A. I.: Research of excess kurtosis sensitiveness of diagnostic signals for control of the condition of the electrotechnical equipment. Tekhnichna elektrodynamika 4, 2017, 79–85 [http://doi.org/10.15407/techned2017.04.079].

[5] Chernykh I. V.: Modeling of Electrical Devices in MATLAB, SimPowerSystems and Simulink. DMK-Press, Moscow 2008.

[6] Goldin A. S.: Vibration of rotating machines. Mashinostroyeniye, Moscow 1999.

[7] Gorban I. I.: Probability theory and mathematical statistics for scientists and engineers. National Academy of Sciences of Ukraine, Kyiv 2003.

[8] Kozochkin M. P., Kochinev N. A., Sabirov F. S.: Diagnostics and monitoring of complex production processes using measurement of vibration-acoustic signals. Measurement techniques 49.7, 2006, 672–678.

[9] Petrukhin V. V., Petrukhin S. V.: Fundamentals of vibrodiagnostics and vibration measuring instruments. Infra-Inzheneriya, Moscow 2010.

[10] Senyo P. S.: Probability theory and mathematical statistics. Znannya, Kyiv 2007.

[11] Sharapov V. M., Musienko M. P., Sharapova E. V.: Piezoelectric Sensors. Tekhospfera, Moscow 2006.

[12] Stakhova A., Kvasnikov V.: Structure Construction of Acoustic Emission System for Diagnostic of Friction Clusters. IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT) 1, 2018, 64–67.

Ph.D. Eng. Anzhelika Stakhova
e-mail: sap@nau.edu.ua

Doctoral student at the National Aviation University (NAU). Associate Professor of Computerized electrical systems and technologies department NAU.
Main scientific direction – systems for measuring mechanical quantities, the control and forecasting of the technical condition.

http://orcid.org/0000-0001-5171-6330

Prof. Volodymyr Kvasnikov
e-mail: kvp@nau.udu.ua

President of Engineering Academy of Ukraine, Head of Computerized electrical systems and technologies department (NAU).
Main scientific direction – development of methods and instruments for measuring mechanical quantities and metrological support for measurement processes.

http://orcid.org/0000-0002-6525-9721

otrzymano/received: 8.05.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2646

THE METHOD OF OBTAINING THE SPECTRAL CHARACTERISTICS OF THE SCANNING PROBE MICROSCOPE

Mariia Kataieva, Vladimir Kvasnikov
National Aviation University, Ukraine, Kyiv

Abstract. *The article discusses methods and algorithms for digital processing and filtering when carrying out nano-measurements using a scanning probe microscope. The paper discusses frequency methods for improving images, in particular, the use of the Fourier transforms with various filtering methods to improve the quality of the resulting image. Stable computational algorithms have been developed for transforming discrete signals based on the Fourier transform. Methods for the interpretation of the numerical results of the discrete Fourier transform in such packages as Matlab, MathCad, Matematica are presented. It is proposed to use a window transform, developed based on the Fourier transform, which makes it possible to single out the informative features of the signal and to reduce the influence of the destabilizing factors that arise when processing signals from a scanning gold microscope in real conditions.*

Keywords: nano-measurement, digital signal processing, scanning probe microscope, Fourier transform

SPOSÓB UZYSKANIA CHARAKTERYSTYKI WIDMOWEJ SONDY SKANUJĄCEJ MIKROSKOPU

Streszczenie: *W artykule omówiono metody i algorytmy cyfrowego przetwarzania i filtracji podczas nano-pomiarów z wykorzystaniem mikroskopu z sondą skanującą. Badane są metody korekcji częstotliwości obrazu, w szczególności wykorzystanie transformaty Fouriera z różnymi metodami filtracji w celu poprawy jakości otrzymanego obrazu. Opracowano stabilne algorytmy obliczeniowe do konwersji sygnałów dyskretnych na podstawie transformaty Fouriera. Przedstawiono metody interpretacji numerycznych wyników dyskretnej transformaty Fouriera w takich pakietach jak Matlab, MathCad, Matematica. Proponuje się zastosowanie transformacji okienkowej opracowanej na podstawie transformaty Fouriera, która pozwala wyodrębnić charakterystykę informacyjną sygnału i zmniejszyć wpływ czynników destabilizujących występujących podczas przetwarzania sygnału z mikroskopu z sondą skanującą w warunkach rzeczywistych.*

Słowa kluczowe: nano-pomiar, cyfrowe przetwarzanie sygnału, mikroskop z sondą skanującą, transformata Fouriera

Introduction

In recent years, the study of submicron, nano-, and cluster materials has developed rapidly due to existing and potential applications in many technological fields, such as electronics, catalysis, magnetic data storage, structural components, etc.

The study of nanoscale structures (nanostructures) belongs to the direction of nanotechnology. The important components of this scientific and technical direction are the development and study of nanostructured materials, the study of the properties of the obtained nanostructures under various conditions.

Many countries are actively searching for optimal methods for measuring nanomaterials. The results of these methods are increasingly demanding. The main ones are accuracy, versatility, and high-speed calculations. Unfortunately, optimal methods have not yet been developed by which it would be possible to analyze different types of nano-materials with equal success, taking into account the above requirements [13].

One of the most modern methods for measuring the characteristics of materials and diagnosing the features of small-sized systems is electron and scanning probe microscopy (SPM). With the use of this device and the developed techniques, many interesting scientific results have been obtained concerning the mechanical, electrical, and tribological properties of new structural materials and unique products. The purpose of microscopic examination of solids using scanning probe microscopy is to obtain an enlarged image of the surface.

SPM images, due to the specificity of the method, usually contain noise. This is due to the vibration of the probe relative to the sample, acoustic interference, noise from electrical equipment, which are always present when measuring weak signals. Image distortions are also introduced due to thermal drift of the probe relative to the sample, nonlinearity, and creep in the piezoelectric elements from which the scanner is made. In this regard, to obtain high-quality SPM images and conduct their quantitative analysis, it is necessary to improve existing ones and develop new methods for the digital processing of these images.

1. Analysis of existing methods for studying nanostructures

Thus, several methods of analysis are often used to analyze the structure of nanomaterials – fractal, textural, and classical amplitude method. Currently, the most popular fractal methods of analysis, but these methods have low efficiency in the study of materials with complex relief. This is because the main characteristic of fractals - fractal dimension [13, 18], has a direct dependence on the complexity of the relief, which is obtained by the interaction of filler particles and the source material. Other commonly used methods of image analysis are methods based on working with the matrix of the image itself, the so-called. amplitude methods [12].

Such as calculating the areas of figures are in the image, brightness differences, highs, and lows, averages, etc. However, there is an even bigger drawback. In matters relating to the analysis of images obtained by optical methods, there is an error related to the quality of the image and the conditions under which it was obtained. So changing the brightness or sharpness of the image leads to a significant change in the result.

These methods of analysis can be applied to images that are the result of the use of electronic, X-ray, and other non-optical methods of studying the structure of the material. Digital filtering of noisy signals and images is important in solving a wide range of scientific and technical problems. Such tasks arise, in particular, in communication technology to improve the quality of reception of transmitted messages [13, 18].

Thus, several methods of analysis are often used to analyze the structure of nanomaterials – fractal, textural, and classical amplitude method. Currently, the most popular fractal methods of analysis, but these methods have low efficiency in the study of materials with complex relief. This is because the main characteristic of fractals – fractal dimension [13, 18], has a direct dependence on the complexity of the relief, which is obtained by the interaction of filler particles and the source material.

