IRSTI 09.02.23

A. Amankossova[1] , C. Turan[2]
[1, 2] Suleyman Demirel University, Kaskelen, Kazakhstan

AN EVALUATION OF UNSUPERVISED OUTLIER DETECTION METHODS FOR UNIVARIATE TIME SERIES DATA IN FINANCIAL TRANSACTIONS

Abstract. An essential problem in finance application areas is identifying abnormal subsequences in time series data. Despite the wide range of outlier detection algorithms, no substantial research has been conducted to thoroughly investigate and assess the various methodologies, particularly in the financial industry. This study focuses on comparing and contrasting the outcomes of various unsupervised algorithms. The findings reveal that the Local Outlier Factor technique outperforms the other methods in terms of precision, recall, and F1-score. The research provides valuable insights for financial institutions and businesses looking to improve their identification of abnormalities systems and highlights the importance of choosing the appropriate unsupervised outlier detection method for financial transaction data. The outcomes of this study can be used to inform future research and development in the area of financial unusual case detection.

Keywords: univariate time series, comparison, detection techniques, anomaly, financial industry.

***

Аңдатпа. Уақыт қатарларының деректерінде аномальді ішкі реттіліктерді анықтау қаржы салаларындағы маңызды міндет болып табылады. Шектеулерді анықтау алгоритмдерінің кең ауқымына қарамастан, әртүрлі әдістемелерді, әсіресе қаржы индустриясында мұқият зерттеу және бағалау үшін айтарлықтай зерттеулер жүргізілген жоқ. Бұл зерттеу әртүрлі бақыланбайтын алгоритмдердің нәтижелерін салыстыруға және салыстыруға бағытталған. Нәтижелер «Жергілікті шектен тыс фактор» әдісі дәлдік, еске түсіру және F1 ұпайлары бойынша басқа әдістерден басым екенін көрсетеді. Зерттеу ауытқулар жүйелерін анықтауды жақсартуға ұмтылатын қаржы институттары мен бизнес үшін құнды түсініктер береді және қаржылық транзакция деректері үшін сәйкес бақыланбайтын ауытқуларды анықтау әдісін таңдаудың маңыздылығын көрсетеді. Бұл зерттеудің нәтижелерін қаржылық ерекше жағдайды анықтау саласындағы болашақ зерттеулер мен әзірлемелер туралы ақпараттандыру үшін пайдалануға болады.

Түйінді сөздер: бір айнымалы уақыт қатары, әдістерді салыстыру, ауытқушылық, анықтау, қаржы саласы.

***

Аннотация. Существенной проблемой в области финансовых приложений является выявление аномальных подпоследовательностей в данных временных рядов. Несмотря на широкий спектр алгоритмов обнаружения выбросов, не проводилось никаких существенных исследований для тщательного изучения и оценки различных методологий, особенно в финансовой отрасли. Это исследование фокусируется на сравнении и противопоставлении результатов различных неконтролируемых алгоритмов. Полученные данные показывают, что метод фактор локального выброса превосходит другие методы с точки зрения точности, полноты и мери F1. Исследование дает ценную информацию для финансовых учреждений и предприятий, стремящихся улучшить свои системы выявления аномалий, и подчеркивает важность выбора соответствующего метода обнаружения неконтролируемых выбросов для данных финансовых транзакций. Результаты этого исследования могут быть использованы для информирования будущих исследований и разработок в области выявления необычных финансовых случаев.

Ключевые слова: одномерный временной ряд, сравнение методов, отклонение, обнаружение, финансовая индустрия.

I.Introduction

The banking industry is one of the most data-driven industries, and monitoring each transaction is a crucial aspect of its operations. Univariate time series data, which represents the historical trend of transaction frequency, can provide valuable insights into the performance of a portfolio. However, this data is often subject to outliers or anomalies, which can have a significant impact on the accuracy of forecasting and decision-making.

Outlier identification in univariate time series data in banking industry refers to the process of detecting data points in a time series that are considerably distinct from the remaining data. These data points are commonly described as outliers or anomalies. There is still a need for robust and effective methods that can handle the complexities. Outlier detection is important in various applications, such as financial forecasting and monitoring of system performance.

