http://doi.org/10.35784/iapgos.5866 received: 05.02.2024 | revised: 07.04.2024 | accepted: 19.06.2024 | available online: 30.06.2024

REVIEW OF THE ACHIEVEMENTS OF EMPLOYEES OF THE LUBLIN UNIVERSITY OF TECHNOLOGY IN THE FIELD OF FUZZY SET UTILIZATION

Maciej Celiński, Adam Kiersztyn
Lublin University of Technology, Faculty of Mathematics and Technical Computer Science, Lublin, Poland

Abstract. *In this paper, we present a review of research on the applications of fuzzy set theory conducted by Lublin University of Technology researchers. We focus on analyzing research trends and practical applications of fuzzy sets in time series analysis and missing data imputation. Fuzzy sets constitute a key methodology for addressing data uncertainty and imprecision. We discuss various techniques within the field of fuzzy sets, including fuzzy classification, outlier detection, and missing data imputation, emphasizing their significance across various fields of science and social life. The presented results indicate the potential for innovative research and further development in this field. The academic community at Lublin University of Technology plays a significant role in promoting and advancing fuzzy set theory, which is crucial for future scientific and technological research.*

Keywords: fuzzy sets, fuzzy methods, outlier detection, fuzzy classification, missing data filling

PRZEGLĄD OSIĄGNIĘĆ PRACOWNIKÓW POLITECHNIKI LUBELSKIEJ W DZIEDZINIE WYKORZYSTANIA ZBIORÓW ROZMYTYCH

Streszczenie. *W niniejszym artykule przedstawiamy przegląd badań nad zastosowaniami teorii zbiorów rozmytych prowadzonych przez naukowców z Politechniki Lubelskiej. Skupiamy się na analizie trendów badawczych i praktycznych zastosowaniach zbiorów rozmytych w analizie szeregów czasowych oraz imputacji brakujących danych. Zbiory rozmyte stanowią kluczową metodologię do rozwiązywania problemów związanych z niepewnością i nieprecyzyjnością danych. Omawiamy różne techniki w dziedzinie zbiorów rozmytych, w tym klasyfikację rozmytą, wykrywanie wartości odstających oraz imputację brakujących danych, podkreślając ich znaczenie w różnych dziedzinach nauki i życia społecznego. Przedstawione wyniki wskazują na potencjał innowacyjnych badań i dalszego rozwoju w tej dziedzinie. Społeczność akademicka Politechniki Lubelskiej odgrywa istotną rolę w promowaniu i rozwijaniu teorii zbiorów rozmytych, co jest kluczowe dla przyszłych badań naukowych i technologicznych.*

Słowa kluczowe: zbiory rozmyte, metody rozmyte, wykrywanie wartości odstających, klasyfikacja rozmyta, uzupełnianie brakujących danych

Introduction

This article provides an overview of current scientific works and research in the field of fuzzy sets. In the second half of the 20th century, L. A. Zadeh introduced the theory of fuzzy sets, which extends the classical set theory [23]. The motivation behind the development of this theory was the mathematical description of ambiguous and imprecise phenomena and concepts, commonly used in a flexible manner in human life. The presented article serves as a review of selected works considered most interesting and also presents the results of specific research utilizing fuzzy sets. This indicates that there is still a wide scope for describing and presenting the most innovative applications of fuzzy sets using data from various scientific disciplines.

Fuzzy sets mathematically describe phenomena and concepts whose nature is ambiguous and imprecise. In fuzzy set theory, unlike Boolean logic, we use linguistic variables instead of [0 and 1], representing true or false. Vague descriptions play a significant role in human thinking, pattern recognition, information transmission, and abstraction. Therefore, such variables take on imprecise values related to certain object features.

Techniques and methods employed in fuzzy systems are designed to represent information that is imprecise, undefined, or indeterminate. Consequently, it becomes possible to describe phenomena with an ambiguous nature, which is impossible to represent in classical set theory and two-valued logic. A characteristic feature of fuzzy systems is that the knowledge processed through them takes a symbolic form and is recorded in the form of fuzzy rules. Fuzzy systems find application where there is insufficient knowledge about a specific mathematical model governing a phenomenon or where recreating such a model is impossible or impractical. Their application is feasible for data from control databases, electrical engineering, and other fields focused on natural language processing.

Our main research goal is to conduct a thorough analysis of scientific works authored by employees of the Lublin University of Technology published in recent years. To achieve this goal, have chosen several key publications and intend to analyze them regarding their contribution to the development of research on fuzzy sets. We aim to identify significant scientific issues, assess innovative approaches, and understand how these works have contributed to the advancement of the field of fuzzy sets. Through my research, We expect to gain a deeper understanding of Lublin University of Technology's contribution to the field of fuzzy sets and identify areas that warrant further research and development.

1. Theoretical background

Fuzzy numbers are a mathematical tool that can represent uncertainty or ambiguity in a more flexible way than traditional real numbers. Unlike classical numbers, which are either true or false, fuzzy numbers can take on values on a continuum, from completely false to completely true, through all degrees of truth in between. They are described using membership functions that determine to what extent a given element belongs to a given set. Thanks to these numbers, we can precisely model situations in which we are not sure whether something is true, false, or somewhere in between. Fuzzy numbers are used in various fields, such as control, artificial intelligence, economics and medicine, where it is necessary to take into account uncertainty and instability of data.

In many cases, linguistic descriptors are used to describe numerical phenomena. The values of numerical variables can be expressed using fuzzy sets. For this purpose, a fuzzy modification of the three-sigma rule was used [11]. In a very intuitive way, it is possible to introduce 2k+1 descriptors describing the position of a given element relative to the analyzed set.