Other commonly used methods of image analysis are methods based on working with the matrix of the image itself, the so-called. amplitude methods [12]. Such as calculating the areas of figures are in the image, brightness differences, highs, and lows, averages, etc.

However, there is an even bigger drawback. In matters relating to the analysis of images obtained by optical methods, there is an error related to the quality of the image and the conditions under which it was obtained. So changing the brightness or sharpness of the image leads to a significant change in the result.

These methods of analysis can be applied to images that are the result of the use of electronic, X-ray, and other non-optical methods of studying the structure of the material. Digital filtering of noisy signals and images is important in solving a wide range of scientific and technical problems.

Such tasks arise, in particular, in communication technology to improve the quality of reception of transmitted messages [13, 18]. Traditionally, approaches that use Fourier transforms have been used to clear information messages from interference.

2. Signal filtering using Fourier transforms

The most convenient from the point of view of the organization of processing and a natural way of sampling is the representation of signals in the form of samples of their values (samples) in separate, regularly located points $T_\partial = \Delta t$. In practice, the sampling operation is performed by measuring the values of the signal using a sensor, the action of which can be described as a convolution with a core $\gamma_\partial = (t)$:

$$u(k\Delta t) = \int_{-\infty}^{\infty} u(t)\gamma_\partial(t - k\Delta t)dt \tag{1}$$

set of values $\{u(k\Delta t)\}$ is a discrete representation of the signal. The core $\gamma_\partial(t)$ is called the sampling aperture. Recovery of a continuous signal from approximate values $\{u(k\Delta t)\}$ performed by interpolation

$$\sum_k u(k\Delta t)\gamma_\partial(t - k\Delta t) \approx u(t) \tag{2}$$

using the interpolating function $\gamma_b(t)$, which is called the recovery aperture.

If we proceed only from the accuracy of the approximation, then there is an important class of signals and the corresponding basic functions for which the distributions (1) and (2) are accurate. These are signals whose Fourier spectrum $U(f) = F\{u(t)\}$ non-zero only within a limited area of the detection area (signals with a limited spectrum). Let the signal sector differ from zero in the interval $\left(\frac{1}{2\Delta t}, \frac{1}{2\Delta t}\right)$, that is

$$U(f) = U(f)rect\left(f\Delta t + \frac{1}{2}\right). \tag{3}$$

$$rect(a) = \begin{cases} 1, & 0 \le \alpha \le 1 \\ 0, & in\ other\ cases \end{cases}$$

For such signals, the sampling and recovery bases are formed from the reference functions:

$$\gamma_\partial(t) = \left(\frac{1}{\Delta t}\right)\sin c\left[\frac{\pi(t-k\Delta t)}{\Delta t}\right],$$

$$\gamma_b(t) = \sin c\left[\frac{\pi(t-k\Delta t)}{\Delta t}\right], \tag{4}$$

a (1) i (2) go into exact equations:

$$u(k\Delta t) = \frac{1}{\Delta t}\int_{-\infty}^{\infty} u(t)\sin c\left[\frac{\pi(t-k\Delta t)}{\Delta t}\right], \tag{5}$$

$$u(t) = \frac{1}{\Delta t}\sum_{-\infty}^{\infty} u(k\Delta t)\sin c\left[\frac{\pi(t-k\Delta t)}{\Delta t}\right], \tag{5}$$

These relations are called the reference theorem. Equation (5) means that the samples of the signal are its values in points $\{k\Delta t\}$, obtained after passing the signal through an invariant to the offset "ideal" filter with pulse and frequency characteristics:

$$h_\partial(t) = \gamma_\partial(t) = \left(\frac{1}{\Delta t}\right)\sin c\left(\frac{\pi t}{\Delta t}\right),$$

$$H_\partial(f) = rect\left(f\Delta t + \frac{1}{2}\right). \tag{6}$$

Equation (6) means that the procedure for restoring a continuous signal from its samples $u\{k\Delta t\}$, can be represented as the transmission through an ideal low-pass filter (7), (8) of a continuous signal of the form

$$u_1(t) = \sum_{-\infty}^{\infty} u(k\Delta t)\,\delta(t - k\Delta t), \tag{7}$$

whose spectrum $U_1(f)$ is periodically extended with the period $\frac{1}{\Delta t}$ spectrum $U(f)$ signal $u(t)$:

$$U_1(f) = \sum_{m=-\infty}^{\infty} U\left(f - \frac{m}{\Delta t}\right). \tag{8}$$

Indeed, such filtering spectrum $U_1(f)$ multiplied by the frequency response of the filter (8), which allocates only one period of the spectrum corresponding $m = 0$ equal to the signal spectrum $u(t)$. Periodic continuation of the spectrum (10) is possible if the rasterization step Δt is less than or equal to the inverse of the length of the spectrum. Otherwise, there is an overlap of adjacent periods of the signal spectrum, $u_1(t)$ and with an ideal low-pass filter, it is impossible to select the signal spectrum in its pure form (Fig. 1).

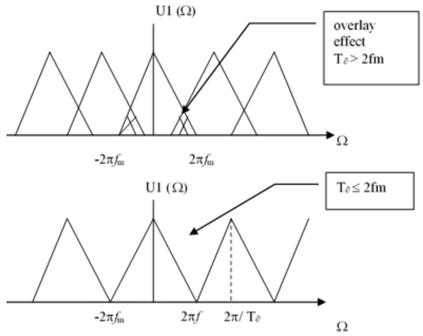


Fig. 1. Overlapping of adjacent periods of the signal spectrum

Although as a result of thinning, each of the time series will be characterized by a frequency range twice less than the signal before filtering, the presence of two sequences (at the output of each filter) allows you to uniquely restore the output signal in reverse conversion. The use of digital low-pass filters in the receiving device reduces the level of additive noise present in the signal transmitted over the communication channel.

Depending on the requirements for receiving information messages and the spectral composition of the signal and noise can also be used band-pass or band-barrier filters [10, 12]. But even though the mathematical apparatus of Fourier transform is an important and useful tool for practical research, it has some limitations.

Thus, filters based on Fourier transforms do not effectively eliminate isolated features of signals. Since this transformation uses infinitely oscillating harmonic functions, information about the isolated features of the signal is contained in all conversion factors.

These problems can be partially solved by applying the window Fourier transform, which allows you to limit the range corresponding to the selected time window.

3. Application of window Fourier transform

The time interval of the signal is divided into subintervals and the conversion is performed sequentially for each subinterval separately. Thus, the transition to the frequency-coordinate representation of the signals, while within each subinterval the signal is "considered" stationary. The result of the window transformation is a family of spectra, which reflects the change in the spectrum of the signal at intervals of the shift of the transformation window.

This allows you to select on the coordinate axis and analyze the features of non-stationary signals. The size of the window function media w(t) is usually set comparable to the stationary interval of the signal. In essence, such a transformation of one nonlocalized basis is divided into some bases localized within the function w(t),

Window conversion is performed according to the expression:

$$S(\omega, b_k) = \int_{-\infty}^{\infty} s(t)\cdot\omega(t - b_k)\cdot e^{-j\omega t}dt, \tag{9}$$

where is the function $w(t - b)$ is a function of the transformation shift window from coordination t, where the parameter b sets fixed offset values.

When shifting windows with a uniform step, the values of b_k are taken equally $b_k = k\Delta b$. Both the simplest rectangular window and special weighted windows (Bartlett, Gauss, etc.) which provide small distortions of a spectrum at the cutting of window segments of signals (neutralization of the Gibbs phenomenon) can be used as transformation windows.