The fundamental purpose is to compare and evaluate the performance of different approaches based on machine learning for spotting outliers in univariate time series data. The research question that will be addressed in this study is “Which approach is best for detecting outliers in time series datasets?”

In recent years, there have been several studies on outlier detection in univariate time series data using unsupervised learning methods. The survey by

Zhang and Zomaya [1] provides a valuable resource for researchers and practitioners in the field of data science, providing a comprehensive overview of the state-of-the-art in outlier detection methods and their applications. The authors' findings highlight the need for further research to develop more effective and efficient outlier detection techniques for large-scale data, as well as to address some of the limitations and challenges of current methods. The paper "Anomaly Detection in Multivariate Time Series Data: A Comparison of Unsupervised and Semi-Supervised Methods" by Shi, Zhang, and Li [2] compares and evaluates the performance of unsupervised and semi-supervised methods for anomaly detection in multidimensional time series data. Some of the related works in this area can be seen in the works of Wang and Yao [3], Motoda [4], and Zimek and Kriegel [5]. These studies have shown promising results and demonstrate the potential of unsupervised learning methods for outlier detection in time series data. However, there is still a need for more comprehensive evaluations of the performance and limitations of these methods in real-world finance organizations. Additionally, the results of these studies have varied depending on the data set used and the evaluation metrics applied, highlighting the need for a more systematic and comparative analysis of unsupervised learning methods for outlier detection in the banking industry.

This investigation will enrich the existing collection of knowledge in the field of outlier detection in univariate time series data and provide practical insights for practitioners in the banking industry. The findings of this study will inform the development of better methods for outlier detection and provide valuable information for decision-makers in the banking industry.

II. Methodology

The methodological part of the research paper on detecting financial transaction outliers in univariate time series data in the banking industry using unsupervised learning techniques includes the following steps: pre-processing time series data of loan amounts, choosing appropriate attributes from pre-processed data and selecting the most efficient algorithm based on performance metrics. This part provides a comprehensive overview of current anomaly detection methods that we look at when evaluating them. This paragraph of the article presents an evaluation of the algorithms' quality.

Machine Learning methods are generally used for complex, high-dimensional problems where the underlying relationships between variables are not well understood. They typically rely on algorithms that learn patterns from data without requiring explicit programming or pre-defined rules. In contrast, statistical methods are often used for simpler, well-understood problems where the underlying relationships between variables are known or assumed.

2.1. One-class Support Vector Machine

One-class Support Vector Machine (One-class SVM) is designed specifically for anomaly detection and can effectively identify deviations from normal

behaviour in time series data. The goal of one-class SVM is to learn the decision boundary that separates the positive class from the rest of the data in a high-dimensional feature space. The algorithm generates a higher dimensional space that best distinguishes the positive class from the origin, which represents the negative class. One-class SVM is particularly useful in applications where the data is highly imbalanced, and it is difficult to find a sufficient number of examples for the negative class. It works by modelling the normal behaviour of a system based on a single class of data and then identifying any observations that are significantly different from the learned normal behaviour as potential anomalies. It requires data from a single class for training, making it ideal for scenarios where data from other classes is not available or is difficult to obtain.

This can be used to identify new and unseen data points that differ from the known data. In this one-class SVM algorithm, an anomalous object is a point in n-dimensional space that does not pass through the hyperplane [6]. One-class SVM assumes that the time series data is stationary, meaning that its statistics have constant characteristics across time. If the data is non-stationary, a one-class SVM may not perform well. Also, it can be computationally expensive, especially for large datasets. This algorithm is limited in its ability to model complex relationships between variables and may not be suitable for all types of time series data. It is suitable when the data in the training set obeys a normal distribution and the test set contains anomalies.

2.2. Isolation Forest

The Isolation Forest algorithm is based on the isolation of sample instances [7, 8].