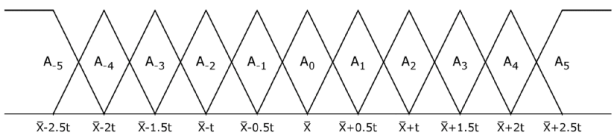

Fig. 1. Example of modification of the three-sigma rule using triangular membership functions for N = 5

After applying the appropriate encoding, each individual value is represented as membership degrees to 2k+1 descriptors.

$$D_{-k}, D_{-k+1}, \dots, D_{-1}, D_0, D_1, \dots., D_{k-1}, D_k \text{ Let} \quad (1)$$

$$\mu(x) = [\mu_{D_{-k}}(x), \mu_{D_{-k+1}}(x), \mu_{D_{-1}}(x) \mu_{D_0}(x), \mu_{D_1}(x), \mu_{D_{k-1}}(x), \mu_{D_k}(x)] \quad (2)$$

denotes the vector of membership degrees to individual descriptors. It is obvious that

$$\mu_{D_i}(x) \geq 0, dla\ i = -k, \dots -1,0,1, \dots k \quad (3)$$

or

$$\sum_{i=-k}^{k} \mu_{D_i}(x) = 1 \quad (4)$$

The distance between two vectors of membership degrees μ(x), μ(y) is expressed using the formula

$$dist(\mu(x), \mu(y)) = \sum_{i=-k}^{k} |\mu_{D_i}(x) - \mu_{D_i}(y)| \quad (5)$$

For any two elements x, y, the following relation holds

$$0 \leq dist(\mu(x), \mu(y)) \leq 2 \quad (6)$$

2. Academic exploration

In article [15], the authors focus on the analysis of time series, which find applications in both scientific research and the everyday operations of many organizations and companies. These areas, relying on time series analysis, encompass a broad spectrum, including the biological sciences and fields related to transportation logistics. In the biological sciences, time series are employed to study the impact of climate change on wildlife and to forecast the necessary management and conservation measures. Meanwhile, in transportation logistics, the primary objective is to achieve maximum efficiency in transportation. In both these domains, changes and trends in various parameters over time are analyzed, and precise comparisons require data, both contemporary and historical, of high quality and reliability. Unfortunately, the quality of historical data often leaves much to be desired, with one of the main challenges being gaps in the data, which are seldom systematically collected over long periods.

The main goal of this work is to propose and test the utility of several fuzzy methods for filling in missing data in studies on the mutual relationships between analyzed time series. The methodology presented by the researchers has been tested on both theoretical and experimental data from various fields, such as ecology (addressing changes in bird arrival dates in the context of climate change) and data related to container transport between ports in the Mediterranean Sea. Importantly, the study applies methods based on fuzzy sets to fill gaps in data from ecological and transportation research.

It is worth mentioning that while the issue of applying fuzzy methods to missing data has been discussed in the literature, the specific areas of application discussed in this work have not been widely explored until now. As part of the proposed solution, the researchers conduct tests on complete data, from which values are randomly removed, and then these artificially introduced gaps are filled. There are several ways of generating missing data:

- Removing a random value in each time series.
- Removing 3 random values in each time series.
- Removing a single common time point for all-time series.
- Removing a maximum of 3 common time points shared by all-time series.

For the introduced artificial gaps in the data, five imputation methods are proposed in two variants: "crisp" and "fuzzy". The "fuzzy" variant is based on the introduction of linguistic descriptors based on fuzzy sets, which are created based on corresponding "crisp" variants. This approach demonstrates high effectiveness, as confirmed in a series of experiments.

In the section addressing the imputation of "crisp" missing data, the authors discuss this method, and its "fuzzy" counterpart is discussed in the next section as a more efficient version. In the general case, it can be assumed that all-time series have values in the set of real numbers. One commonly used technique for imputing missing data in time series (but not limited to them) involves calculating the arithmetic mean for the studied time series in a specified time window, using only those time points for which the series values are known.

Another approach is to use complete information about the behavior of all analyzed series at consecutive time points. If the investigated time series describe similar phenomena, it can be assumed that they behave similarly at specific time points in terms of growth or decline. Based on this assumption, average growths for individual time points or median growths at consecutive time points can be used. Additionally, if there is knowledge of specific groups in the examined subsets, growths for each of these subgroups can be applied independently.

An alternative approach to using traditional statistical measures for imputing missing data is the utilization of fuzzy sets. This approach is logically justified, especially when there is uncertainty about the exact values of measurements. For field data, for example, the precise value of a feature or phenomenon may be challenging to determine, and approximations must be used. Therefore, the use of fuzzy numbers is justified and often recommended.

Fuzzy methods involve applying fuzzy numbers instead of exact values of time series. To simplify the model and streamline the analysis, linguistic descriptors based on the analysis of available data results are introduced. Membership in these descriptors depends on the fuzzy characterization of the analyzed values. Different numbers of descriptors can be used, but 5 or 7 are most commonly employed, providing a good compromise between accuracy and computational complexity. To facilitate calculations, each descriptor is assigned a corresponding numerical value, as presented in the table.

For such symbolic representations of linguistic descriptors, similar analyses can be conducted as in the case of "crisp" descriptors, with a fuzzy representation of time series values. Additionally, each achieved time series value is fuzzified according to the nature of the analyzed data. Triangular membership functions are most commonly used due to their intuitiveness, but other fuzzy membership functions can also be employed.

Within each of the proposed data imputation methods, a comprehensive analysis of linguistic descriptors is carried out, and artificially introduced missing values are imputed, which are then compared with fuzzy input values.

In the section on numerical experiments, the authors present the results of experiments conducted on various types of data, including synthetic data, ecological data concerning bird arrival, and data related to container transport between ports in the Mediterranean Sea. Overall, the experiments showed that methods based on fuzzy sets can significantly increase the effectiveness of imputing missing data compared to "crisp" methods. The use of fuzzy numbers generally yields much better results, especially in the case of ecological and transportation-related data.

In this section, detailed results of experiments and analyses of the effectiveness of different missing data imputation methods are presented, depending on the type of data and experimental conditions. Additionally, the results of experiments for various cases of missing data, such as randomly removing one value in each time series or removing a common time point in multiple time series, are outlined.

In summary, the experimental results suggest that methods based on fuzzy sets can be effective tools for imputing missing data in various fields, especially where data availability is insufficient or data is incomplete.

In article [3], the authors proposed an approach based on advanced fuzzy techniques, such as Fuzzy C-Means and Fuzzy Cognitive Maps, for clustering bird species based on information about the dates of first appearance. Birds are a suitable object for modeling climate change, and the first appearance date

is a widely used indicator for predicting bird migration dynamics, proven to be very useful in previous studies. However, there is still a lack of precise methods for clustering birds in a satisfactory way that allows for detailed information about species and their relationships.