An example of a window conversion for a non-stationary signal at a high noise level is shown in Fig. 2. According to the signal spectrum, it is possible to judge the presence of harmonic oscillations in its composition at three frequencies, to determine the relationship between the amplitudes of these oscillations, and to specify the locality of oscillations in the signal interval.

The signal is the sum of three consecutive radio pulses with different frequencies without pauses, with a signal-to-noise ratio close to 1. The window function wi is given with an effective window width $b \simeq 34$ and a full-size $M = 50$.

The frequency step $\Delta\omega = 0.1$ is set for the results. slightly higher than the actual resolution $\frac{2\pi}{M} = 0.126$.

The coordinate resolution of the window transformation is determined by the width of the window function and is inversely proportional to the frequency resolution. When the width of the window function is equal to b, the frequency resolution is determined by the value $\Delta\omega = \frac{2\pi}{b\Delta\omega}$.

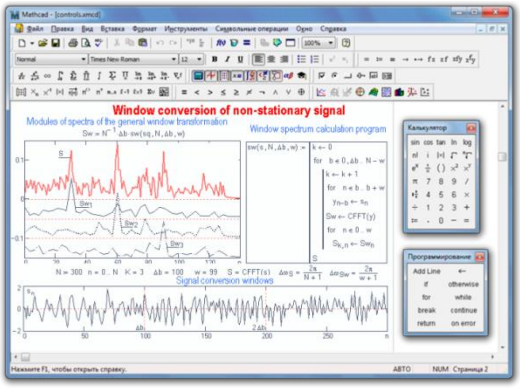


Fig. 2. Example of a conversion window for a transient signal at the high

At the required value of the frequency resolution $\Delta\omega$ accordingly, the width of the window function must be equal to $b = \frac{2\pi}{\Delta\omega}$. For the window Fourier transform, these restrictions are fundamental. Yes, for Fig. 2 when the size of the data array $N = 300$ and the width of the window function $\Delta b = 100$ the frequency resolution of the conversion results decreases $\frac{N}{\Delta b} = 3$ times compared to the original data and graphs $S(\omega, b_k)$ $S(n\Delta\omega\_s)$ on the coordinate n for visual comparison with the graph Σ (νΔωΣ) are constructed with a step on the frequency $\Delta\omega_{S_\omega} = 3\Delta\omega_S$, that is, at points $n = 0, 3, 6, \ldots, N$.

However, this option has significant drawbacks – it is excessive, and many coefficients of decomposition of wavelet functions contain information that is duplicated in other coefficients. Such redundancy is not always a disadvantage, but it leads to a significant increase in computational time due to the lack of effective fast calculation algorithms of the non-stationary random process.

It is also worth remembering that when calculating the value of any fractal dimension, the selected step (window size) has a great influence, by which the image is calculated. For example, the fractal dimension of the same image calculated by a window of 5x5 dots will be less than when using a window of size 3x3.

Frequency-time window conversion is used to analyze non-stationary signals if their frequency composition changes over time. The window transformation function (11) can be translated into a two-dimensional version with independent variables in time and frequency:

$$S(t,\omega) = \int_\tau s(t-\tau) \cdot \omega(\tau) \cdot e^{-j\omega t} d\tau, \qquad (10)$$

In Fig. 3 shows an example of calculation and presentation (module of the right part of the main range of the spectrum) of the frequency-time spectrogram for a discrete task of the input signal $Sq\ (n)$.

To ensure the operation of the window function throughout the signal interval, the initial and final conditions of the calculations were set (continuation of both ends of the signal by zero values at M points).

As can be seen from the results of calculations, the window conversion allows you to highlight the informative features of the signal and time and frequency. The localization resolution is determined by Heisenberg's uncertainty principle, which states that it is impossible to obtain an arbitrarily accurate discrete representation of a signal.

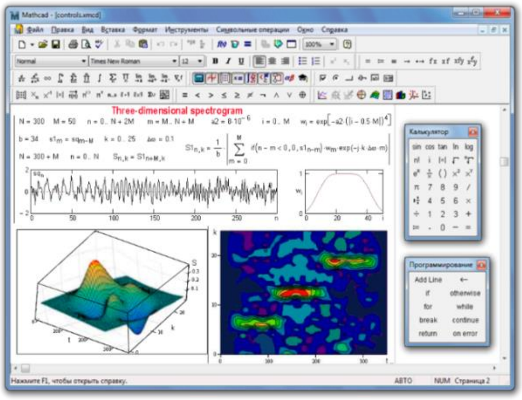
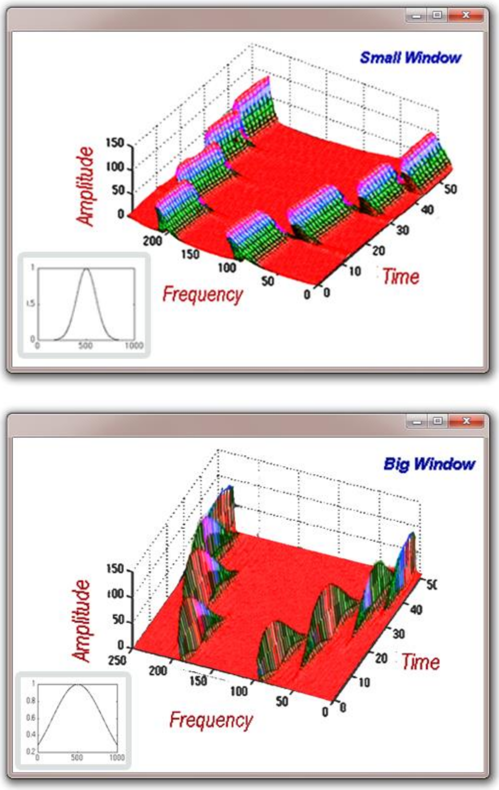


Fig. 3. An example of calculating and presenting a time-frequency spectrogram for a discrete problem of an input signal

The larger the window, the better the temporary resolution, but worse the frequency, and vice versa (Fig. 4).



a)

b)

Fig. 4. An example of frequency-time window conversion of a signal consisting of 4 disjoint intervals: a) Small Window, b) Big Window

Figure 4 shows an example of frequency-time window conversion of a signal consisting of 4 disjoint intervals, each of which is the sum of two harmonics of different frequencies. The Gaussian function of different widths is applied as windows.

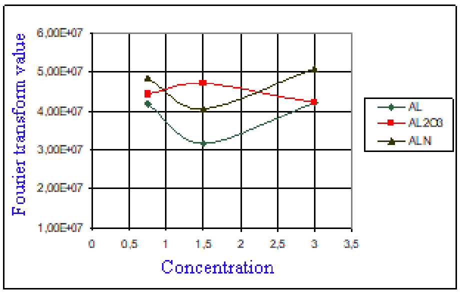
The narrow window provides better temporal resolution and clear fixation of the boundaries of the intervals, but wide frequency peaks within the intervals. A wide window opposite – clearly indicates the frequency of intervals, but with overlapping boundaries of time intervals. When solving practical problems, you have to choose a window to analyze the entire signal, while different parts of it may require the use of different windows.

If the signal consists of distant frequency components, then you can sacrifice spectral resolution in favor of time, and vice versa.

4. Illustrative examples

Using the analytical mechanism, areas with different textures were selected, each type of texture was assigned its unique index. This index, when creating the texture of the image, acts as a label for a particular color, we have the opportunity for both the visual analysis of the texture picture and mathematical analysis of the matrix of indexes.