By utilising isolation, the suggested approach, iForest, is able to utilise sub-sampling to a degree that is not practical with current methods, resulting in a linear time complexity method with a low constant and storage demand [9]. With the "random" way of building trees, outliers will hit the leaves at early stages (at small tree depth), i.e., outliers are easier to "isolate." The extraction of anomalous values occurs at the first iterations of the algorithm. Isolation Forest is a simple and fast algorithm that does not require extensive tuning or complex parameter selection. If the data size is extremely large, the algorithm may become slow. This is because the time complexity of Isolation Forest is linear with the number of samples in the data.

2.3. Local Outlier Factor

Local outlier factor refers to unsupervised machine learning methods, which are important in the context of our study because it is not required to know whether the analysed data samples are normal at the outset, which simplifies calculations. The Local Outlier Factor algorithm is based on calculating the deviation of the local density of a point with respect to its k nearest neighbours [9]. The main parameter of the algorithm is the number of nearest neighbors, Nk(p), where p is the object from which the distance is measured. This algorithm is relatively robust to noise, as it considers the density of data points, rather than relying on

global statistics such as the mean or variance. In the context of a bank, Local Outlier Factor algorithm can be used to detect unusual patterns in financial transactions, such as fraudulent activities or unusual expenditures. The algorithm works by comparing a data point's density in relation to its surrounding data points, and marking points that have a significantly lower density as outliers. The LOF algorithm can be implemented in various programming languages, such as Python or R, and can be applied to time series data in a bank's database to identify potential outliers of transaction frequency. Local outlier factor has a relatively low computational cost, which makes it appropriate for real-time anomaly applications.

2.4. Interquartile Range

The interquartile range (IQR) is a statistical method used to measure the spread or dispersion of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset, which contains 50% of the data. The IQR is therefore a measure of the middle 50% of the data. In outlier detection, the IQR is used to identify values that are significantly different from the rest of the dataset. Specifically, an outlier is defined as a value that falls below Q1 - 1.5*IQR or above Q3 + 1.5*IQR. Values outside this range are considered potential outliers. However, it is important to note that the IQR method is not foolproof and can still produce false positives and false negatives. In addition, it assumes that the dataset follows a normal distribution, which may not always be the case in real-world scenarios. Therefore, it is important to use the IQR method in conjunction with other outlier detection techniques and to carefully examine any values identified as potential outliers.

III. Results

Experimental findings comparing the effectiveness of the strategies previously mentioned are reported in this part. For experimental evaluation, our algorithms were executed on several datasets.

Due to the deficit of a data set for detecting anomalies in the financial sector, univariate time series with anomalies were generated. So, in this research work, synthetic data generated using Python was used as support. Synthetic datasets are artificially generated data sets that are used in research work to simulate real-world scenarios. It is a useful tool for studying the banking industry, as it allows for the simulation of real-world scenarios and data patterns without compromising data privacy and security, and without being limited by data availability or scope. Generating a synthetic univariate timeseries dataset using Python can be done using the NumPy library. First, we created an array of random values using the NumPy random module. This array represents the values of our timeseries. After it, we used the NumPy linspace function to generate a sequence of evenly spaced values that will represent the time intervals in our timeseries.

This is how our Python-generated dataset with 100,000 samples looks.

