DOI: https://dx.doi.org/10.4314/gjpas.v27i1.7

GLOBAL JOURNAL OF PURE AND APPLIED SCIENCES VOL. 27, 2021: 55-60
COPYRIGHT© BACHUDO SCIENCE CO. LTD PRINTED IN NIGERIA ISSN 1118-0579
www.globaljournalseries.com, Email: info@globaljournalseries.com

COMPARISON OF OUTLIER DETECTION TECHNIQUES IN NON-STATIONARY TIME SERIES DATA

SAMPSON TWUMASI-ANKRAH, SIMON KOJO APPIAH, DORIS ARTHUR, WILHEMINA ADOMA PELS, JONATHAN KWAKU AFRIYIE, DANIELSON NARTEY

(Received 28 September 2020; Revision Accepted 4 December 2020)

ABSTRACT

This study examined the performance of six outlier detection techniques using a non-stationary time series dataset. Two key issues were of interest. Scenario one was the method that could correctly detect the number of outliers introduced into the dataset whiles scenario two was to find the technique that would over detect the number of outliers introduced into the dataset, when a dataset contains only extreme maxima values, extreme minima values or both. Air passenger dataset was used with different outliers or extreme values ranging from 1 to 10 and 40. The six outlier detection techniques used in this study were Mahalanobis distance, depth-based, robust kernel-based outlier factor (RKOF), generalized dispersion, Kth nearest neighbors distance (KNND), and principal component (PC) methods. When detecting extreme maxima, the Mahalanobis and the principal component methods performed better in correctly detecting outliers in the dataset. Also, the Mahalanobis method could identify more outliers than the others, making it the "best" method for the extreme minima category. The kth nearest neighbor distance method was the "best" method for not over-detecting the number of outliers for extreme minima. However, the Mahalanobis distance and the principal component methods were the "best" performed methods for not over-detecting the number of outliers for the extreme maxima category. Therefore, the Mahalanobis outlier detection technique is recommended for detecting outlier in non-stationary time series data.

KEYWORDS: Outlier, time series, mahalanobis method, depth-based method, generalized dispersion method.

INTRODUCTION

There are two notable definitions of an outlier in literature. According to Barnett and Lewis (1994), an outlier is an observation that appears to deviate evidently from observations of the sample in which it occurs. Similarly, Johnson (1992) defines an outlier as an observation in a dataset that appears inconsistent with the rest of the observations in that dataset. The sources of outliers are mainly due to human error, instrument error, natural deviations in populations, fraudulent behavior, changes in systems' behavior, and/or faults in systems (Hodge and Austin, 2004).

Outlier detection refers to the task of identifying patterns in data that do not conform to expected behaviors (Ané et al., 2008; Angiulli and Pizzuti, 2002). Because an outlier can reveal unexpected but useful patterns in a dataset, it plays a crucial role in decision making, clustering, and pattern classification. Outlier detection is widely applied in public health anomaly, credit card fraud, intrusion detection studies, and has become of great interest to the data mining area (Barnett and Lewis, 1994; Fox, 1972; Glendinning, 1998). In literature, there are several outlier detection algorithms. Some popular categories of outlier detection techniques include z-score or extreme value

Sampson Twumasi-Ankrah, Department of Statistics and Actuarial Science, College of Science, Kwame Nkrumah University of Science and Technology, PMB, UPO, Kumasi, Ghana.
Simon Kojo Appiah, Department of Statistics and Actuarial Science, College of Science, Kwame Nkrumah University of Science and Technology, PMB, UPO, Kumasi, Ghana.
Doris Arthur, Department of Statistics and Actuarial Science, College of Science, Kwame Nkrumah University of Science and Technology, PMB, UPO, Kumasi, Ghana.
Wilhemina Adoma Pels, Department of Statistics and Actuarial Science, Collehe of Science, Kwame Nkrumah University of Science and Technology, PMB, UPO, Kumasi,Ghana.
Jonathan Kwaku Afriyie, Department of Statistics and Actuarial Science, College of Science, Kwame Nkrumah University of Science and Technology, PMB, UPO, Kumasi, Ghana.
Danielson Nartey, Department of Statistics and Actuarial Science, College of Science, Kwame Nkrumah University of Science and Technology, PMB, UPO, Kumasi, Ghana.