Going from the original image to its texture map, it is possible to level the contribution of external factors affecting the resulting image. Fig. 5 shows an image obtained using an electron microscope, as well as its texture map. In the future, based on the resulting texture, the Fourier transform is calculated, which allows you to select the main peak values, that distinguish images with different fillers and concentrations.

a) 

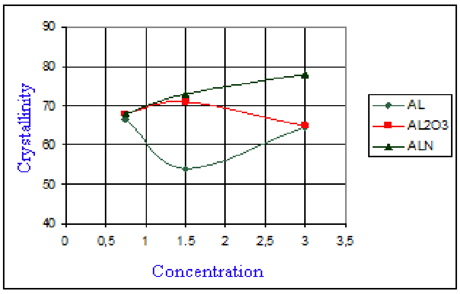
b) 

Fig. 5 a) Dependences of the value of the Fourier transform on the concentration and b) the dependence of the degree of crystallinity on the concentration

The meaning of the Fourier transform is that it allows us to distinguish from the textural picture of the amplitudes and their periodicity, which characterize the distribution and magnitude of texture variability. From a physical point of view, this transformation allows us to assess how much the structure of the object changes, and to link this change with the physicochemical parameters that cause this change.

The results of evaluations with textured cards (quantified by the value of the Fourier transform) were compared with the values of the degree of crystallinity obtained for the selected objects of study. It turned out that the nature of the dependences on the concentration of the value of the Fourier transform is similar to the pattern of changes in the concentration of the degree of crystallinity of the studied objects (Fig. 5).

Since the degree of crystallinity directly affects the structure of the material, we can assume that the value of the Fourier transform responds to structural changes.

5. Conclusions

In the course of analyzing the methods and algorithms for digital processing and filtering of signals during nano-measurements using a scanning probe microscope, frequency methods of image enhancement were studied. It is proposed to use a window transform, developed based on the Fourier transform, which makes it possible to single out the informative features of the signal and reduce the influence of destabilizing factors arising from the processing of signals from a scanning gold microscope in real conditions. The use of the Fourier transform allows both a qualitative analysis of the spectrum and its quantitative interpretation.

Analysis data show that the method not only depends on the degree of crystallinity of the polymer but also distinguishes materials by their properties.

References

[1] Addison P. S.: Secondary transform decoupling of shifted nonstationary signal modulation components: application to photoplethysmography. Int. J. Wavelets Multires. Inf. Proc. 2, 2004, 43–57.
[2] Falvo M. et al.: The nanomanipulator: A teleoperator for manipulating materials at the nanomerter scale. Proc. of Int. Symp. On Science and Technology of Atomically Engineered Materials, 1996, 579–586.
[3] Hyon C. K. et al.: Application of atomic-force-microscope direct patterning to selective positioning of InAs quantum dots on GaAs. Applied Physics Letters 77, 2000, 2607–2609.
[4] Ito K. J. et al.: Servomechanism for locking scanning tunneling microscope tip over surface nanostructures. Rev. of Sci. Inst. 71(2), 2000, 420–423.
[5] Iwasaki H., Yoshinobu T., Sudoh K.: Nanolithography on SiO2/Si with a scanning tunneling microscope. Nanotechnology 14, 2003, 55–62.
[6] Majumdar A. et al.: Nanometer-scale lithography using the atomic force microscope. Applied Physics Letters 61, 2002, 2293–2295.
[7] Mokaberi B., Requicha A. A. G.: Drift compensation for automatic nanomanipulation with scanning probe microscopes. IEEE Trans. on Automation Science and Engineering 3(3), 2006, 199–207.
[8] Mokaberi B., Requicha A. A. G.: Towards automatic nanomanipulation drift compensation in scanning probe microscopes. IEEE Int. Conf. on Robotics and Automation, New Orleans, LA, 2004.
[9] Ohji H. et al.: Fabrication of a beam-mass structure using single-step electrochemical etching for micro structures (SEEMS). J. Micromech. Microeng. 10, 2000, 440–444.
[10] Roth S., Dellmann L., Racine G. A., de Rooij N. F.: High aspect ratio UV photolithography for electroplated structures. J. Micromech. Mecroeng. 9, 2009, 105–108.
[11] Sahoo D. R. et al.: Transient signal based sample detection in atomic force microscopy. Applied Physics Letters 83(26), 2003, 5521–5523.
[12] Said R. A.: Microfabrication by localized electrochemical deposition: experimental investigation and theoretical modeling. Nanotechnology 15, 2004, 867.
[13] Salapaka S., De T.: A new sample-profile estimate for faster imaging in atomic force microscopy. Proceedings of the American Control Conference, Boston, MA, 2004.
[14] Salapaka M. V. et al.: Multimode noise analysis of cantilevers for scanning probe microscopy. Journal of Applied Physics 81(6), 1997, 2480–2487.
[15] San Paulo A., Garcia R.: Tip-surface forces, amplitude and energy dissipation in amplitude–modulation (tapping mode) force microscopy. Physical Review B. 64, 2002, 041406 (1–4).
[16] Sebastian A. et al.: Robust control approach to atomic force microscopy. Proceedings of the IEEE Conference on Decision and Control, Hawai, 2003.
[17] Staub R. at al.: Drift elimination in the calibration of scanning probe microscopes. Rev. Sci. Inst. 66(3), 1995, 2513–2516.
[18] Yang Q., Jagannathan S.: Nanomanipulation using atomic force microscope with drift compensation. Proceedings of the 2006 American Control Conference, Minneapolis, Minnesota, USA, 2006.
[19] Yang S. et al.: Block phase correlation-based automatic drift compensation for atomic force microscopes. IEEE Int. Conf. on Nanotechnology, Nagoya, Japan, 2005.
[20] Yaseen A. S. at al.: Speech signal denoising with wavelet-transforms and the mean opinion score characterizing the filtering quality. Proc. SPIE. 9707, 2016, 970719.

Prof. Vladimir Kvanikov
e-mail: kvp@nau.edu.ua

Professor of the Department of Computerized Electrical Systems and Technologies, National Aviation University, Kyiv, Ukraine.

http://orcid.org/0000-0002-6525-9721

Ph.D. Mariia Kataieva
e-mail: mariia.kataeva@gmail.com

Associate professor of the Department of Computeized Electrical Systems and Technologies, National Aviation University, Kyiv, Ukraine.

http://orcid.org/0000-0002-1586-1861

otrzymano/received: 7.05.2021 *przyjęto do druku/accepted: 7.06.2021*

http://doi.org/10.35784/iapgos.2630

BROADBAND SATELLITE DATA NETWORKS IN THE CONTEXT OF THE AVAILABLE PROTOCOLS AND DIGITAL PLATFORMS

Jacek Wilk-Jakubowski
Kielce University of Technology, Department of Information Systems, Kielce, Poland

Abstract. *Satellites are a transmission medium for providing connectivity and building global, continental, or regional networks around the world (satellite operators effectively use satellites to support Internet traffic), and point-to-point connections are also possible. In practical use, there are combinations of VSAT networks with terrestrial wireless extensions, allowing end users to increase the capabilities offered with the use of satellite. This paper provides selected information on broadband satellite networks using VSAT technology, including available protocols and transmission platforms. The objective of the article is also to present the chosen technical aspects of satellite networks operating with the use of VSAT technology.*

Keywords: information systems engineering, satellite networks, satellite systems, VSAT

