Outlier Detection in Multivariate Time Series Data using a Fusion of K-Medoid, Standardized Euclidean Distance and Z-Score

Nwodo Benita Chikodili (✉) 🆔, Mohammed D. Abdulmalik 🆔,
Opeyemi A. Abisoye 🆔, and Sulaimon A. Bashir 🆔

Department of Computer Science, Federal University of Technology, Minna, Nigeria.

bennychika2@gmail.com, drmalik@futminna.edu.ng,
o.a.abisoye@futminna.edu.ng,
bashirsulaimon@futminna.edu.ng

Abstract. Data mining technique has been used to extract potentially useful knowledge from big data. However, data mining sometimes faces the issue of incorrect results which could be due to the presence of an outlier in the analyzed data. In the literature, it has been identified that the detection of this outlier could enhance the quality of the dataset. An important type of data that requires outlier detection for accurate prediction and enhanced decision making is time series data. Time series data are valuable as it helps to understand the past behavior which is helpful for future predictions hence, it is important to detect the presence of outliers in time series dataset. This paper proposes an algorithm for outlier detection in Multivariate Time Series (MTS) data based on a fusion of K-medoid, Standard Euclidean Distance (SED), and Z-score. Apart from SED, experiments were also performed on two other distance metrics which are City Block and Euclidean Distance. Z-score performance was compared to that of inter-quartile. However, the result obtained showed that the Z-score technique produced a better outlier detection result of 0.9978 F-measure as compared to inter-quartile of 0.8571 F-measure. Furthermore, SED performed better when combined with both Z-score and inter-quartile than City Block and Euclidean Distance.

Keywords: Outlier Detection, Time Series Data, Multivariate, Outliers, K-Medoid, Euclidean distance, Z-Scores, City block

1 Introduction

Currently, the volume of data is significantly increasing with each passing day in various application fields, hence, it is essential to use more effective technology to examine and manage these volumes of data to discover the integral, previously unknown, and potentially useful knowledge. One key data processing technology that is used to examine and manage data is data mining. Data mining is a method of obtaining valuable information from unprocessed-data [1]. With data mining, a search engine can be used

to analyze massive quantities of data and detect interesting trends automatically without needing human involvement [2]. Data mining is generally about finding non-trivial, concealed, and important information from different data types. There are many fields in which data mining is commonly used: retail, network security, financial data processing, biochemical data analysis, telecommunications industry, and other research applications [1].

One of the fundamental issues of data mining is to achieve a profound and accurate conclusion of the study. And outlier detection is a crucial step in solving this basic problem [3, 4]. An outlier is a data point that does not fit into the usual data set classification points. Outlier detection has major applications in data preprocessing and the elimination of abnormal points for stock price prediction, credit scoring, advertisement, intrusion detection, system evaluation, and e-mail spam identification among others [5]. Outlier detection is also known in some researches as anomaly detection and is a critical longstanding research concern in the fields of data mining and statistics. The principal principle behind finding an outlier is to locate unusual points in the data points. Outlier detection points out artifacts that often deviate from a given set of data[6] and Outliers may potentially contain some valuable information [7]. For example, the identification of unusual trends in patient databases can be helpful for disease identification in clinical applications [5, 8].

Time series data is an essential form of data that needs outlier detection for accurate prediction and effective decision making. Data from time series are valuable because it helps to explain the past actions that would be useful for future predictions [9, 10]. Time series data can be used for business forecasting which in turn can improve business decision making. These data are applicable in various fields such as mathematics, sensor processing, pattern recognition, economics, mathematical finance, weather monitoring, earthquake prediction, electromyography, industrial engineering, astronomy, telecommunications and, in large part, any field of applied engineering and science[11] . The identification of outliers in time series data is therefore critical for better decision making and prediction. Considering the importance of outlier detection in time series data, numerous researches have been carried out using various types of algorithms in the success of an outlier detection function in time series data. Algorithms that have been proposed in the literature for outlier detection in multivariate time series dataset include genetic algorithm, Mahalanobis distance, Euclidean distance, autoregressive integrated moving average with exogenous inputs (ARIMAX) model and k-means algorithm while z-score and inter-quartile(box plot) algorithm have been used for univariate time series dataset. This paper proposes a method that combines K-Medoid which is a cluster-based algorithm, Standard Euclidean Distance which is a distance-based, and Z-Score algorithm for outlier detection in multivariate time series (MTS) data. This study's principal contributions are as follows:

1. A fusion of cluster-based and distance-based algorithms for improved outlier detection.
2. Comparative experimentation of different distance-based algorithm on the data cluster obtained from the clustering algorithm.
3. Extension of Z-score and Box plot algorithms for effective outlier detection in MTS dataset.