As indicated in the experimental section, the proposed approach allows researchers and practitioners in the field of ecology to observe subtle dependencies between different bird species. Furthermore, this work sheds light on new applications of both Fuzzy C-Means and Fuzzy Cognitive Maps as effective tools for analyzing ecological data collected in a changing climate environment.

The study utilized data on the first appearance dates of birds to investigate phenological changes in response to climate change. Birds are widely used as models for studying climate change, and it is assumed that climate change will affect the shape and dynamics of periodic processes in bird populations, such as migration and breeding. One of the most popular indicators used to study bird migration is the First Appearance Date (FAD) on breeding grounds. The work of this study adds valuable insights into the potential disruptions in food chains caused by phenological changes in FAD, especially in relation to short-distance migratory birds.

The paper also describes bird migration periods and highlights challenges related to the accuracy of FAD data. For example, the argument is often raised that population size can affect the detectability of bird appearance time since there is a higher probability of earlier appearance in larger populations with increased bird activity. Therefore, there is an ongoing search for new statistical tools that can help reduce the limitations of data collection methods and the indicators calculated based on them. The use of fuzzy numbers may be one such solution.

In the further part of the work, the authors described research methodologies such as Fuzzy C-Means (FCM) Clustering, Fuzzy Cognitive Maps, and Fuzzy Rule Classifier. FCM is a classic method for analyzing spatial and temporal data clusters. It divides data into clusters and determines membership in those clusters. It is used to analyze data related to bird migration time.

Next, Fuzzy Cognitive Maps (FCM) were discussed, effective tools for modeling complex structures and data. FCM is a directed graph in which nodes represent factors, and edges reflect cause-and-effect relationships between them. Each node has a value expressing the strength of that factor. The value of a concept is calculated based on the previous state and weight matrix. FCM allows for modeling dependencies between different factors and is useful for analyzing ecological data.

The Fuzzy Rule Classifier was then discussed, a type of fuzzy set-based classification algorithm. It can be used, for example, to classify data based on multiple "if-condition-then-conclusion" rules. This classifier is used in the analysis of bird migration data and allows for assigning birds to different categories based on their characteristics and behaviors.

In the next stage of the work, the authors included the results of experiments, including data descriptions, clustering results using FCM and Fuzzy C-Means, and classification results using the Fuzzy Rule Classifier. Charts representing bird clusters and cognitive maps were presented. The classification results of years into "warm" and "cold" based on bird appearance dates were also discussed, analyzing the effectiveness of this process. It turns out that the classification results of birds by species seem to be more promising than the classification of years based on bird migration dates.

It is worth noting that the entire work is innovative and opens new possibilities for ornithologists. These results suggest that the analysis of bird migration data based on appearance dates can be very promising and opens new avenues for researchers, especially in the context of changing climates. Ultimately, the developed methods and research results have the potential for further development and application in ecology and environmental sciences.

The authors of article [1] highlight that research on the condition of road networks worldwide is a crucial aspect of road infrastructure management. Poland serves as an example, with over 300,000 kilometers of roads primarily under the supervision of local government units, and subsidies being a common practice for financing road maintenance activities. Continuous monitoring of road conditions is essential, especially considering heavy traffic and the need to maintain road surfaces in adequate technical condition.

To effectively manage road conditions and detect various damages such as potholes, many researchers propose the use of crowdsourcing, involving citizens as a data source. With the increasing popularity of smartphones and IoT devices, this approach becomes more attractive. One approach to monitor road conditions involves using vibration measurement systems, which measure acceleration during vehicle travel on the road surface. This process collects vibration data that can indicate the presence of damages like potholes.

Another approach is the use of visual recognition systems, analyzing images from cameras mounted on vehicles or smartphones. These systems can detect road surface damages based on image analysis.

It's worth noting that road surface damages pose a complex problem as potholes vary in shape, depth, and location on the road. Additionally, different types of vehicles have different parameters, such as axle spacing, influencing how they respond to road damages.

In the presented article, the authors describe their approach to detecting road surface damages by utilizing vibration-based data as the main source of information. Their method relies on fuzzy thresholding, aiming to increase the accuracy of damage detection and reduce the number of false alarms.

The research methodology also includes numerical experiments confirming the effectiveness of this method. The researchers analyze various scenarios, including different sizes of time windows for data analysis and different membership functions in the fuzzy system.

An important research aspect is the comparison of the effectiveness of different approaches to road condition monitoring. The authors analyze which methods are more efficient and precise in detecting road damages.

The conclusions drawn from the conducted research suggest that the proposed approach is more effective in terms of accuracy and false alarm reduction compared to previous threshold-based methods. This indicates a promising direction for the development of road condition monitoring systems and detection of road surface damages.

However, there are certain limitations in the research, such as not considering the individual driving style of the driver and the inability to detect road surface damages directly from vehicle data. The authors plan to continue their research, including the implementation of the system in IoT devices and considering more advanced data analyses, such as driver's driving style and techniques for avoiding damages.

In the work [20], the authors present an advanced analysis of fuzzy cognitive maps (FCM), a special tool used for analyzing and modeling various systems in diverse fields such as control systems, risk analysis, and decision-making processes in engineering, IT, medicine, and social sciences. The authors particularly emphasize the modeling of eHealth in the context of Generation Z.

The authors conduct research on the impact of the initial state of the matrix, obtained using both fuzzy and conventional methods, on FCM results. They employ fuzzy techniques such as Fuzzy C-Means and Fuzzy Robust Gamma Rank Correlation, comparing FCM analysis results with clustering outcomes obtained using Fuzzy C-Means and Hierarchical Cluster Analysis algorithms.

When analyzing the behaviors of Internet users, the authors focus mainly on Generation Z. They use the UTAUT model, assuming that there are four constructs significantly influencing the direct determinants of information technology usage: performance expectancy (PE), effort expectancy (EE), social influence (SI), and facilitating conditions (FC). These constructs are analyzed in the context of their impact on the acceptance and use of available health information on the Internet by Generation Z.