SZEROKOPASMOWE SATELITARNE SIECI DANYCH W KONTEKŚCIE DOSTĘPNYCH PROTOKOŁÓW I PLATFORM CYFROWYCH

Streszczenie. *Satelity stanowią medium transmisyjne dla zapewnienia łączności i budowy sieci globalnych, kontynentalnych czy regionalnych na całym świecie (operatorzy satelitarni efektywnie wykorzystują satelity do obsługi ruchu internetowego), przy czym możliwe są także połączenia typu punkt-punkt. W praktycznym użyciu są kombinacje sieci VSAT z bezprzewodowymi rozszerzeniami naziemnymi, co pozwala zwiększyć użytkownikom końcowym możliwości oferowane za pośrednictwem satelity. W artykule zamieszczono wybrane informacje na temat szerokopasmowych sieci satelitarnych z wykorzystaniem technologii VSAT, z uwzględnieniem dostępnych protokołów i platform transmisyjnych. Celem artykułu jest także przedstawienie wybranych aspektów technicznych sieci satelitarnych pracujących z wykorzystaniem technologii VSAT.*

Słowa kluczowe: inżynieria systemów informacyjnych, sieci satelitarne, systemy satelitarne, VSAT

Introduction

In practice, satellite networks are used to provide access to multimedia services as well as to build disaster resilient computer networks. According to a recent research, the expenditure on disaster management is increasing every year [32], and robots are used for disaster management [31]. Besides, satellite networks are the cheapest method for providing global connectivity and communication wherever, for technical (lack of terrestrial infrastructure) and economic reasons, connectivity would be impossible. An example can be seas and oceans, inaccessible areas of taiga or desert, mountainous areas, and many others [33]. In the European Union countries, due to numerous programs aimed at eliminating digital exclusion, satellites are becoming an accelerator of changes aimed at improving the reliability of modern telecommunications systems. In addition to access to numerous multimedia services, it becomes possible to provide access to the Internet with the use of satellite links. For reception, or transmission and reception on the client side, cheap VSAT (Very Small Aperture Terminal) terminals are usually used with antenna diameters usually below 1 m. Besides the antenna, such systems include a receiver with a converter, a modem, cabling, and optionally a router. On the other side, there is an NOC (Network Operations Centre) station. In order to ensure transmission, transponders placed on geostationary satellites are applied. With their use, it becomes possible to transmit data in two ways and to build interactive networks. It is also possible to work with multiple satellites simultaneously and with different frequency bands [34]. During the transmission to VSAT terminals, one of the multiplexing techniques is used. Therefore, the implementation of appropriate modulation, compression, and coding techniques is crucial as described in [35].

Currently, satellite service providers operate their networks at the highest level of redundancy and reliability, which guarantees the Quality of Service (QoS). Figure 1 shows the share of satellite communications operators in the global market. According to Euroconsult data, over ten percent in market share has been recorded for the three largest satellite service operators (SES, Intelsat, and Eutelsat) [14].

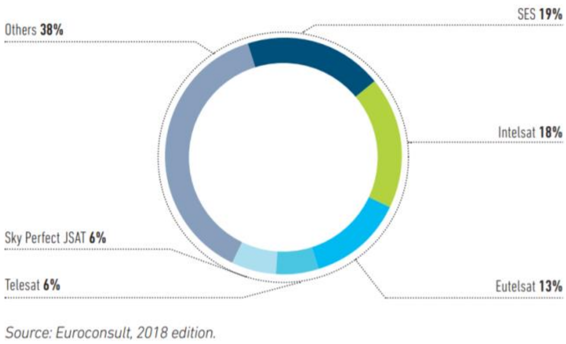


Fig. 1. Operator share in the global market based on revenues [14]

Due to multiple layers of redundancy and backup systems, the actual satellite lifetime is close to 100%, and even often exceeds this value. In many cases, these values match or exceed military specifications, making satellite communications the only technology capable of providing communications with a reliability in excess of 99% (some networks offer 99.9921% availability).

Since 2013, Federal Communications Commission (FCC) reports presented the transmission speeds offered by satellite communication systems besides DSL technology, fiber optic technology, and cable TV operators' networks [18, 19]. Based on the above reports, it was identified that the actual data rates are typically higher than the declared ones. The differences observed are at the level of 40% for 90% of users during PTH (Peak Traffic Hour), which allows to rank the service using satellites at the first place for broadband services offered in the USA.

1. Literature review including IEEE databases

Analyzing the literature in the scope of the article, Benjamin A. Pontano presents the possibilities of the *Linkway™ system for 21st century military communications* [27]. Gerson Souto and John Stevenson discuss *Technical features of the @Intelsat Internet product suite* (Intelsat's technical solutions can be found in [28]). Hiroyasu Obata, Kazuya Tamehiro, and Kenji Ishida primarily describe *Experimental Evaluation of TCP-STAR for Satellite*

Internet over WINDS [22], Horst D. Clausen and Bernhard Nocker present *Internet Services via Direct Broadcast Satellites* [11], Hyoung-Kee Choi, Osama Qadan, Dolors Sala, John O. Limb, and Jeff Meyers introduce *Interactive Web Service via Satellite to the Home* [10]. In addition, the following publications should be indicated as crucial, taking into account the IEEE database: Norman Abramson: *Internet Access Using VSATs* [1], Leonid Volkov: *VSAT Networks of Russian Satellite Communications Company* [29], Mamadou A. Barry, James K. Tamgno, Claude Lishou, and Renaud KK Maleka: *Challenges of Integrating a VoIP Communication System on a VSAT Network* [6], Paul D. Bacsich: *JANUS: one year's experience with a TCP/IP VSAT Network* [4], Noriharu Suematsu, Suguru Kameda, and Shigeru Eguchi et al.: *Multi-Mode Portable VSAT for Disaster-Resilient Wireless Networks* [13], Suguru Kameda, Tetsuya Okuguchi, Shigeru Eguchi, and Noriharu Suematsu: *Development of Satellite-Terrestrial Multi-Mode VSAT Using Software Defined Radio Technology* [24], Tao Huang: *A Ka and Ku Band Feed Horn for Satellite Broadband and TV Integrated IP Solution* [21]. Additionally, a literature review has been presented below, taking into account domestic publications in this area.

Analyzing the literature on satellite communications, modulation techniques, IP traffic transmission (IP transit both domestic and international) and satellite data communication networks (including VSAT technology), the source of knowledge can be a book [38] by Ryszard Zieliński, published in 2009 and articles by the same author, e.g. [37]. A separate matter is the analysis of the propagation conditions of microwaves in satellite communications. These issues are partly consistent with the research carried out by the author at the Faculty of Electrical Engineering, Automatics and Computer Science of the Kielce University of Technology (the most recent results are from a research entitled: *The analysis of solar activity and factors affecting the propagation of microwave signals in the troposphere and satellite signal reception in rainy weather in the area of Kielce*, which has been initiated in 2010). Many studies related to the engineering of information systems using VSAT networks, as well as their technical parameters, link budgets, and radio wave propagation were included in the author's monograph [35]. However, due to the subject of this article, these issues are omitted.

2. Technical aspects of using networks based on VSAT technology

In the case of using a geostationary orbit, one of the many benefits is that the satellite placed on it is perceived to be stationary relative to the Earth's surface. For this reason, satellite systems in geostationary orbit have simplified satellite tracking devices compared to systems with satellites located in low orbit. These devices are installed in the systems of satellite platform operators. Exemplification can be the largest satellite platform in Central and Eastern Europe – Cyfrowy Polsat S.A. (see Fig. 2).