The organization of these studies is as follows: section two presents the relevant work, section three presents the methods used, section four describes the findings and discussion and finally, in section five and section six conclusions were drawn and recommendations for future works were presented.

2   Related Work

The presence of outliers in time series and non-time series data could provide important information that can be used by an analyst in drawing conclusions. Also, an outlier presence could alter the result produced during analysis. Hence several types of research have been carried to detect these outliers and even remove them when necessary.

Jones et al. [12] performed outlier detection in physical activity using the k-means clustering algorithm. In this work, a FilterK technique was introduced to boost the quality of clusters of k-means based on physical activity. The FilterK algorithm uses an outlier score function that allocates a level of abnormality based mostly on the standardized score obtained from each of its 3 tests which are average distance to neighbors, distance to the closest centroid, and density of the neighborhood. This allows the setting of a standard outlier threshold score as an exclusion criterion when searching the data for abnormalities on the accelerometer. However, the robustness, strength, and weakness of this proposed method were not tested as the method was tested on a small range of accelerometer dataset.

Souza, Aquino and Gomes [13] integrated tensor decomposition with data categorization to outliers in urban spaces detect applications and provide useful information for urban planning and operation. This method consists of three stages which are dimensionality reduction stage in which several latent variables are derived from the contraction, classification of latent variables stage in which the latent variables produced are being used for classification to obtain high-quality groups from the factorization stage, and finally, the production of a polished environmental sequence identification phase that deals with the design of a monitoring process statistics to detect events outside the regularity trends of the measured dataset. A drawback of this proposed approach is that it needs a large sample size to classify the outliers and this necessity has an effect on applications in real-time.

Erkus and Purutc [4] proposed a non-parametric method for quasi-periodic outliers detection in time series data based on a combination of frequency-domain and Fourier transform definition. The result of the algorithm was compared to four existing algorithms namely: Grubbs, box-plot, autonomous anomaly detection, and z-score method, and it was proven that the developed approach worked even better if the data had quasi-periodic structures from start to finish. However, this method did not consider outlier detection for high dimensional datasets as only outliers in univariate dimensions were considered.

Wang et al. [14] considered outliers detection based on its importance in air pollution forecasting. It was found that the identification and correction of the outlier point of the original time series affect air pollution prediction. To improve their air quality index forecasting, a novel hybrid approach based on outlier detection, corrective algorithm, and a heuristic intelligent optimizing algorithm was developed. To achieve outlier identification and correction, the Hampel recognition system was based on measured local

median and standard deviation was adopted. Only the air quality index time series was considered while other influencing factors were not considered in this work.

Borah and Nath [15] proposed a method for solving the problem of outlier detection for incremental medical data, as most of the current outlier recognition algorithms are capable of controlling just static data, and must therefore perform incremental data from scratch. Rare pattern-based outlier detection (RPOD) technique was used to carry out this task. The RPOD consists of two stages: the full info on the database is stored in a compact prefix-tree structure at the first stage. So then the desired set of uncommon patterns is extracted progressively from the prefix-tree by skimming down the index only once. While the second phase includes the identification of outliers based on the unusual patterns obtained during the first stage. The outliers were identified using 3 techniques of outlier detection: Rare Pattern Support Deviation Factor (RPSDF), Transaction Outlier Factor (TOF), and Rare Pattern Outlier Factor (RPOF), respectively. The drawback of this approach is that by setting a minimum support level, RPOD creates outliers based upon the notion of uncommon patterns. And the number of uncommon patterns produced is basically too large and it can be very cumbersome to assign an acceptable support value for outlier detection

Ghallab, Fahmy and Nasr [16] model called NRDD-DBSCAN focused on density-dependent spatial clustering of noise-based applications (DBSCAN) algorithms and using resilient distributed data sets (RDDs) to identify anomalies that affect Internet of Things (IoT) technology data quality. The NRDD-DBSCAN has been used to solve the RDD-DBSCAN system's low-dimensionality problem and also to solve the DBSCAN algorithm issue not being able to handle IoT data. And NRDD-DBSCAN can also be used to boost the efficiency of the present data in IoT devices and applications. The primary drawback of this proposed system is that the data reduction and distribution process uses principal component analysis (PCA) to reduce the dimensionality of the datasets, however, PCA works well by reducing the dimensionality of a strongly correlated linear n-dimension data that prevents the system from processing non-linear data.