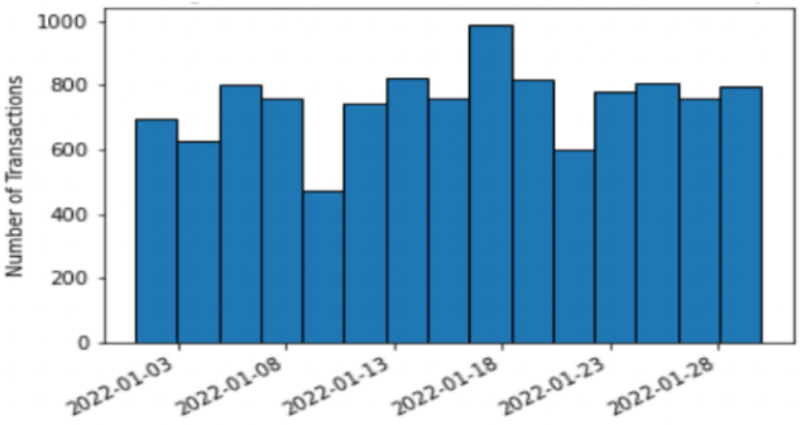

Figure 1: Histogram of Number of Transactions vs Timestamp

Figure 1 presents a histogram visualization of the distribution of the frequency of transactions in the time series for a sample of 100,000 generated data. The X-axis displays the time series data, with each point representing a unique hour. The Y-axis displays the frequency of transactions for each hour, with the frequency represented as the number of transactions. The purpose of this visualization is to investigate the relationship between the frequency of transactions and time series data.

Another dataset that is used in our experiment is Credit card fraud detection dataset. This is a publicly available dataset that contains credit card transactions labeled as fraudulent or genuine. It can be used to evaluate outlier detection methods for detecting abnormal transactions. The dataset has a total of 284,807 transactions, of which 492 are fraudulent. The dataset is highly imbalanced, with the fraudulent transactions accounting for only 0.172% of the total transactions. The dataset contains 30 features, most of which are the result of a PCA transformation to anonymize sensitive information. The features include time, which is the number of seconds elapsed between each transaction and the first transaction in the dataset, as well as various amounts, such as the transaction amount and the amount of money in the cardholder's account.

In this comparison study of outlier detection methods in the banking industry, three algorithms of machine learning and one statistical method were tested. In order for the tests to apply the potentially optimal configuration for each method, I globally optimised the settings of all algorithms. The first step is to identify the hyperparameters that need to be tuned. For Isolation Forest, the main hyperparameters are the number of trees (n_estimators) and the maximum samples used to build each tree (max_samples). For LOF, the main hyperparameter is the number of neighbors used to calculate the outlier score (n_neighbors). For One-Class SVM, the main hyperparameters are the kernel function (kernel) and the regularization parameter (nu). There are several methods for hyperparameter optimization, including grid search, random search,

and Bayesian optimization. Grid search is a simple approach that exhaustively searches the entire hyperparameter space. Random search is similar to grid search, but samples hyperparameters randomly from the search space. Bayesian optimization is a more sophisticated approach that uses a probabilistic model to choose hyperparameters that are likely to improve performance.

The evaluation of the results was carried out utilizing precision, recall and F1 score metrics, as demonstrated in Table 1.

Table 1: Evaluation of unsupervised outlier detection methods

	Recall		**Precision**		**F1-Score**	
	Dataset 1: Synthetic dataset	Dataset 2: Real-world dataset	Dataset 1: Synthetic dataset	Dataset 2: Real-world dataset	Dataset 1: Synthetic dataset	Dataset 2: Real-world dataset
Isolation Forest	86.05%	83.94%	100%	95.38%	92.5%	89.3%
Local Outlier Factor	95.14%	88.06%	95.1%	92.4%	95.11%	90%
One-class SVM	50.4%	90.2%	100%	82.32%	67%	87.6%

To calculate the evaluation metrics for unsupervised machine learning methods on a univariate time series dataset with columns containing timestamps and numerical values, first preprocess the data to convert the timestamps into numerical representations, such as the number of seconds or milliseconds since a certain point in time. Then the data is divided into a training set and a testing set, and the model is used to make predictions on the testing set. Scatter plots demonstrating the results of algorithms for detecting outliers, such as One-class Support Vector Machine, Isolation Forest, and Local Outlier Factor, Interquartile range are presented in Figures 2, 3, 4 and 5, respectively.