As part of the study, a survey was conducted among students at the Technical University of Lublin, with questions related to the predicted variables (PE, EE, SI, FC, and BI) along with 13 indicators. Based on the responses, a cluster analysis was carried out using Fuzzy C-Means and HCA methods, allowing the identification of three different clusters, each representing different variables.

Detailed results obtained from Fuzzy C-Means in conjunction with Fuzzy Cognitive Maps provide a very clear energy graph with several relationships between concepts, allowing for a deep understanding of the impact of different variables on each other. The results of the UTAUT model analysis are also important, showing which constructs have a significant impact on behavioral intentions and which do not.

Importantly, the authors' research provides valuable information and insights into how different variables and constructs affect the acceptance and use of health information on the Internet by Generation Z. These findings are crucial for the development of future eHealth strategies and solutions that will need to be tailored to the unique attitudes and behaviors of this generation.

Below is a table containing various analysis methods and techniques used by researchers, including their main applications and results/discoveries.

Table 1. Analysis methods and techniques used by researchers

Method/Technique	Application	Main Results/Discoveries
Fuzzy Cognitive Maps (FCM)	Modeling complex structures in various fields	Allows for the analysis of relationships between different concepts
Agent-Based Modeling (ABM)	Simulation of the impact of one individual's thinking on the behavior of others	Allows for the analysis of influences between different units
Unified Theory of Acceptance and Use of Technology (UTAUT)	Explaining and predicting Internet users' behaviors	Analysis of constructs influencing the acceptance and use of information technology
Fuzzy C-Means (FCM)	Clustering data in a clustered form	Enables the identification of clusters in data
Robust Gamma Rank Correlation	Determining correlation coefficients for the initial weight matrix in Fuzzy Cognitive Maps (FCM)	Provides the weight matrix for FCM (Fuzzy Cognitive Maps)
Hierarchical Cluster Analysis (HCA)	Splitting the data into different clusters	Reveals natural groupings in the data

In summary, the authors presented a multidimensional approach to analyzing and modeling user behaviors in the context of online health information. They combined advanced data analysis techniques, such as fuzzy cognitive maps, with a deep understanding of Generation Z psychology. The results could have broad implications for public health, psychology, and information technology, providing crucial insights for creating more effective interventions and technology-based solutions.

In [5], the authors conducted detailed research on anomaly detection in the functioning and development of startups. Detecting anomalies in startup activities is crucial for identifying companies with the highest potential for success, given the high failure rate among newly established firms. Anomaly analysis included evaluating different funding methods and entrepreneurial methodologies. The fuzzy set-based classification used in the study allows precise determination of whether a startup exhibits features deviating from the norm, indicating greater potential for success or increased risk of failure. The modified AHP method was a key tool for assessing the innovativeness of funding sources.

In [12], the authors focused on applying fuzzy methods to analyze the genetic population of red deer, particularly examining homozygosity and heterozygosity levels in different population groups. Fuzzy methods, crucial to the study, were used to model uncertainty and imprecision in genetic data. By using fuzzy sets, researchers estimated the degree of genotype membership in homozygous or heterozygous categories, enabling a deeper analysis of genetic dynamics within the population.

In [19], the authors explored challenges and needs in software engineering, IT project management, and programming paradigms. They aimed to understand the essential skills and knowledge required for future IT graduates. The study utilized an enhanced version of the Analytic Hierarchy Process (AHP) with graphical tools and fuzzy logic to efficiently collect group decisions. This approach highlighted differences in perception between experienced IT professionals and students regarding software engineering challenges and where attention is focused in the software life cycle.

In [18], the authors concentrated on anomaly detection in datasets, focusing on improvements to the Isolation Forest method. They discussed various sources of anomalies and their impact on societies and business organizations. The Isolation Forest method, built on binary search tree construction and scoring, was presented, and the authors proposed a modification to improve anomaly detection. The modification involves calculating the cluster's center based on the average value of a selected attribute and proved more intuitive and efficient in identifying anomalies.

These studies demonstrate the diverse applications of fuzzy logic and advanced data analysis techniques in fields such as health information analysis, startup anomaly detection, genetic population analysis, and software engineering research. The authors provided valuable insights and proposed improvements that could contribute to the development of more effective strategies and solutions in these domains.

In [9], the authors emphasize anomaly detection as a significant issue in data analysis due to its numerous applications, spanning fields from medicine to industry. Anomalies can manifest in various ways, such as unusual objects in medical images, suspicious behaviors in videos, erroneous data entered with the intent of sabotage, attempted fraud, or measurement results deviating from the norm. Detecting these anomalies is crucial for minimizing financial losses and preventing disasters in critical systems. The article introduces an enhanced version of the Isolation Forest method, incorporating the K-Medoids algorithm.

The primary goal of the study is to develop and compare the enhanced version of the Isolation Forest method with its original version, aiming to increase anomaly detection effectiveness across different domains. The research methodology and experimental results confirming the utility and efficiency of the new method are the focal points.

The Isolation Forest method involves constructing multiple decision trees, each built on a random subset of data. Anomaly detection relies on assessing how deep in a tree a given observation can be found, with deeper placements indicating a higher likelihood of anomaly. There are several versions of this method with various improvements and extensions.

The enhanced Isolation Forest version utilizes the K-Medoids algorithm for more precise data clustering, helping determine the optimal number of groups into which data can be divided. This, in turn, affects shorter tree searches and a more natural fit for the data, eliminating the need for random data splits.

Experiments were conducted on various datasets, including medical, industrial, and financial data. Results of the new method were compared with the original Isolation Forest and other popular anomaly detection methods. The analysis confirmed that the enhanced method is an effective tool for anomaly detection across different domains, marking a significant development in anomaly detection tools with potential applications in areas where anomaly detection is critical.

In [10], the authors focus on anomaly detection in databases, a key challenge in the field of data analysis. This problem occurs in various types of databases containing numerical, categorical, temporal, mixed, and graphical data. The article introduces the Isolation Forest as a classical anomaly detection technique, relying on the creation of a forest of decision trees based on random data samples. These trees are then used to calculate anomaly scores for individual records in the database.