3   Methodology

This section discusses the techniques that were used to carry out this research work. Figure 1 shows the processes and techniques that were used to achieve the research goal. Each of the steps shown in the diagram in Figure 1 is discussed in the sub-sections below.
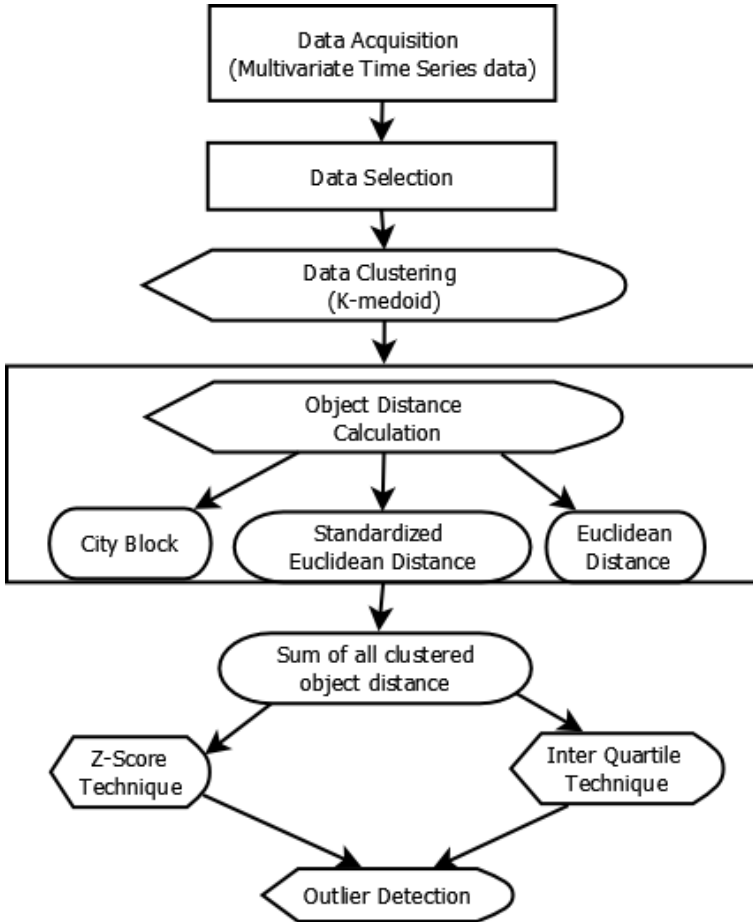
Fig. 1. Proposed System

3.1   Data Acquisition

The time series dataset used in this study was gotten from the Time series section in the UCI repository. Two labeled multivariate time series datasets were used which are the occupancy dataset and the MHEALTH dataset. The MHEALTH which is a multivariate time series dataset consists of observations of body movement and vital signs for 10 individuals of different personalities while undertaking 12 physical activities, and it consists of 23 feature vectors. The occupancy data set defines the dimensions of a room and is intended to determine whether space is occupied or not and it has 20,560 one minute observations taken over a period of few weeks.

3.2   Data Selection

In this work, not all instances in the dataset were used and this was done in order to make the dataset suitable for the outlier detection task. For the MHealth dataset 2000 instances were used out of 161280 instances contained in the MHealth dataset. Out of

the 2000 instances used, 1975 instances are labeled as a normal class while 25 instances were inserted as outliers. For the occupancy dataset, 1500 observations were taken out of 9753 observations contained in the dataset and the 1500 observations used consist of 1485 occupied rooms observations labeled as a normal class and 15 empty room observations identified as outliers.