analysis, probabilistic and statistical modeling, linear regression models, proximity-based models, and information theory models. Graphically, the box plots and the scatter plots are also used to detect outliers in a given dataset. Several studies in literature compared some of these outlier detection methods. Notable but recent ones are as discussed in the following sequel:

Hodge and Austin (2004) surveyed the outlier detection methods that are used in machine learning and statistics, whiles Chandola et al. (2009) also reviewed the outlier detection techniques with respect to different assumptions. According to Xiaodan et al. (2018), other literature on outlier detection mainly focused on applications, such as network data (Gogoi et al., 2011) and temporal data (Gupta et al., 2014), or particular learning techniques, such as subspace learning and ensemble learning. The critical question is, which method can better detect outliers in a given time series dataset?

This study seeks to compare the performance of six outlier detection methods concerning their ability to correctly identify the exact number of outliers that are introduced in the dataset. The study is different from the literature reviewed in these ways: (1) several numbers (or sample size) of outliers are introduced to the dataset; and (2) two dimensions of outliers that are extreme minima and extreme maxima are considered in the dataset.

METHODS AND MATERIALS

Data Source and Nature

The performance of six outlier detection methods was compared using the air passenger dataset, which spans from 01/1960 to 12/1971, consisting of 144 observations or data points, which exhibits both trend and seasonality patterns. The dataset was obtained from the Time Series Analysis (TSA) package in R software (Cryer and Chan, 2012).

The analysis was performed following the below steps of an algorithm for outlier detection:

Step 1: Check to see if the datasets contain any outlier using the classical box plot approach to create lower and upper fences. Hence any value below the extreme minima or above the extreme maxima fence is an outlier. The extreme minima and maxima values are given by $Q_1 - (1.5 \times IQR)$ and $Q_3 + (1.5 \times IQR)$, respectively, where Q_1 is the first quartile, Q_3 is the third quartile and IQR is the inter-quartile range of the dataset.

Step 2: Check the minimum value of the dataset and separately introduce, in each data, arbitrary extreme minima and maxima values (i.e., the number of outliers) $1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \; and \; 40$, in each case of the data, the sample size will increase depending on outliers introduced.

Step 3: Compare the performance of the six outlier detection methods to find which method can correctly detect all the outliers that were introduced into the dataset. In context, an outlier detection method is considered to be the "best" performing method if it identifies all or maximum number of outliers that were introduced in the dataset.

Step 4: Introduce both extreme maxima and minima into a particular dataset, which in this study is termed as the mixture dataset with sample sizes (i.e., the number of outliers) $2, 4, 6, 8, 10, 12, 14, 16, 18, 20, and \; 80.$ Thus, sample size 2 would contain one value of extreme minima and one value of extreme maxima.

Step 5: Compare the performance of the six outlier detection methods for this mixture dataset. This is to check the methods' performance when both extreme maxima and minima are present in the data.

Step 6: Repeat steps 1-5 using same dataset.

OUTLIER DETECTION METHODS

The outlier detection techniques considered in this study are Mahalanobis distance, depth-based, robust kernel-based outlier factor (RKOF), generalized dispersion, Kth nearest neighbors distance (KNND), and principal component (PC) methods. The Mahalanobis distance method is a well-known criterion which depends on estimated parameters of the multivariate distribution. Given *n* observations from a *p*-dimensional dataset X, we define the sample covariance matrix by

$$V_n = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(x_i - \bar{x}_n)^T \qquad (1)$$

where $\bar{x}_n$ denotes the sample mean vector. All observations with a large V$_n$ values are indicated as outliers.

The depth-based method is defined as

$$0.5(1 + \binom{n}{p}^{-1} \sum (Volume\left(S\left(u, x[i_1,], \ldots, x\left[i_p,\right]\right)\right))^{-1} \qquad (2)$$

where $S(.)$ denotes the simplex generated by args, and the sum and average are taken over all p-pelts $x[i_1,], \ldots, x\left[i_p,\right]$ such that $1 \le i_1 < \cdots < i_p \le n.$

For the robust kernel-based outlier factor (RKOF), the local kernel density estimate of *p* is defined by:

$$kde(p) = \frac{\sum_{0 \in N_k(p)} \left\{h^{-\gamma} {\lambda_0}^{-\gamma} K\left(h^{-1} {\lambda_0}^{-1} (p-0)\right)\right\}}{|N_k(p)|} \qquad (3)$$

where *h* is the smoothing parameter, γ is the sensitivity parameter, $K(x)$ is the multivariate kernel function, $\lambda_0 = \{f(0)/g\}^{-\alpha}$ is the local bandwidth factor, $f(x)$ is a pilot density estimate that satisfies $f(x) > 0$ for all the objects, α is the sensitivity parameter that satisfies $0 \le \alpha \le 1$, and g is the geometric mean of $f(x)$.