Importantly, the authors propose combining the Isolation Forest method with the Fuzzy C-Means (FCM) clustering technique. FCM allows obtaining degrees of record membership in clusters, which can be used to improve anomaly detection effectiveness. This is significant as it enables considering record membership degrees in groups when calculating anomaly scores. Numerous numerical experiments were conducted on different datasets to investigate the effectiveness of combining Isolation Forest with FCM. Results suggest that FCM can significantly impact anomaly detection effectiveness and improve result quality. In particular, this method appears stable and effective in detecting anomalies in various types of data. In conclusion, combining Isolation Forest with FCM could be a promising approach in the field of anomaly detection in data. Experimental results suggest that this method may be effective across different domains, such as logistics, where anomaly detection is crucial for safety and economy.

In the study [2], the authors focus on data obtained from an array of acceleration sensors installed in vehicles. Their objective is to identify locations where acceleration values differ from standard readings, indicating the presence of road irregularities such as bumps or speed humps. The authors aim to create a road monitoring system based on data collected by drivers using smartphones. They encourage drivers to collect and transmit data to a system that will analyze it for road quality and report potential issues.

The main goal of the study is the application of the well-known "Isolation Forest" algorithm and its extensions in the context of anomaly detection on roads. The authors compare the results obtained with various variants of the isolation forest and analyze the structure of detected anomalies. Experiments on datasets from real road measurements confirm the effectiveness and applicability of isolation forest-based techniques in analyzing road quality and evaluating other spatial and temporal data.

In the examined area, the authors previously dealt with the classification of road surface profiles using sound data, not acceleration data. Other studies focused on classification and grouping to locate specific anomalies on road maps. Various mobile measurement platforms enable the use of inexpensive devices, such as smartphones, for collecting acceleration data and determining road surface quality indicators.

The study utilizes data from acceleration measurements using smartphones mounted in different vehicles. This data is collected through a specially designed data collection application. The software allows the collection of acceleration measurements in three axes (X, Y, Z) and in the global coordinate system (N, E, Z2). Additionally, GPS location and time data are recorded.

Over the course of several years, data from various types of roads in Poland have been accumulated. The locations of real road artifacts, such as bumps or speed humps, were manually marked by drivers using the application. These data were used to assess the performance of algorithms. Various variants of the "Isolation Forest" method (IF) and its extensions are employed in the study. IF relies on two main stages: the training process, where binary search trees are created based on randomly selected data, and the evaluation process, where the deviation of each record is assessed based on searching these trees. Extensions of this method consider different aspects, such as K-means clustering or fuzzy membership functions. The research results indicate that extensions of the "Isolation Forest" method, especially those based on fuzzy membership functions, significantly improve the effectiveness of detecting road artifacts in acceleration data. High accuracy and a low false alarm rate were achieved, which is crucial for identifying problems on roads. The findings confirm that extensions of the "Isolation Forest" method are a promising tool for detecting road artifacts and other anomalies in acceleration data. They can contribute to improving road quality and safety. In the future, the authors plan to continue research in this area, exploring other method extensions and testing them on different datasets.

In the study [13], the authors aimed to investigate the effectiveness of a proposed modification of the Analytic Hierarchy Process (AHP) method. This modification combines flexible assessment of pairs of alternatives with intuitive graphical interfaces and a fuzzy approach. Instead of the traditional 1-9 scale, values indicated using a slider were converted into degrees of membership to classical scale equivalents. This method takes into account the full range of user uncertainty and is more intuitive than other approaches.

Experiments involved surveying 195 students and university staff regarding the risks associated with the COVID-19 pandemic. Each participant made 10 pairwise comparisons, and the results were analyzed in terms of the impact of different aggregation methods on the final weights assigned to responses. The most crucial element of the study was detecting anomalies in the data. Groups of respondents who completed the survey very quickly or returned to it after some time were identified. The experiment results showed that the choice of data aggregation method had a significant impact on the final results and the hierarchy of the importance of individual issues. It was indicated that the most reliable and comparable results are obtained by using weighted averages or considering CR or μ values. The study suggests that the proposed modification of the AHP method may be useful, especially in situations where participant preferences are less clear-cut. Detecting anomalies in data, such as the time taken to complete the survey, can be an important element in result analysis. The final conclusions in the study emphasize the potential for further development of the method, including parameter optimization and the development of more advanced result aggregation methods.

In the paper [14], the authors note that in recent years, they have confirmed that acoustic features related to patients' voices, obtained through smartphones, constitute promising indicators and can support the diagnosis of Bipolar Disorder (BD). Bipolar Disorder is a serious mental illness characterized by mood swings from health (euthymia) through depression, mania, to mixed states. Early detection of a new episode of this disease is crucial for effective treatment. In the context of the possibility of continuous data collection through smartphone applications, voice analysis has significant potential in monitoring this disorder. Various machine learning approaches have been applied in this field; however, recent works indicate a wide range of accuracies (from 67% upwards) in the classification of phases of Bipolar Affective Disorder. Furthermore, there is still a need to establish common standards and clear guidelines for metrics to be used in designing and evaluating systems supporting the prediction of this mental disorder.

The aim of the study is an experimental evaluation of the performance of various predictive methods in different validation approaches concerning the monitoring of Bipolar Affective Disorder using smartphones. It's important to note that data collected from sensors and target classes (psychiatric labels)

are subject to various sources of uncertainty. Therefore, in addition to the best results achieved by benchmark algorithms, the study considers "fuzzy" approaches and an ensemble approach that combines selected benchmark algorithms. Comprehensive comparisons of the results of different methods are also conducted using a metric based on fuzzy numbers.

A set of benchmark algorithms is employed in the study to address the classification task. These algorithms include State-of-the-Art (SOTA) algorithms, Fuzzy Rule (FR), Probabilistic Neural Network (PNN), Decision Tree (DT), Gradient Boosted Tree (GBT), Random Forest (RF), and Tree Ensemble (TE).

All analyses are conducted using the open-source KNIME platform, chosen for its versatility and access to a wide range of analytical tools and visualization capabilities. These algorithms are used to predict the state of patients based on data collected from smartphones.

A fuzzy number-based metric is applied because the patient's mental state is defined by psychiatrists based on the severity of depression and mania symptoms. These symptoms are measured using two rating scales, namely the Hamilton Depression Rating Scale (HDRS) and the Young Mania Rating Scale (YMRS). Subsequently, the mood state is classified into phases of Bipolar Affective Disorder, such as depression, hypomania/mania, euthymia, and mixed state. In previous studies, researchers adopted different cutoff points on HDRS and YMRS scales to train classifiers.