3.3   Data Clustering (K-Medoid)

In the literature, the clustering result of the k-medoid clustering method cannot be influenced by the presence of outliers as data points are selected to be the medoids. K-medoid algorithm is a partitioning clustering algorithm that divides a set of n objects into k number of clusters. This approach is widely used in areas that demand robustness to outer data, arbitrary distance measures, or those that do not have a precise definition of mean and median [17]. Given objects with p variables that will be grouped into k(k < n) clusters. Defining the jth vector of object i as Yi j (i = 1, ...,n; j = 1, ...,p).K-medoid algorithm consists of three steps which are:

1. Selecting initial medoids: using Euclidean distance as a metric of dissimilarity, compute the difference between each pair of objects using the formula below:

$$D_{ij} = \sqrt{\sum_{a=1}^{p}(Y_{ia} - Y_{ja})^2} \qquad i = 1, \dots, n; j = 1, \dots, n \qquad (1)$$

2. Calculate p_{ij} by using the formula in equation 2 below to make an initial guess at the cluster centers.

$$p_{ij} = \frac{d_{ij}}{\sum_{i=1}^{n} d_{ij}} \qquad i = 1, \dots, n; j = 1, \dots, n \qquad (2)$$

3. Replace current medoid with the object in each cluster which minimizes the distance to other objects in its cluster.

In this study, k was assigned a value of three (3).

3.4   Object Distance

After grouping data into clusters using the k-medoid algorithm, the distance of each object from the medoid of each cluster was calculated using three distance-based algorithms which are: the SED, City Block, and the Euclidean distance. Considering the m-by-n data matrix X, which will be handled as m (1-by-n) row vectors x1, x2,...xm, the different distances between both the xs and xt vectors are stated as follows:

1. **Euclidean Distance:** A linear distance in Euclidean space between two points[18].

$$D_{st}^2 = (x_s - x_t)(x_s - x_t)' \qquad (3)$$

2. **Standardized Euclidean Distance (SED):**

$$D_{st}^2 = (x_s - x_t)U^{-1}(x_s - x_t)' \qquad (4)$$

Where U is the n-by-n diagonal matrix with diagonal elements given by s_j^2 , which signifies the variance of the variable x_j over the m objects.

3. **City Block Distance**: the city block distance between two vectors x_s and x_t in an n-dimensional vector space with static Cartesian coordinate is the addition of the lengths of the projections of the line section between the points onto the coordinate axes. The formula is as follows:

$$D_{st} = \sum_{j=1}^{n} |x_{sj} - x_{tj}| \tag{5}$$

3.5   Distance Summation

After using the above distance metrics to calculate the distance of objects from each cluster medoid, this calculated distances for all clusters are summed together for each object. Hence a single distance gotten from the summation of all an object distance from each cluster medoid is used as an input for the inter-quartile and z-score technique. Given that there are i (i =1,…,n) instances grouped into 3 clusters with various calculated distance (Ci1, Ci2, Ci3). This distance summation is given as:

$$d_s = C_{i1} + C_{i2} + C_{i3} \tag{6}$$

3.6   Outlier Detection Techniques

Two techniques were considered for outlier detection and these techniques are the z-core and inter-quartile techniques.

1. **Z-core technique:** This is a statistical measurement of a score's correlation to the mean in a collection of scores [19]. A Z-score of zero indicates that the score is the same as the average. It can also be a negative/positive value, signifying if it is below or above the average and by how many standard deviations [20]. Z-score makes use of average and standard deviation values to detect outliers in a dataset. Z-score is calculated as follows:

$$Z_{score}(i) = \frac{x_i - \mu}{sd} \tag{7}$$

Where μ = distribution mean, sd= standard deviation and x_i = each object in the distribution. The standard deviation can be calculated using:

$$sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2} \tag{8}$$

To detect outliers based on z-score a cut-off value of 3 was used. That is any distance value with an integer value greater than 3 are detected as outliers.

2. **Inter-quartile technique:** this technique is also known as box plot method. It is called the box plot technique because it forms a box chart, which shows information such as the upper and lower outliers of the univariate data set, lower and upper quartiles, and the median. This method has only been used to detect outliers in a univariate dataset but this study extends its usage to a multivariate dataset. This extension is possible as the sum of the distances was used as a single univariate input to the

box-plot and based on input the technique was able to detect outliers in a multivariate dataset. The rule for this method is as follows.

a. Calculate the interquartile range (IQR)

$$IQR = Q_3 - Q_1 \tag{9}$$

b. Calculate the lower and upper internal boundaries using the formulas :

$$LIB = Q_1 - 1.5IQR \tag{10}$$

$$UIB = Q_3 - 1.5IQR \tag{11}$$

c. Calculate the lower and upper external boundaries using the formulas.

$$LEB = Q_1 - 3IQR \; Q\_1\text{-}3IQR \tag{12}$$

$$UEB = Q_3 - 3IQR \tag{13}$$

d. Observation values between internal and external boundaries are defined as possible outliers.
e. Observation values outside the external boundaries are determined as outliers.