Fig. 2. One of the antennae of the platform Cyfrowy Polsat S.A. (Warsaw)

Although the equipment is located close to the antenna, the position of the satellite can be monitored from the main monitoring center (see Fig. 3).

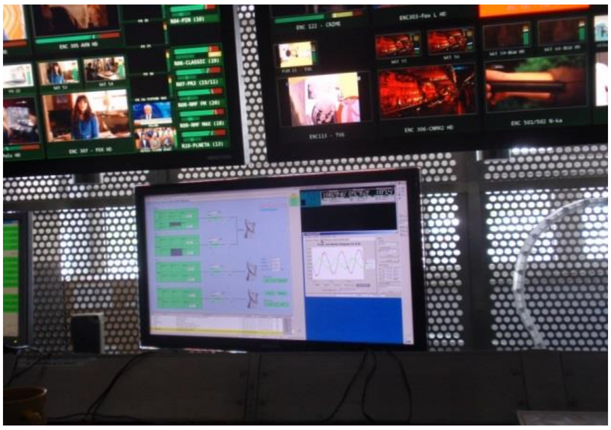


Fig. 3. Monitoring Centre of Cyfrowy Polsat S.A. (Warsaw) [29]

Antennae are oriented in the selected direction, while the position of the satellite in a station keeping window is changing constantly, which translates into the correction of azimuth and elevation angles by satellite traffic engineers (for example, data for two hours for the Ka-Sat satellite are presented in Table 1). It is worth noting that the Ka-Sat is the first HTS (High Throughput Satellite) satellite in Europe and in the Mediterranean Region with considerable capacity and efficiency, operating in the Ka-band [15]. It is anticipated that the advent of satellites in the Ka-band will lead to a reduction in the cost of access to Internet services using satellites in the long term, while increasing bit rates to the levels offered by fibre optic technology.

Table 1. Examples of changes in azimuth and elevation angles, as well as distances from the receiving station for the Ka-Sat satellite (April 2019, adapted from Eutelsat [17])

Time(gmt)	Azimuth(deg)	Elevation(deg)	Distance(km)
16:30	-165.29	30.763	42157.216
17:00	-165.293	30.759	42160.826
17:30	-165.294	30.756	42164.504
18:00	-165.294	30.752	42168.186
18:30	-165.292	30.748	42171.81

When the position of the selected satellite to the Earth is stabilized, the antennae on board of the directional satellites have a very high energy gain. This results from the need to compensate for signal degradation due to the long distance between the satellite and the ground station. The accuracy of the Ka-Sat keeping window equals +/- 0.1 degree in elevation and azimuth. The parameters of the satellite's actual position can be determined by using 3D tools. A sample 3D diagram for the Ka-Sat satellite is shown in Figure 4.

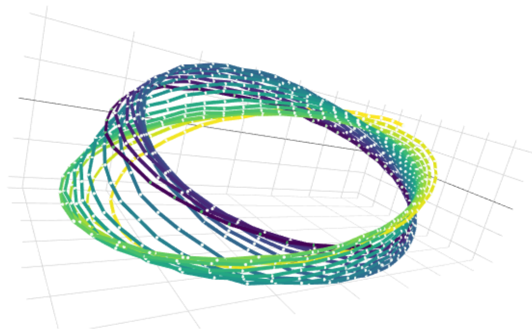


Fig. 4. Examples of three-dimensional visualization of positional changes for the Ka-Sat satellite (adapted from Eutelsat [16])

On the other hand, a certain inconvenience is that the use of a geostationary satellite results in significant delays (for example, in corporate networks it is several hundred milliseconds, as the time of each individual hop between the satellite and the Earth is between 120 milliseconds and 140 milliseconds [35]). Undesirable delay affects the use of the TCP (Transmission Control Protocol) /IP (Internet Protocol) protocol stack, which was developed for low delay and low bit error rate terrestrial links [25, 33, 36, 37]. It is difficult to use in both network and transport layers [2]. Information on this will be presented in the next section.

3. Overview of protocols and digital satellite platforms

Several types of protocols and transmission platforms are currently used for the provision of broadband multimedia and IT (Information Technology) services, particularly Internet access. In this regard, the following should be pointed out: (1) the TCP/IP protocol stack (protocol tunneling); (2) the DVB-S (Digital Video Broadcasting – Satellite) / DVB-S2 (Digital Video Broadcasting – 2nd generation) platforms; (3) as well as the ATM (Asynchronous Transfer Mode) protocol.

DVB-S digital platform is usually used for multimedia services and data transmission via IP tunneling. A newer standard is DVB-S2, established in 2003. The platform, aside from the specification of protocols, includes, among others: coding of signals based on MPEG (Motion Picture Export Group) standard, providing additional information (mainly for configuration and synchronization of the receiver), protection of data against the impact of undesired factors, i.e. interferences that cannot be predicted using DC (Direct Current) and AC (Alternating Current) analysis methods. Through packet transmission, it is possible to access coded programs in MPEG streams. DVB platforms may be an alternative to VOD (Video On Demand) services available via terrestrial links by distributing content to specific groups of recipients. In the case of Internet services, IP tunneling is applied. In practice, IP packets are subjected to MPE (Multiprotocol Encapsulation) after which they get into DVB transport streams. Because MPE packets are segmented, their size is 184 B (the additional 4 B is the MTE packet header). The total container capacity of the DVB-S transport stream is therefore equal to 188 B. The Section Packing (SP) technique aims to improve the encapsulation efficiency of IP packets by placing them optimally in MPEG packets [38]. Sometimes, depending on the interference, it is necessary to change the channel coding, increase the transmitter power, or modify the modulation technique to provide the required QoS level. This often occurs at the expense of reduced data transfer efficiency. DVB-S and DVB-S2 platforms using different modulations allow to transmit data at high speed and to protect the data from errors. The return channel often includes an ATM protocol (the interface may be installed in routers). IP datagrams are encapsulated into packets and then segmented into small ATM cells.

Depending on the type of service, several data transmission profiles can be distinguished: (1) Data Pipe (DP) – used for transmission between end users; (2) Data Streaming (DS) – used for transmission of synchronous, asynchronous or synchronous data streams; (3) Data Carousel (DC) – allows for cyclic transmission of data sets; and the previously mentioned (4) MPE encapsulation, which allows for various types of transport services (IP packet transmissions).

In turn, the use of the ATM protocol, unlike the TCP/IP protocol stack, allows the establishment of a virtual channel between selected points and the transmission of data in cells having a fixed length of 53 B at rates ranging from about 2 Mbps to about 2.4 Gbps [9, 12]. Since the cell-tax occupies 5 B, this protocol is not recommended for links with low capacity and data rates. Systems integrated with ATM networks use similar solutions (S-ATM protocol), additionally focused on providing access to the MAC (Medium Access Control) and the physical layer of the link [38]. In comparison to the ATM protocol, the S-ATM protocol has a modified cell-tax format, which is coded redundantly with BCH (Bose-Chandhuri-Hocquenghen) codes. The advantage of the protocol incluces a guarantee of QoS. Depending on the kind of service in ATM networks, the following transmission types can be distinguished: (1) CBR (Constant Bit Rate) – transmission with uniform bandwidth requirements (e.g., for link emulation); (2) VBR (Variable Bit Rate) – transmission of a data stream with variable speed (we can distinguish in this respect: (A) RT VBR – transmission of a data stream with non-uniform distribution of traffic over time; (B) N-RT VBR – transmission of a data stream that does not require real-time support); (3) GFR (Generic Frame Rate) – transmission similar to VBR, however the main measure is not cells but frames; (4) ABR (Available Bit Rate) – transmission with undefined bit rate (without significant time requirements) with the possibility of fair bandwidth allocation; (5) UBR (Unspecified Bit Rate) – transmission with undefined bit rate (without transmission quality requirements).