The generalized dispersion method computes Leave-One-Out (LOO) dispersion matrix for each observation (without considering the current observation) and, based on the difference between determinant of LOO dispersion matrix and determinant of actual dispersion matrix, labels an observation as an outlier.

The principal component outlier statistic is defined, and the extremity of observation concerning a particular group is evaluated with this statistic:

$$ {D_{W(k)}}^2(x) = \frac{{D_{1(k)}}^2(x)/(p-k)}{\sum_{i=1}^{n} {D_{1(k)}}^2(x_i)/((p-k)(n-k-1))} \qquad (4)$$

to assess a new observer x. The numerator in this expression (equation (4)) is the variance of the observation x from the principal component model, and the denominator as the variance of the deviations. This statistic is compared to some critical values of the *F*-distribution

.

The K^{th} nearest neighbors distance (KNND) method uses the distance-based method in finding outliers in a dataset, thus using the k nearest neighborhood method. For a set of each point in the KNND, the local outlier factor (LOF) uses the local reachability density (LRD) and compares it with those of the neighbors of each participant of that KNND set. The LRD (a density estimate that reduces the variables) of an object *p* is defined as:

$$LRD(p) = 1 / \frac{\sum_{o \in KNN(p)} reach - dist_k(p \leftarrow o)}{|KNN(p)|} \qquad (5)$$

where$reach - dist_k(p \leftarrow o) = \max\{k - dist(o), d(p, o)\}$. The final local outlier factor score is given as:

$$LOF_{k(p)} = \frac{1}{|KNN(p)|} \sum_{o \in KNN(p)} \frac{lrd_{k(o)}}{lrd_{k(p)}} \qquad (6)$$

where $lrd_{k(p)}$ and $lrd_{k(o)}$ are the local reachability density of *p* and o respectively.

RESULTS AND DISCUSSION

In this study, the two issues of interest are the correct detection of the number of outliers introduced into a non-stationary time series dataset and over detection of the number of outliers introduced into the dataset. Therefore, an outlier detection method is considered the "best" performing method if it identifies all or the maximum number of outliers introduced in the dataset.

In Table 1, the descriptive summary of the data set is presented. Using the classical box plot approach, it was evident that the air passenger dataset has no outlier. Therefore, artificial outliers would be introduced into the dataset.

Table 1: Descriptive summary of the Air Passenger time series data sets

Statistic	Air Passenger
Minimum Statistic	104
Maximum Statistic	622
Mean	280.3
Median	265.5
1st Quartile	180
3rd Quartile	360.5
Number of Outliers	0*

* zero number of outlier means there is no outlier

Correct Detection of Number of Outliers

Artificial outliers of several sizes were introduced into the air passenger dataset; therefore, the technique that could identify them was considered the "best" outlier detection method. In all, ninety-five (95) outliers were introduced in the extreme minima and maxima categories and one-hundred and ninety (190) outliers for the mixture dataset (containing both minima and maxima). Therefore, the method with the maximum number of detections is considered the "best" method at each extreme category.

In detecting the appropriate method for extreme minima with varying outliers, the Mahalanobis method could identify 36 out of 95 outliers in the air passenger dataset (see Table 2). However, the worst performed method was the principal component method that could not detect any outliers. For the extreme maxima category, the generalized dispersion method was the worst in detecting the outliers with only 3 out of 95 outliers, whiles the Mahalanobis distance and principal component methods were the "best" in correctly detecting the number of outliers.