3.7   Performance Metrics

Four performance measures were used to evaluate the proposed method. These measures are explained below.

- **Precision:** is a measure that evaluates the number of accurate predictions made correctly. It is determined as the ratio of correctly predicted positive examples, divided by the total number of predicted positive examples.

$$Precision = True\_Positives / (True\_Positives + False\_Positives) \tag{14}$$

- **Recall:** is a statistic that evaluates the amount of accurate positive predictions that could have been made from all the positive predictions.

$$Recall = True\_Positives/(True\_Positives + False\_Negatives) \tag{15}$$

- **F-measure:** It is the harmonious measure of precision and the recall

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{16}$$

- **Accuracy:** Accuracy is basically defined as the rate of correct classifications. And it is calculated as follows:

$$Accuracy = \frac{True\ Positive\ +\ True\ negative}{True\ Positive\ +\ True\ negative\ +\ False\ Positive\ +\ False\ negative} \tag{17}$$

This study deals with imbalanced classification problems where the number of examples in the dataset for each class label is not balanced, that is the distribution of examples across the known class is skewed. For example, in the occupancy dataset, only 15 examples are of class 1 (outliers) while the remaining 904 examples are class 0 (normal). Hence the most important performance metrics considered for evaluation are precision, recall, and f-measure which evaluates imbalanced classification problems effectively.

3.8   Algorithm Fusion

After performing the data acquisition and data selection process. The K-medoid, Standard Euclidean Distance (SED), and Z-score were combined as follows: The data were then grouped into 3 clusters using the k-medoid algorithm. After clustering the distance of each object from the medoid of each cluster was calculated using Standard Euclidean Distance. Given that there are 3 clusters 3 different SED distances were obtained for each object from each cluster. In order to utilize the 3 SED distances obtained for each object a single distance for each object was generated by summing all of the 3 cluster distances. This generated single distance was used as input data to the z-score technique. The z–score of the SED distances was computed and based on a threshold value of the z-score a data was identified as an outlier.

4   Results and Discussion

In this study, outlier detection was performed using K-Medoid, SED, and Z-Score. The performance of SED was compared to that of Euclidean Distance and City clock. And the performance of z-score was compared to that of inter-quartile. The algorithms were tested using the MHealth and Occupancy multivariate time series dataset. The results obtained are presented in tables 1, 2, 3, and 4 below. Table 1 and table 2 shows the results obtained for MHealth multivariate time series dataset.

Table 1. Results of inter-quartile Outlier Detection technique for the three different distance metric for MHealth dataset

Outlier Detection Algorithm	Precision	Recall	F-Measure	Accuracy
K-Medoid + SED + inter-quartile	0.2778	1.0000	0.4348	0.9133
K-Medoid + Euclidean + inter-quartile	0.2400	0.933	0.3889	0.9022
K-Medoid + City Block + inter-quartile	0.2727	1.0000	0.4286	0.9111

The results obtained for inter-quartile outlier detection technique based on the three different distance metric is shown in table 1 above. Table 1 shows that standardized Euclidean distance performed better with an f-measure of 0.4348 as compared to Euclidean and City block distance with f-measure of 0.3889 and 0.4286 respectively.

Table 2. Results of Z-score Outlier Detection technique for the three different distance metric for MHealth dataset

Outlier Detection Algorithm	Precision	Recall	F-Measure	Accuracy
K-Medoid + SED + Z-score (Proposed System)	0.9704	0.9793	0.9748	0.9511
K-Medoid + Euclidean + Z-score	0.9748	0.9770	0.9759	0.9533
K-Medoid + City Block + Z-score	0.9725	0.9747	0.9736	0.9489

Table 2 shows the results obtained for the z-score detection technique based on the three different distance metrics for the MHealth dataset. Based on the precision, recall, and f-measure obtained in table 2 it can be seen that the z-score technique performs better with f-measures of 0.9748, 0.9759, and 0.9736 for Standardized Euclidean, Euclidean, and city block respectively than the inter-quartile technique with f-measure of 0.4348, 0.3889 and 0.4286 for Standardized Euclidean, Euclidean and city block respectively. Table 3 and table 4 shows the results obtained for Occupancy multivariate time series dataset.