In the context of this study, the goal is to predict the intensity of symptoms (e.g., the sum of points from HDRS) using data collected from smartphones regarding phone conversations. Additionally, fuzzy numbers are considered to represent acceptable deviations.

Experiments are conducted on real datasets representing speech signals collected from two patients suffering from Bipolar Affective Disorder, who participated in a prospective observational study. For standardization, these patients are labeled as patient A and patient B. During the observational study, each patient was accompanied by a psychiatrist who assessed their mental state, and interviews were conducted with varying frequency, depending on the needs identified by the doctor or the patient. As part of this work, labels obtained from psychiatric assessments are assigned to smartphone data, assuming a ground-truth period of 7 days before the psychiatric assessment and 2 days after it.

In the experiments, various approaches were employed to split the data into a training set and a test set, including:

Random Split, data were randomly divided into training and test sets, with different proportions of elements in the test set (50%, 60%, 70%, 80%, 90%). To eliminate the impact of random division, 10 independent sampling repetitions were conducted for each split, and the results were averaged.

Chronological Split, in this approach, data were divided into training and test sets chronologically, with 50%, 60%, 70%, 80%, and 90% of observations assigned to the training set, and the remaining to the test set. The newest data were assigned to the test set.

Patient-wise Validation, the model was trained on one patient's data and then tested on another patient's data. As a result, two models were built: one based on patient A's data and the other based on patient B's data.

Universal Model with Random Split, this approach involved randomly dividing the data into training and test sets, but data from both patients were combined to maintain the split proportions between training and testing for both patients. Similarly, 10 independent sampling repetitions were conducted for each trial.

The results of the experiments were then "fuzzified" using various membership functions, including triangular and trapezoidal fuzzification functions. The impact of the type and parameters of these functions on prediction results was investigated.

The experimental results presented in this study for two patients are promising. However, further experiments are planned on larger datasets. Future work also includes considering other aggregation methods, such as Choquet integration.

In the paper [6], the authors highlight the growing role of digitization in contemporary society, especially in advanced medicine, and the challenges associated with the vast amount of data generated in these fields. They emphasize the importance of computerization in managing and analyzing this data. Eye-tracking technology is underscored as a valuable tool with various applications, including the assessment of website ergonomics, smart homes, and cognitive load evaluation.

The authors discuss the role of eye-tracking in predicting cognitive load levels, studying cognitive disorders such as Alzheimer's disease, and researching customer behaviors. Eye-tracking signals are utilized to assess cognitive load, and various machine learning algorithms are applied for this purpose.

Anomalies in eye-tracking data are discussed, including errors related to defects and mistakes, as well as anomalies indicating different values of certain features in the examined individuals. The authors mention that detecting anomalies in eye-tracking can be useful in diagnosing diseases such as cerebrovascular disease, schizophrenia, and disorders in children with strabismus and visual impairment.

Various anomaly detection techniques are explored in the paper, such as k-nearest neighbors algorithm, principal component analysis, neural networks, density analysis, tree-based algorithms, clustering algorithms, and statistical methods. The Choquet aggregation operator is introduced as a tool for combining the results of anomaly detection algorithms.

The main objective of the study is to apply anomaly detection algorithms and the Choquet operator to support the decision-making process in the analysis of eye-tracking data. The authors compare the results of machine learning-based techniques with expert assessments and investigate whether aggregation operators, such as the Choquet operator, can identify anomalies similar to those detected by experts.

The paper presents anomaly detection methods used in the study, such as angle-based outlier detection, Isolation Forest, k-nearest neighbors detector, and others. The Choquet operator is explained as a tool for combining the results of these methods.

The results of experiments are presented, including anomaly detection by different methods. Some participants consistently showed anomalies in multiple tests, suggesting potential characteristics of eye movements worth further investigation. Anomalies are also classified based on the number of positive detections by the algorithms.

Researchers also discuss the impact of anomaly removal on classification results and note that removing anomalies usually improves accuracy. The results align with expert assessments, as anomalies tend to exhibit lower signal quality and a higher number of blinks.

In the concluding section of the paper, the authors summarize their research, emphasizing the complementary nature of intelligent techniques and expert assessments in anomaly detection. The authors plan to expand their experiments to larger datasets and explore a wider range of aggregation functions.

In the paper [16], the authors observe that detecting outliers is crucial as they may result from systemic failures or human errors. There are many methods for outlier detection, many of which rely on distance measures between data points or data distributions.

In this article, the authors present an innovative approach that deviates from traditional methods by independently analyzing individual dimensions of the dataset. They then draw conclusions about the entire dataset based on appropriate data aggregation techniques. This approach allows for a more detailed analysis of the dataset and emphasizes specific data features.

The proposed approach utilizes fuzzy modifications of the three-sigma rule to determine the degrees of membership of individual elements to descriptions describing the degree and direction of deviation from the norm. These descriptions characterize the strength and direction of the deviation from the norm, enabling a detailed analysis of the degree of deviation of individual elements. The approach identifies the dimension with the greatest impact on the occurrence of outliers, and the choice of aggregation functions for membership to descriptions for different dimensions allows for a more versatile and widespread application of the proposed method.

The presented approach is a modification of the Intuitively Adaptable Outlier Detector (IOAD) method, which uses fuzzy sets to improve its accuracy. The Fuzzy Rule-Based Outlier Detector (FROD) method employs fuzzy modifications of the three-sigma rule, independently analyzing each dimension of the dataset to determine the degrees of membership to descriptions describing deviations from the norm. Descriptions are labeled as Dk, where k indicates the direction and magnitude of the deviation. Combinations of measures such as arithmetic mean or median combined with standard deviation or quartile deviation are used to define the norm.

The FROD method uses random samples of size N to calculate statistical measures for degrees of membership to descriptions. The sample size (N) depends on the dataset size (L), and multiple iterations (M) are recommended for reliable results. Different aggregation functions, such as mean, maximum, or minimum, can be used in successive iterations.