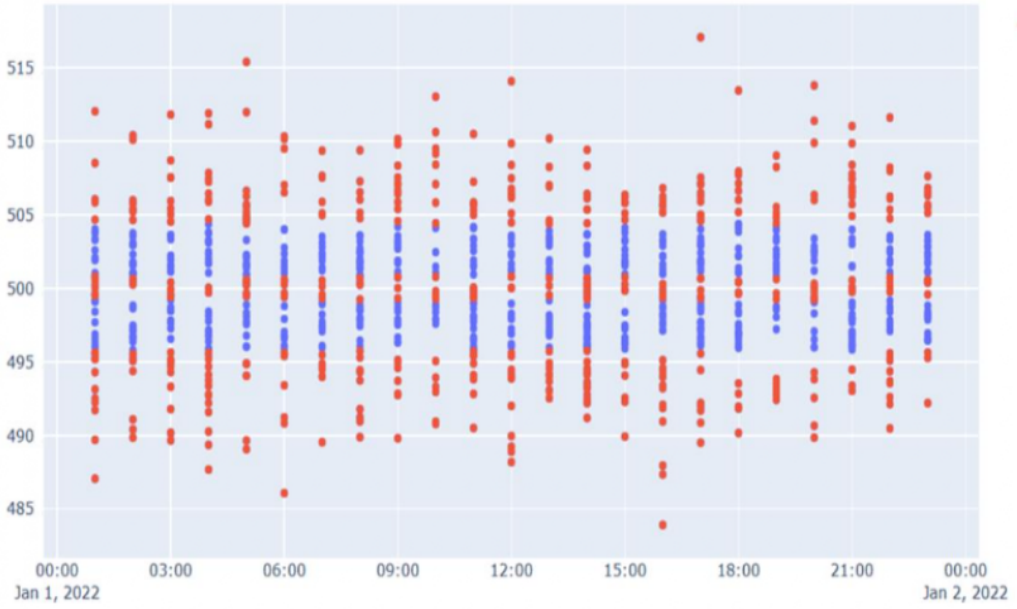

Figure 2: Outlier detection using One-Class SVM

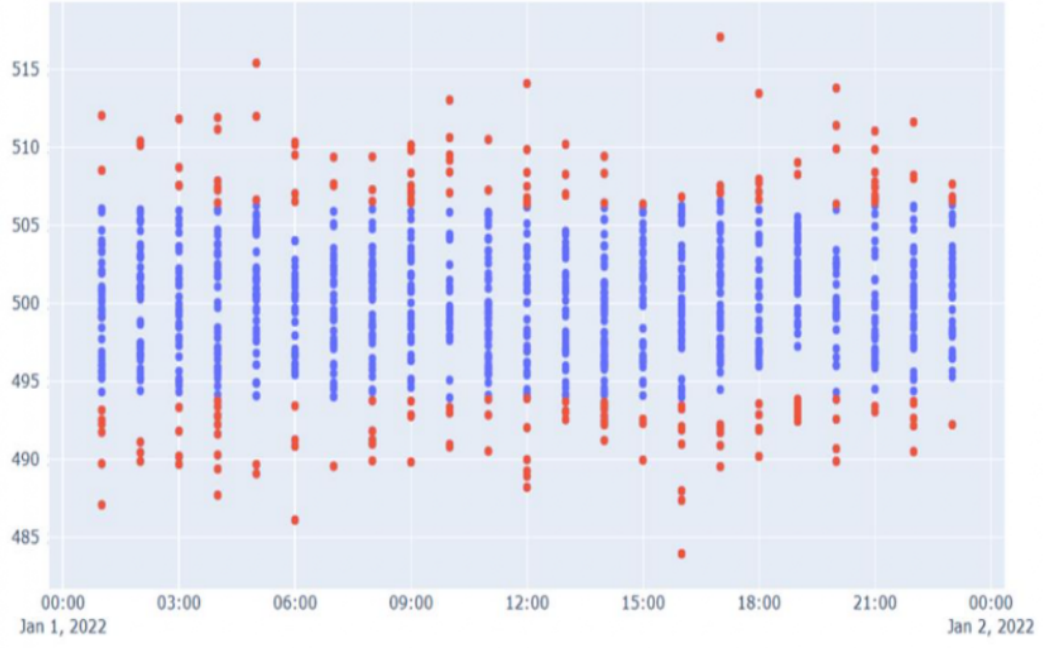

Figure 3: Outlier detection using Isolation Forest

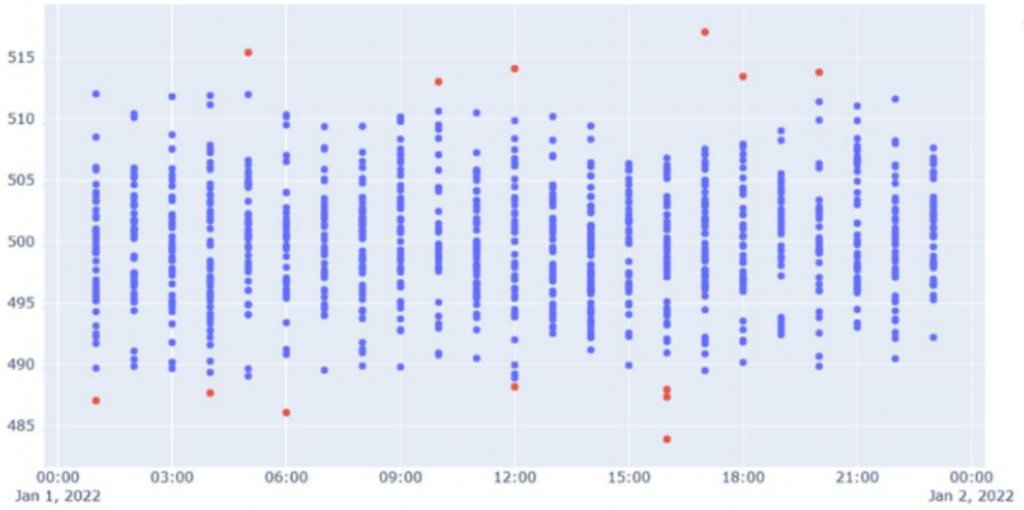

Figure 4: Outlier detection using Local Outlier Factor

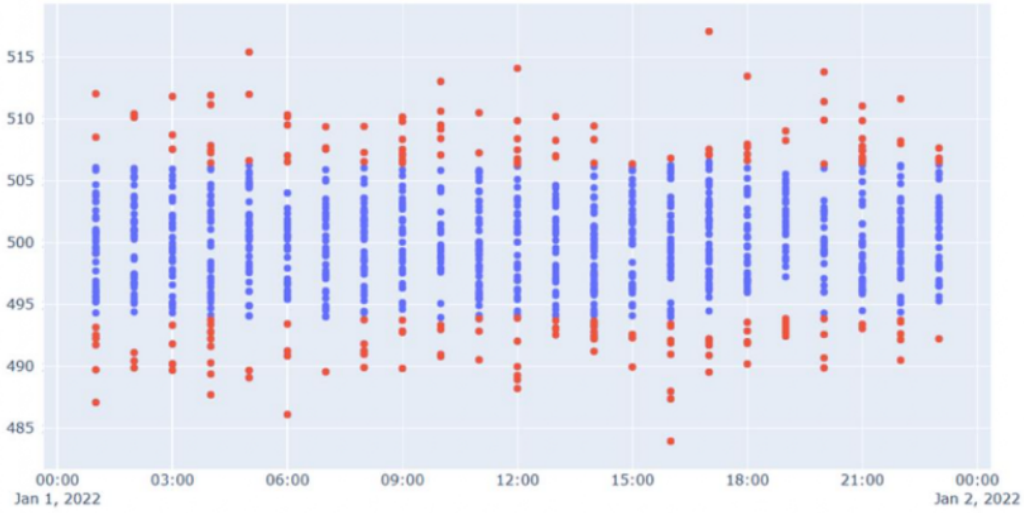

Figure 5: Outlier detection using IQR

In this visualization, the normal instances are plotted as blue dots, while the outliers identified by the method are marked in red. The scatterplot was chosen as the most appropriate type of visualisation for this analysis because it allows for the examination of the relationship between two continuous variables. The x-axis and y-axis can be used to represent the feature values of the instances, and the colour of each point can indicate whether it is a normal instance or an outlier. Visualizing the results of outlier detection methods using scatter plots with red dots to mark outliers can provide a useful tool for understanding and interpreting the results. In each case, the scatter plot can display the normal instances as well as the outliers identified by the method.