Table 3. Results of inter-quartile Outlier Detection technique for the three different distance metric for Occupancy dataset

OUTLIER DETECTION ALGORITHM	PRECISION	RECALL	F-MEASURE	ACCURACY
K-Medoid + SED + inter-quartile	0.7500	1.0000	0.8571	**0.**9946
K-Medoid + Euclidean + inter-quartile	0.4286	1.0000	0.6000	0.9782
K-Medoid + City Block + inter-quartile	0.7500	1.000	0.8571	0.9946

Table 3 shows the results obtained for the inter-quartile outlier detection technique based on the three different distance metrics for the occupancy dataset. From table 3 it can be seen that standardized Euclidean distance and city block distance performed better with an f-measure of 0. 8571 as compared to Euclidean with an f-measure of 0. 6000.

Table 4. Results of Z-score Outlier Detection technique for the three different distance metric for Occupancy dataset

Outlier Detection Algorithm	Precision	Recall	F-Measure	Accuracy
K-Medoid + SED + Z-score (Proposed System)	1.000	0.9956	0.9978	0.9956
K-Medoid + Euclidean + Z-score	1.000	0. 9934	0.9967	0. 9935
K-Medoid + City Block + Z-score	1.000	0.9956	0.9978	0.9956

Table 4 shows the results obtained for z-score detection technique based on the three different distance metrics for the Occupancy dataset. Based on the precision, recall, and f-measure obtained in table 4 it can be seen that the z-score technique still performs

better with f-measures of 0.9978, 0.9967, and 0.9978 for Standardized Euclidean, Euclidean and city block respectively than the inter-quartile technique with f-measure of 0.8571, 0.6000, and 0.8571 for Standardized Euclidean, Euclidean, and city block respectively.

From the results obtained for the inter-quartile technique for both occupancy and MHealth dataset as shown in table 3 and table 1 respectively, it can be seen that squared Euclidean distance performs better than Euclidean and city block distance. Hence it can be concluded that the inter-quartile technique performs better when combined with squared Euclidean distance. Also from table 2 and table 4, it can be seen from the precision, recall, and f-measure that all three distance metrics produce good and satisfactory results when combined with z-score. However, in both Mhealth and occupancy dataset z-score produces a better result as compared to inter-quartile technique. With high precision, recall, f-measure, and accuracy it can be deduced that outliers can be detected effectively with z-score when combined with any of the distance metric especially with the Standardized Euclidean distance. Hence a system which combines k-medoid, SED and z-score is proposed for effective outlier detection in multivariate time series dataset.

5   Conclusion

This study was able to extend the inter-quartile and z-score technique for outlier detection on multivariate time series dataset as compared to existing works which used these techniques for outlier detection on only univariate datasets. This is due to the ability of the algorithm to combine the multivariate features into a univariate feature. Also, this study was able to detect outliers more effectively as compared to existing works on outlier detection. From the study, it can be concluded that the z-score technique performs better than the inter-quartile technique for outlier detection.

6   Future Works

In this study, only the extreme outlier values generated by the inter-quartile techniques were identified as outliers while the mild identified outlier values were ignored and assigned to the normal class because the data used are skewed it caused a large number of observations to be determined as outliers in the data. This is due to the use of the interquartile distances and lower and upper quadrants measured without considering the skewness of the data set. Hence is it recommended that the skewness of the dataset should be considered to improve the performance of the inter-quartile technique. Three distance metric was used in this work, for future work other distance metrics can be used. This work can be extended for outlier detection for other non-time series datasets.

7   References

1.   S. Jain, F. Sahib, A. Kaur, and F. Sahib, "A Review Paper on Comparison of Clustering Algorithms based on Outliers," vol. 3, no. 05, pp. 178–182, 2016.

2. S. Rajagopal, "Customer Data Clustering using Data Mining Technique," vol. 3, no. 4, 2011, doi: 10.5121/ijdms.2011.3401.

3. K. B. Ramesh and K. K. V Aljinu, "A Survey on Outlier Detection Techniques in Dynamic Data Stream," *Int. J. Latest Eng. Manag. Res. IJLEMR*, vol. 02, no. 08, pp. 23–30, 2017.