Table 2: Comparison of various outlier techniques correctly detecting the number of outliers introduced into the air passenger dataset

Number of Outliers	Depth based		Generalised Dispersion		KNND		Mahalanobis		RKOF		PC	
	min	max	Min	max	Min	max	Min	max	Min	max	Min	Max
1	**1**	**1**	0	0	0	**1**	**1**	**1**	**1**	**1**	0	**1**
2	**2**	1	0	0	**2**	1	**2**	1	**2**	1	0	1
3	2	1	0	0	0	1	**3**	1	**3**	1	0	1
4	1	2	0	0	0	**4**	**4**	**4**	**4**	**4**	0	**4**
5	2	2	0	0	1	4	**5**	4	**5**	4	0	4
6	2	2	0	0	1	4	**6**	4	0	4	0	4
7	2	2	1	0	0	4	**7**	4	0	4	0	4
8	2	2	1	0	0	4	6	4	0	4	0	4
9	2	2	1	0	0	4	2	4	0	4	0	4
10	2	2	1	0	0	4	0	4	0	4	0	4
40	3	3	0	3	0	4	0	16	0	4	0	16
Sum	**21**	**20**	**4**	**3**	**4**	**35**	**36**	**47**	**15**	**35**	**0**	**47**

Correct detection is shown in boldface

In the mixture dataset (containing both minima and maxima) category, it was evident in Table 3 that for the extreme minima, the generalized dispersion, Mahalanobis, and principal component methods could not detect any of the outliers introduced into the dataset. However, the depth-based method was "best" in detecting the number of outliers introduced into the dataset. For extreme maxima, the principal component and Mahalanobis methods were the "best" in detecting the number of outliers introduced into the dataset. The generalized dispersion could not detect any outlier in the dataset.

Table 3: Comparison of various outlier techniques correctly detecting the number of outliers for the mixture dataset

No. of Outliers	Depth-based		Generalised Dispersion		KNN		Mahalanobis		Robust Kernel		P C	
	min	max	Min	max	Min	max	min	Max	min	Max	min	max
2	**1**	**1**	0	0	**1**	**1**	0	1	**1**	**1**	0	1
4	**2**	**2**	0	0	**2**	**2**	0	2	**2**	**2**	0	2
6	2	2	0	0	0	6	0	3	**3**	**3**	0	3
8	2	2	0	0	0	4	0	4	0	4	0	4
10	2	2	0	0	0	5	0	4	0	5	0	4
12	2	2	0	0	0	5	0	5	0	5	0	5
14	2	2	0	0	0	0	0	6	0	5	0	6
16	2	2	0	0	0	0	0	6	0	5	0	6
18	2	3	0	1	0	0	0	7	0	4	0	7
20	2	3	0	0	0	0	0	7	0	4	0	7
80	3	3	0	3	0	5	0	19	0	6	0	19
Sum	**22**	**24**	**0**	**4**	**3**	**28**	**0**	**64**	**6**	**44**	**0**	**64**

(Correct detection is shown in boldface)

Over Detection of Number of Outliers

The performance of the six methods for over detecting outliers is assessed in the dataset. The "best" detection technique is the technique that records the minimum value.

From Table 4, it was evident that the kth nearest neighbor distance method was the "best" method for not over-detecting the number of outliers for the extreme minima. However, the principal component method was worst in performance since it recorded the highest number of outliers for over detection when having extreme minima. In introducing extreme maxima, the kth nearest neighbor distance method was the "worst" performing method since it had the highest number of over-detection of outliers. The Mahalanobis distance and the principal component methods were the "best" performing method with only three over-detections.

Table 4: Comparison of various outlier techniques for over detection of the number of outliers introduced into the Air Passenger dataset

Number of Outliers	Depth based		Generalised Dispersion		KNND		Mahalanobis		RKOF		PC	
	min	max	min	max	Min	Max	min	max	min	max	min	Max
1	3	3	4	4	**0**	5	5	**0**	3	2	6	**0**
2	**2**	3	3	5	2	3	5	1	2	4	7	1
3	2	**0**	3	**0**	**0**	4	5	2	1	3	8	2
4	1	**0**	3	**0**	**0**	1	5	**0**	**0**	1	9	**0**
5	2	**0**	**0**	**0**	1	**0**	5	**0**	**0**	1	10	**0**
6	2	**0**	1	**0**	1	**0**	5	**0**	**0**	**0**	11	**0**
7	2	**0**	5	**0**	2	**0**	5	**0**	**0**	1	12	**0**
8	2	**0**	5	**0**	2	**0**	5	**0**	5	**0**	11	**0**
9	2	**0**	5	**0**	2	2	5	**0**	6	1	7	**0**
10	2	**0**	5	**0**	2	1	5	**0**	5	1	5	**0**
40	3	3	4	6	5	**0**	5	**0**	6	**0**	5	**0**
Sum	**23**	**9**	**38**	**15**	**17**	**16**	**55**	**3**	**28**	**14**	**91**	**3**

(Correct detection is shown in boldface)

From Table 5, the principal component and Mahalanobis methods were the "best" methods since they could not over-detect any outlier. In contrast, the generalized dispersion method was the worst in performance since it recorded the highest number for over detection regarding the mixture scenario.