The effectiveness of the proposed method is evaluated by researchers on artificially generated and empirical datasets, including data on taxi movement in New York and Chicago and data from bike-sharing systems in Chicago. The experiments show that the method performs well in classifying outliers, especially for multidimensional datasets.

Aggregation functions, dataset dimensions, and norm determination measures are factors influencing the method's effectiveness. Numerical experiment results demonstrate that the proposed approach achieves high precision and accuracy compared to other known outlier detection methods. The authors conclude by emphasizing the effectiveness of their proposed approach for detecting outliers in multidimensional datasets. They also highlight that their future work will focus on refining method parameters, investigating the impact of the number of descriptions on outlier detection, and exploring the application of the approach to non-monomodal datasets. This method shows potential as a powerful tool for detecting outliers and deserves further research and development.

In the paper [17], the authors focus primarily on the problem of anomaly detection in data and present an innovative method based on the concept of information granules in the context of telemetry data from European bison in the Białowieża Forest.

The authors begin by introducing the problem of anomaly detection in data, emphasizing its complexity and the extensive research in this field over the years. They explain that there are many anomaly detection methods, including those based on statistical analysis, artificial intelligence, and fuzzy sets.

The authors highlight the importance of anomaly detection in ecology and nature conservation. They point out that ecological data is typically highly variable, making the detection of deviations from the norm crucial. These anomalies may arise from both data errors and atypical animal behavior, impacting the management of human-wildlife coexistence risk and the protection of endangered species.

The main goal of the paper is to present the potential of using information granules for detecting anomalies in telemetry data from European bison. The authors assume that this innovative approach will enable the identification of atypical animal behaviors.

The authors describe the key concept of their proprietary method, which involves transforming input data into elements of the information granule space. This space is multidimensional, where N indicates the number of elements, and k indicates the number of dimensions in the new space.

As part of the proposed method, information granules receive fuzzy semantics. This means that the values of input data are transformed into degrees of membership to sets describing the deviation from the norm. The authors propose classifying anomalies based on the number of dimensions in which the information granule is considered anomalous. They discover various types of anomalies, including 0-dimensional (typical), 1-dimensional (outliers only in one dimension), and n-dimensional (outliers in multiple dimensions). The authors conduct an analysis considering contexts, such as months and different times of the day, enabling more advanced anomaly detection.

In the section on numerical experiments, the authors analyze the behavior of five European bison based on telemetry data. They show that appropriately selecting contexts during data transformation is crucial for anomaly detection.

In summary, the presented work describes an advanced anomaly detection method in the movement of European bison, utilizing information granules and fuzzy semantics. This method allows for a more advanced analysis of telemetry data and the detection of various types of anomalies in animal behavior. Numerical experiments confirm the effectiveness of this method and its usefulness in ecological sciences.

In the paper [7], the authors discuss the development of transport systems in the 20th century and emphasize their significant role in developed societies. However, they also highlight that transport systems have a significant impact on the natural environment, leading to negative consequences such as emissions of greenhouse gases, excessive noise, habitat destruction, and disturbance of ecological balance. Therefore, there is a need to control and optimize transport systems to reduce their environmental impact.

Contemporary transport systems use information technology and monitoring systems for sustainable transport applications. However, such systems must be prepared to capture and process outlier data. Despite the rich variety of algorithms, there is still a lack of tools that can be boldly called universal due to the specificity of each application domain or dataset. For example, seemingly correct transport data may contain peculiar, undesirable anomalies, such as a train departure time later than its arrival time at the destination station. Such anomalies are relatively challenging to detect without knowledge from the application domain or the context of their occurrence.

The authors define the main goal of their work as proposing a tool that allows for an effective assessment of the degree of anomaly of data records in sets containing logistic data. An innovative solution may be the use of the Choquet integral and its general versions known as preaggregation functions to aggregate results from different classifiers that assign a given record to a certain class of anomalies or normal data. An additional goal of this work is to find the most suitable function for a wide range of transport data. In a series of numerical experiments, the authors use various classifiers and four different databases.

In the paper, the authors analyze the possibility of applying the Choquet integral and its enhancements in anomaly detection problems in databases of transport systems. In a series of comprehensive experiments, they demonstrate that the proposed solution is effective in detecting deviations and anomalies in data, not only in individual records but also after analyzing dependencies between record attributes. Finding such discrepancies, errors, defects, and incorrect data in transport-related datasets is one of the main tasks associated with data analysis, and the Choquet integral demonstrates its practical application in this context.

The authors envision future research directions, which may include the application of other types of aggregation functions, such as OWA operators, or the use of other quadratures, such as Simpson's rule. They also note that it is worthwhile to consider the application of other classifiers, such as models based on granular computing or deep learning.

Moreover, it is worth mentioning that research on the use of fuzzy sets and numbers is also conducted by other university employees who do not belong to the discipline of Information Technology and Telecommunications [3, 8, 21, 22].

3. Summary

This work provides a deep insight into the development and applications of fuzzy set theory. Particularly interesting are the areas of application that encompass time series analysis and missing data imputation, which have broad implications in various scientific fields and daily life for organizations and businesses. These methods enable a more flexible and ambiguous approach to data analysis, which can be crucial in situations where data is incomplete, imprecise, or contains errors.

The prospects for further work and research in this field are promising. Areas such as fuzzy classification, outlier detection, and missing data imputation techniques point towards potential new paths of exploration and innovation. There is substantial potential for fuzzy sets in the future development of various scientific domains, emphasizing that further research and exploration in this field are not only possible but also necessary.

In summary, this work highlights the significance of fuzzy sets in modern data analysis and points to diverse and innovative possibilities for their application. The information contained in it serves as a foundation for the continued development of fuzzy set theory and its practical applications, potentially leading to the discovery of new data analysis methods and the development of novel technologies.

References

[1] Badurowicz M., Montusiewicz J., Karczmarek P.: Detection of Road Artefacts Using Fuzzy Adaptive Thresholding. IEEE International Conference on Fuzzy Systems, 2020 [https://doi.org/10.1109/FUZZ48607.2020.9177822].

[2] Badurowicz M., Karczmarek P., Montusiewicz J.: Fuzzy Extensions of Isolation Forests for Road Anomaly Detection. IEEE International Conference on Fuzzy Systems, 2021 [https://doi.org/10.1109/FUZZ45933.2021.9494469].