4. E. C. Erkus and V. Purutc, "Journal Pre-proof," *Eur. J. Oper. Res.*, 2020, doi: 10.1016/j.ejor.2020.01.014.

5. R. Pamula, J. K. Deka, and S. Nandi, "An outlier detection method based on clustering," in *Proceedings - 2nd International Conference on Emerging Applications of Information Technology, EAIT 2011*, 2011, no. February, pp. 253–256, doi: 10.1109/EAIT.2011.25.

6. H. N. Akouemo and R. J. Povinelli, "Time series outlier detection and imputation," pp. 1–5, 2014.

7. R. Jiadong, L. Hongna, H. Changzhen, and H. Haitao, "ODMC: Outlier Detection on Multivariate Time Series Data based on Clustering," *J. Converg. Inf. Technol.*, vol. 6, no. 2, pp. 70–77, Feb. 2011, doi: 10.4156/jcit.vol6.issue2.8.

8. J. Ren, H. Li, C. Hu, and H. He, "ODMC : Outlier Detection on Multivariate Time Series Data based on," *J. Converg. Inf. Technol.*, vol. 6, no. 2, pp. 70–77, 2011, doi: 10.4156/jcit.vol6.

9. S. Liu, A. Wright, and M. Hauskrecht, "Online Conditional Outlier Detection in Nonstationary Time Series," *Assoc. Adv. Artif. Intell.* 2017.

10. A. Abayomi-Alli, M. O. Odusami, O. Abayomi-Alli, S. Misra, and G. F. Ibeh, "Long Short-Term Memory Model for Time Series Prediction and Forecast of Solar Radiation and other Weather Parameters," in *2019 19th International Conference on Computational Science and Its Applications (ICCSA)*, Saint Petersburg, Russia, Jul. 2019, pp. 82–92, doi: 10.1109/ICCSA.2019.00004.

11. E. A. Hasan, "A Method for Detection of Outliers in Time Series Data," *Int. J. Chem. Math. Phys. IJCMP*, vol. 3, no. 3, pp. 56–66, 2019, doi: 10.22161/ijcmp.3.3.2.

12. P. J. Jones *et al.*, "FilterK : a new outlier detection method for k-means clustering of physical activity," *J. Biomed. Inform.* p. 103397, 2020, doi: 10.1016/j.jbi.2020.103397.

13. T. I. A. Souza, A. L. L. Aquino, and D. G. Gomes, "analysis A Method to Detect Data Outliers from Smart Urban Spaces via Multiway Analysis," *Future Gener. Comput. Syst.*, 2018, doi: 10.1016/j.future.2018.09.062.

14. J. Wang, P. Du, Y. Hao, X. Ma, T. Niu, and W. Yang, "An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting," *J. Environ. Manage.*, vol. 255, no. July 2019, p. 109855, 2020, doi: 10.1016/j.jenvman.2019.109855.

15. A. Borah and B. Nath, "Jo urnal Pre-Proof," *Appl. Soft Comput. J.*, pp. 1–51, 2019, doi: 10.1016/j.asoc.2019.105824.

16. H. Ghallab, H. Fahmy, and M. Nasr, "Detection outliers on internet of things using big data technology," *Egypt. Inform. J.*, no. xxxx, pp. 1–8, 2019, doi: 10.1016/j.eij.2019.12.001.

17. A. Hudaib, M. Khanafseh, and O. Surakhi, "An Improved Version of K-medoid Algorithm using CRO," *Mod. Appl. Sci.*, vol. 12, no. 2, pp. 116–127, 2018, doi: 10.5539/mas.v12n2p116.

18. I. Dokmani, R. Parhizkar, J. Ranieri, and M. Vetterli, "Essential Theory, Algorithms and Applications," pp. 1–17, 2015.

19. A. Kolbaşi and P. A. Ünsal, "A Comparison of the Outlier Detecting Methods : An Application on Turkish Foreign Trade Data," *J. Math. Stat. Sci.*, vol. 5, pp. 213–234, 2015.

20. C. Anuradha, P. S. R. C. Murty, and C. S. Kiran, "Detecting Outliers in High Dimensional Data Sets Using Z-Score Methodology," *Int. J. Innov. Technol. Explor. Eng. IJITEE*, vol. 9, no. 1, pp. 48–53, 2019, https://doi.org/10.35940/ijitee.a3910.119119