Historically, IP packet encapsulation was applied in ATM cells or transport containers. The overarching goal was to try to use standard protocols with minimal hardware investment. A gateway was responsible for encapsulating traffic from terminals. The implementation of packet techniques influenced the availability of multimedia services and the optimization of the system architecture. In practice, there are numerous network topologies, which are selected depending on the nature of the services provided [34]. The network may be adapted to support the TCP/IP protocol stack, and it is important to ensure the required QoS level, which is done by providing appropriate traffic classification and stream structure. When stations are Web clients, TCP is typically the primary transport protocol. An acknowledgement is awaited before the packet transmission to the remote site is complete, and the network functions as if the delay is the result of congestion in the link. Since this results in packets being sent at a low start rate due to the Slow Start Algorithm (SAA) used, this mechanism is very inefficient [7]. Moreover, when the data is transmitted before the segment length optimization is completed, the transmission will not be realized at close to the maximum speed. To minimize this risk, it is assumed to start the transmission with a double-length segment [38]. Since the time required to achieve the desired throughput can be several hours, the Selective Acknowledge (SACK) mechanism is in common use [38]. The TTL (Time-To-Live) of a datagram is valid for the IP protocol. Due to the fragmentation of IP packets, it is usually not recommended to generate the same datagram for several minutes. Moreover, during the maximum TTL lifetime of a datagram, IP packets that are sent to the same recipient should have different identifiers (timestamps are helpful in this regard). To improve the channel efficiency, some implementations allow to apply the MFWS (Multiplication Factor of Window Size) scaling mechanism. This technique is useful to increase the length of the transmission window by multiplying its size by a certain factor [23, 38].

In modern satellite network, the process of optimization and acceleration of TCP sessions is performed using PEP (Performance Enhancing Proxies) mechanism that allows modification of TCP headers. This technique is implemented either in the satellite modem software or is made available in the operator resources and hardware on the part of the client and operator. TCP spoofing makes it possible to speed up TCP/IP protocol stack using modern processors, routers, and switches. Reduction of ACK (Acknowledgement) enables an increase in the offered data rate in the satellite links. Acknowledgements are returned quickly, thus hiding significant delays in the link (Fig. 5).

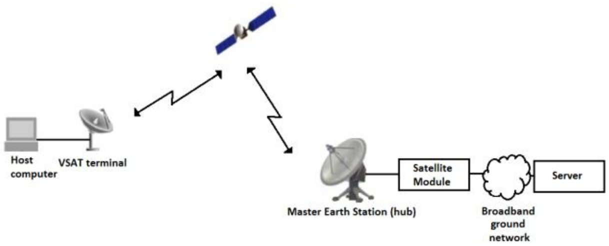


Fig. 5. VSAT network using TCP Spoofing technique

The acknowledgement packets are sent to the local server using the central station module. This reduces the delay, as the data is transmitted to the central station at maximum speed over the terrestrial link. Data transmitted via satellite is buffered. In turn, the VSAT terminal intercepts acknowledgements that are sent at the same time by the host and sends a response to the central station modem. The hardware used is designed to suppress true acknowledgments that arrive from a distant station. When an acknowledgement is not received, the packet is sent from the buffer. The task of the satellite module is removing subsequent packets from the buffer and sending packets (in the case of erroneous packets) to the VSAT terminal. In addition to the TCP protocol, the connectionless UDP protocol is also used in local networks, but is not accelerated due to its connectionless operation. The minimization of the effect of delay on a TCP session is given in Figure 6.

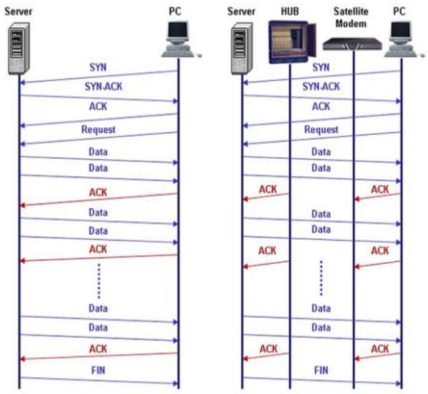


Fig. 6. Illustration of minimizing the impact of delay on a TCP session [38]

The TCP session establishment is based on three packet exchanges that occur between the host and the server (an acceleration example is shown in Figure 7). First, the host sends an SYN request to establish a connection with the server. The server sends an SYN-ACK packet to the host in response to the request, which means that the SYN packet has been received and the server is ready to connect. The host, upon receiving the SYN-ACK packet, sends an ACK-REQUEST packet to the server in order for the server to begin transmission. Due to the significant distance between the satellite and the Earth, the time required to transmit the first packet to the host is long (more than 1 second), which is especially noticeable when transmitting small amounts of data. To counteract this, the acceleration technique is used for WEB applications.

In practice, the VSAT station responds to the SYN packet that was sent by the host by sending an SYN-ACK acknowledgment packet. The request is sent to the server, and then the satellite module in response to the SYN-ACK packet received from the server, using the broadband link, sends an ACK-REQUEST packet to the server. Through a synchronous cheating mechanism between the host and the server, the time required to establish a connection is reduced. This way, data can be sent from the server to the host-coupled VSAT station without unnecessary time delay [38].

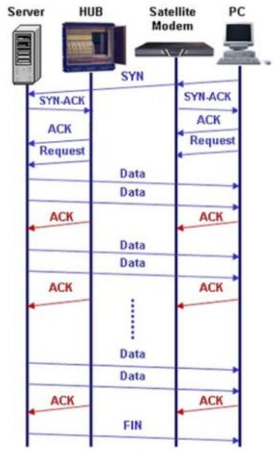


Fig. 7. Illustration of the acceleration technique (own elaboration on the basis of [20])

4. Conclusions

Communication satellites are successfully used to provide Internet services from the beginning. Due to fast transmission speeds and the possibility of providing almost global access using satellites without the need for costly terrestrial technical infrastructure, satellite data transmission systems become more and more popular [5, 8].

Many protocols and digital platforms are used for the implementation of broadband services via satellite links. In this respect, there are mainly solutions based on TCP/IP protocol stack (protocol tunneling), ATM protocol, and DVB-S and DVB-S2 platforms. It is worth noting that ETSI (European Telecommunications Standards Institute) standards apply to all types of VSATs. In general, the use of IP protocol is associated with very good availability for many applications, while ATM protocol – with the need to guarantee the required level of transmission quality. Due to the degradation of satellite signals, it is recommended to use adaptive systems, as they allow to ensure the required QoS and to change the modulation technique and increase the channel coding (if necessary). The architecture of broadband systems requires an appropriate design of satellite modules (there are two basic methods of handling data streams: bent pipe and on-board processing and switching). Service providers often use satellite links as backup links. They can be used as access and backbone networks. In the case of connection of remote networks to a main backbone, satellite systems are often the only way to ensure reliability above 99%. Because satellites are designed with multiple layers of redundancy and backup systems, and their lifetimes often exceed 100%, the levels achieved are equal to or even exceed those offered in military specifications.

References

[1] Abramson N.: Internet access using VSATs. IEEE Communications Magazine 38(7), 2000, 60–68.