[3] Bojanowska A. B., Kulisz M.: Using Fuzzy Logic to Make Decisions Based on the Data From Customer Relationship Management Systems. Advances in Science and Technology Research Journal 17(5), 2023 [https://doi.org/10.12913/22998624/172374].

[4] Czerwinski D. et al.: An Application of Fuzzy C-Means, Fuzzy Cognitive Maps, and Fuzzy Rules to Forecasting First Arrival Date of Avian Spring Migrants. IEEE International Conference on Fuzzy Systems, 2020 [https://doi.org/10.1109/FUZZ48607.2020.9177763].

[5] Czerwinski D. et al.: Influence of the Fuzzy Robust Gamma Rank Correlation, Fuzzy C-Means, and Fuzzy Cognitive Maps to Predict the Z Generation's Acceptance Attitudes Towards Internet Health Information. IEEE International Conference on Fuzzy Systems, 2021 [https://doi.org/10.1109/FUZZ45933.2021.9494596].

[6] Dolecki M. et al.: On the Understanding of Anomalies in the Oculography Data and Their Classification with an Application of Fuzzy Aggregators. IEEE International Conference on Fuzzy Systems, 2022 [https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882877].

[7] Gola A., Kłosowski G.: Development of computer-controlled material handling model by means of fuzzy logic and genetic algorithms. Neurocomputing 338, 2019 [https://doi.org/10.1016/j.neucom.2018.05.125].

[8] Jasiulewicz-Kaczmarek M., Żywica P., Gola A.: Fuzzy set theory driven maintenance sustainability performance assessment model: a multiple criteria approach. Journal of Intelligent Manufacturing 32(5), 2021 [https://doi.org/10.1007/s10845-020-01734-3].

[9] Karczmarek P. et al.: K-Medoids Clustering and Fuzzy Sets for Isolation Forest. IEEE International Conference on Fuzzy Systems, 2021 [https://doi.org/10.1109/FUZZ45933.2021.9494460].

[10] Karczmarek P. et al.: Fuzzy C-Means-based Isolation Forest. Applied Soft Computing 106, 2021 [https://doi.org/10.1016/j.asoc.2021.107354].

[11] Kiersztyn A. et al.: Detection and Classification of Anomalies in Large Datasets on the Basis of Information Granules, in IEEE Transactions on Fuzzy Systems 30(8), 2022, 2850–2860 [https://doi.org/10.1109/TFUZZ.2021.3076265].

[12] Kiersztyn A. et al.: Analysis of the Homozygosity of Microsatellite Markers by Using Fuzzy Sets. IEEE International Conference on Fuzzy Systems, 2022 [https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882630].

[13] Kiersztyn A., Kiersztyn K.: Fuzzy Modification of Analytic Hierarchy Process Using GUI Tools. IEEE International Conference on Fuzzy Systems, 2022 [https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882579].

[14] Kaczmarek-Majer K., Kiersztyn A.: Experimental Evaluation of the Accuracy of an Ensemble of Fuzzy Methods for Classification of Episodes in Bipolar Disorder. IEEE International Conference on Fuzzy Systems, 2022 [https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882582].

[15] Kiersztyn A. et al.: Data Imputation in Related Time Series Using Fuzzy Set-Based Techniques. IEEE International Conference on Fuzzy Systems, 2020 [https://doi.org/10.1109/FUZZ48607.2020.9177617].

[16] Kiersztyn K., Kiersztyn A.: Fuzzy Rule-based Outlier Detector. IEEE International Conference on Fuzzy Systems, 2022 [https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882567].

[17] Kiersztyn A. et al.: The use of information granules to detect anomalies in spatial behavior of animals. Ecological Indicators 136, 2022 [https://doi.org/10.1016/j.ecolind.2022.108583].

[18] Karczmarek P., Kiersztyn A., Pedrycz W.: Fuzzy Set-Based Isolation Forest. IEEE International Conference on Fuzzy Systems, 2020 [https://doi.org/10.1109/FUZZ48607.2020.9177718].

[19] Karczmarek P. et al.: The Assessment of Importance of Selected Issues of Software Engineering, IT Project Management, and Programming Paradigms Based on Graphical AHP and Fuzzy C-Means. IEEE International Conference on Fuzzy Systems, 2020 [https://doi.org/10.1109/FUZZ48607.2020.9177591].

[20] Kiersztyn A. et al.: Classification of Companies Based on Fuzzy Levels of Innovation. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2022 [https://doi.org/10.1109/FUZZIEEE55066.2022.9882734].

[21] Witczak M. et al.: A fuzzy logic approach to remaining useful life control and scheduling of cooperating forklifts. In: IEEE CIS International Conference on Fuzzy Systems 2021: Conference Proceedings, 2021, 1–8 [https://doi.org/10.1109/FUZZ45933.2021.9494562].

[22] Wittbrodt P. et al.: Identification of the Impact of the Availability Factor on the Efficiency of Production Processes Using the AHP and Fuzzy AHP Methods. Applied Computer Science 18(4), 2022 [https://doi.org/10.35784/acs-2022-32].

[23] Zadeh L. A.: Fuzzy Sets. Information and Control, 1965.

Ph.D. Maciej Celiński
e-mail: m.celinski@pollub.pl

He received the degree in M.Sc. informatics, faculty of exact sciences at the John Paul II Catholic University of Lublin, Poland. He is currently an assistant at the Faculty of Technology Fundamentals, Lublin University of Technology, Lublin. His current research interests ICT and new technology in education.

http://orcid.org/0000-0001-8412-207X

Ph.D. Adam Kiersztyn
e-mail: a.kiersztyn@pollub.pl

He received the Ph.D. degree in mathematics from Faculty of Mathematics, Physics, and Computer Science, Maria Curie-Skłodowska University, Lublin, Poland, in 2012. He is currently an assistant professor with the Faculty of Mathematics and Technical Computer Science, Lublin University of Technology, Lublin, Poland. His current research interests include fuzzy measures, data mining, data exploration, quantitative methods, and decision-making theory.

http://orcid.org/0000-0001-5222-8101