Overall, Isolation Forest is designed to handle datasets with complex structures, but it may struggle with complex dependencies in time series datasets. In the IQR method, any value that falls outside the range of 1.5 times the interquartile range (IQR) below the lower quartile or above the upper quartile is considered an outlier. IQR is typically based on modeling the distribution of the data and identifying values that are outside a certain range of expected values. This method is useful when the data is normally distributed and there is a clear understanding of the underlying distribution. One-class SVM is computationally expensive. Isolation Forest and LOF are suitable for datasets with complex geometries and high-dimensional data, while One-class SVM is suitable for datasets with non-linear separable data. IQR is a simple and robust method that can be used for datasets with simple distributions.

In these cases, Local Outlier Factor algorithm was found to be the most effective in detecting outliers in the banking industry datasets, with the highest F1 score.

This suggests that the Local Outlier Factor algorithm could be a useful tool for banks to detect and monitor the frequency of transactions. It is worth noting that the results of this study are specific to the data set and methodology used, and further testing may be required for generalisation to other data sets or industries. Nevertheless, the results provide valuable insights for the use of outlier detection algorithms in the banking industry.

IV. Conclusion

The conclusion of the paper highlights the results of a comprehensive analysis and evaluation of various unsupervised outlier detection machine learning and statistical methods. The study compared and evaluated the performance of multiple approaches. The findings of this research contribute to the field of time series analysis by providing a systematic comparison of outlier detection methods and offering valuable insights into their strengths and limitations. Based on the results, the author recommends the use of the Local Outlier Factor method. In practice, the selection of an outlier detection method should take into consideration both the unique requirement and the features of the dataset. The results of this study serve as a valuable resource for practitioners in the field of data analysis, providing a useful reference for future research in this area.

References

1 Zhang, Zhiyuan, and A. Y. Zomaya. "A Survey on Outlier Detection in Large-Scale Data." Frontiers of Computer Science in China, vol. 4, no. 2, 2010, pp. 167-175.

2 Shi, H., Zhang, H., and Li, C. "Anomaly Detection in Multivariate Time Series Data: A Comparison of Unsupervised and Semi-Supervised

Methods." Journal of Ambient Intelligence and Humanized Computing, vol. 8, no. 1, 2017, pp. 45-61.

3 Wang, Yijun, and Xingquan Yao. "A Time Series Outlier Detection Method Based on Ensemble Empirical Mode Decomposition and Extreme Value Theory." Pattern Recognition, vol. 42, no. 11, 2009, pp. 2763-2774.

4 Hu, Y., Liu, B., and Motoda, H. "Outlier Detection Techniques." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 1, 2015, pp. 69-84.

5 Zimek, Arthur, Hans-Peter Kriegel, and Erich Schubert. "On the Evaluation of Outlier Rankings and Outlier Scores." Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 483-491.

6 Rasulova, A., & Izmailova, A. (2022). Identification of Unique Lakes of Different Origin by Machine Learning Methods. Bulletin of Science and Practice, 8(12), 180-194. (in Russian). https://doi.org/10.33619/2414-2948/85/2

7 Liu F. T., Ting K. M., Zhou Z. H. Isolation forest // 2008 eighth IEEE international conference on data mining. IEEE, 2008. P. 413-422. https://doi.org/10.1109/ICDM.2008.17

8 Liu F. T., Ting K. M., Zhou Z. H. Isolation-based anomaly detection //ACM Transactions on Knowledge Discovery from Data (TKDD). 2012. V. 6. №1. P. 1-39. https://doi.org/10.1145/2133360.2133363

9 Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest." Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422. 2008. IEEE, doi:10.1109/ICDM.2008.17.

10 Rocke, David M. and David L. Woodruff. "Identification of Outliers in Multivariate Data." Journal of the American Statistical Association 91 (1996): 1047-1061