[2] Allman M., Glover D., Sanchez L.: Enhancing TCP Over Satellite Channels using Standard Mechanisms. BCP 28, RFC 2488, 1999.

[3] Allman M. (ed.): Ongoing TCP Research Related to Satellites. IETF draft, 1999.

[4] Bacsich P. D.: JANUS: one year's experience with a TCP/IP VSAT network. IEE Colloquium on Networking Aspects of Small Terminal Satellite Systems, London 1994.

[5] Baras J. S., Corson S., Papademetriou S., Secka I., Suphasindhu N.: Fast asymmetric Internet over wireless satellite-terrestrial networks. Monterey 1997.

[6] Barry M. A., Lishou C., Maleka R. K. K., Tamgno J. K.: Challenges of integrating a VoIP communication system on a VSAT network. International Conference on Advanced Communication Technology (ICACT), Bongpyeong 2017.

[7] Becerra A., Berberana I., Gavilán J.: LAN Internetworking Using VSAT Systems. Proceedings of 1994 3rd IEEE International Conference on Universal Personal Communications, San Diego 1994.

[8] Bobrov A., Bobrov N., Bobrov S.: Organization and economical aspects of satellite networks with asymmetric Internet access. 14th International Crimean Conference Microwaveand Telecommunication Technology, Sevastopoll 2004.

[9] Charalambous C. P., Frost V. F., Evans J. B.: Performance evaluation of TCP extensions on ATM over High Bandwidth Delay Product Networks. IEEE Communications Magazine 37 (7), 1999, 57–63.
[10] Choi H-K., Qadan O., Sala D., Limb J. O., Meyers J.: Interactive Web Service via Satellite to the Home. IEEE Communications Magazine 39(3), 2001, 182–190.
[11] Clausen H. D., Nocker B.: Internet services via direct broadcast satellites. IEEE International Performance, Computing and Communications Conference, Phoenix 1997.
[12] Cuevas E. G.: The development of performance and availability standards for satellite ATM networks. IEEE Communications Magazine 37(7), 1999, 74–79.
[13] Eguchi S., Kameda S., Kuroda K., Oguma H., Sasanuma M., Suematsu N.: Multi-mode portable VSAT for disaster-resilient wireless networks. Asia Pacific Microwave Conference (APMC 2014), Sendai 2014.
[14] Euroconsult, 2018 Edition: https://www.eutelsat.com/files/live/sites/eutelsat-internet/files/PDF/investors/2017-18/Eutelsat_Communications_Reference_Document_2017-18.pdf (14.09.2018).
[15] Eutelsat: Ka-Sat, URL: https://www.eutelsat.com/sites/eutelsat-internet/home/satellites/9-east.html#ka-sat (07.06.2016).
[16] Eutelsat: Services, URL: https://services.eutelsat.fr/deploy_Sorbet_SSO/pages/changeGraphParameters.do (30.04.2019).
[17] Eutelsat: Services, URL: https://services.eutelsat.fr/deploy_Sorbet_SSO/pages/displaySelectAELocation.do (30.04.2019).
[18] FCC Office of Engineering and Technology i Consumer and Governmental Affairs Bureau, http://www.fcc.gov (31.04.2013).
[19] Gajewski M.: Internet satelitarny przoduje na liście usług szerokopasmowych w USA, https://www.chip.pl/2013/04/internet-satelitarny-przoduje-na-liscie-uslug-szerokopasmowych-w-usa (25.04.2013).
[20] Gannon P.: Understanding VPNs over broadband satellite, https://www.bcsatellite.net/blog/understanding-vpn-over-broadband-satellite (21.09.2014).
[21] Huang T.: A Ka and Ku Band Feed Horn for Satellite Broadband and TV Integrated IP Solution. 46th European Microwave Conference (EuMC), London 2016.
[22] Ishida K., Obata H., Tamehiro K.: Experimental Evaluation of TCP-STAR for Satellite Internet over WINDS. 10th International Symposium on Autonomous Decentralized Systems, Tokyo 2011.
[23] Jacobson V., Braden R., Borman D.: TCP Extensions for High Performance. RFC 1323, 1992.
[24] Kameda S., Okuguchi T., Eguchi S., Suematsu N.: Development of satellite-terrestrial multi-mode VSAT using software defined radio technology. Asia-Pacific Microwave Conference, Sendai 2014.
[25] Marchese M.: Performance analysis of the TCP behavior in a GEO Satellite Environment. Computer Communications Journal 24(9), 2001, 877–888.
[26] Mathis N., Mahdavi J., Floyd S., Romanow A.: TCP selective Acknowledgment Options, RFC 2018, 1996.
[27] Pontano B. A.: Linkway for 21st century military communications. MILCOM 2000 Proceedings – 21st Century Military Communications. Architectures and Technologies for Information Superiority (Cat. No. 00CH37155), Los Angeles 2000.
[28] Souto G., Stevenson J.: Technical features of the @INTELSAT Internet product suite. IEE Colloquium on Current Developments in Intelsat (Ref. No: 1997/367), London 1997.
[29] Volkov L.: VSAT networks of Russian Satellite Communications Company. 3rd International Conference on Satellite Communications (IEEE Cat. No. 98TH8392), Moscow 1998.
[30] Wilk J., Marciniak M.: Systemy geostacjonarne we współczesnej telekomunikacji. Zastosowania technologii informatycznych. Teoria i praktyka. Wydawnictwo Naukowe Instytutu Technologii Eksploatacji – Państwowego Instytutu Badawczego, Radom 2015.
[31] Wilk-Jakubowski G., Harabin R., Ivanov S.: Robotics in crisis management: a review of the literature, submitted for publication. Technology in Society 2020/2021.
[32] Wilk-Jakubowski G.: Normative Dimension of Crisis Management System in the Third Republic of Poland in an International Context. Organizational and Economic Aspects, Wydawnictwo Społecznej Akademii Nauk, Łódź-Warszawa 2019.
[33] Wilk-Jakubowski J.: Information systems engineering using VSAT networks. Yugoslav Journal of Operations Research, online first, http://yujor.fon.bg.ac.rs/index.php/yujor/article/view/833 (22.06.2020).
[34] Wilk-Jakubowski J.: Overview of broadband information systems architecture for crisis management. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska – IAPGOS 10(2), 2020, 20–23.
[35] Wilk–Jakubowski J. Ł.: Propagacja fal radiowych w łączności satelitarnej. Radiowaves propagation in satellite communications systems. Wydawnictwo Politechniki Świętokrzyskiej, Kielce 2018.
[36] Yurong H., Lee V. O. K.: Satellite-based Internet: a tutorial. IEEE Communications Magazine 39(3), 2001, 154–162.
[37] Zieliński R. J.: Nowe techniki w systemach VSAT, http://absta.pl/ryszard-j-zieliskinowe-techniki-w-systemach-vsat.html (25.07.2019).
[38] Zieliński R. J.: Satelitarne sieci teleinformatyczne. Wydawnictwa Naukowo-Techniczne, Warszawa 2009.

Ph.D. Jacek Łukasz Wilk-Jakubowski
e-mail: j.wilk@tu.kielce.pl

Kielce University of Technology, Ph.D. (technical) Faculty of Electrical Engineering, Automatic Control and Computer Science, Department of Information Systems.
Research interests: computer science, ICT systems, data transmission, signal processing, electrical engineering, wave propagation.

http://orcid.org/0000-0003-1275-948X

otrzymano/received: 7.04.2021 *przyjęto do druku/accepted: 7.06.2021*