Table 5: Comparison of various outlier techniques for over detection of the number of outliers of the mixture dataset

Number of Outliers	Depth based	Generalised Dispersion	KNND	Mahalanobis	RobustKernel	PC
1	2	4	2	0	3	0
2	0	4	2	0	0	0
3	0	2	1	0	0	0
4	0	3	1	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	5	0	0	0	0
8	0	7	0	0	0	0
9	0	4	0	0	1	0
10	0	0	0	0	3	0
40	0	0	0	0	1	0
Sum	**2**	**29**	**6**	**0**	**8**	**0**

(Correct detection is shown in boldface; thus, no over detection) [Indicated the bold figures]

CONCLUSION

The performance of six different methods for detecting outliers was compared using the air passenger dataset. The air passenger dataset did not have any outlier; therefore, artificial outliers were introduced into the dataset. The performance was evaluated by the highest number of outliers that a detection method could correctly specify. For the extreme minima category, the “best” performed outlier detection technique was the Mahalanobis method, whiles the worst performed method was the principal component method. Again, for the extreme maxima category, the generalized dispersion method was the worst performed detection technique, whiles the Mahalanobis distance and principal component methods were the "best" in correctly detecting the number of outliers.

Also, for the mixture dataset (containing both minima and maxima) category, the Mahalanobis and principal component methods were the “best” performed methods in correctly detecting outliers.

Lastly, the kth nearest neighbor distance method was the "best" method for not over-detecting the number of outliers for extreme minima. However, the Mahalanobis distance and the principal component methods were the "best" performed methods for not over-detecting the number of outliers for the extreme maxima category.

Therefore the Mahalanobis outlier detection technique is recommended for detecting outlier in a non-stationary time series dataset.

REFERENCES

Barnett V and Lewis T., 1994. Outliers in statistical data. 3rd Edition. John Wiley and Sons, Chichester, p. 584.

Johnson R., 1992, Applied Multivariate Statistical Analysis. Prentice-Hall, 3rd Edition

Hodge V J., and Austin J., 2004. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review,**22** (2) 85–126.

Ané T, Ureche-Rangau L, Gambet J. B, Bouverot J., 2008. Robust outlier detection for Asia- Pacific stock index returns. Journal of International Finance, Market Inst. Money, 18: 326-34

Angiulli, F and Pizzuti C., 2002. Fast outlier detection in high dimensional spaces. In European conference on principles of data mining and knowledge discovery pp. 15-27.

Fox A. J., 1972. Outliers in time series. Journal of the Royal Statistical Society: Series B (Methodological), 34(3), 350-363.

Glendinning R. H., 1998. Determining the order of an ARMA model from outlier contaminated data. Communication in Statistics- Theory and Methods, 27: 13-40.

Chandola VA, Arindam B, and Kumar V., 2009. Anomaly detection: A survey. ACM Computing Surveys, 41 (3) 1–58.

Xiaodan X, Huawen L, Li L and Minghai Y., 2018. A Comparison of Outlier Detection Techniques for High-Dimensional Data. International Journal of Computational Intelligence Systems, 11 (652–662)

Gogoi P, Bhattacharyya DK, and Borah B., 2011. A Survey of Outlier Detection Methods in Network Anomaly Identification. Computer Journal, **54** (4) 570–588.

Gupta M, Gao J, Aggarwal C and Han J., 2014. Outlier Detection for Temporal Data: A Survey. IEEE Transactions on Knowledge and Data Engineering, **26** (9) 2250–2267.

Jonathan Cryer and Kung-Sik Chan, 2012. TSA contains R functions and datasets detailed in the book ``Time Series Analysis with Applications in R (second edition)'. https://cran.microsoft.com/snapshot/2018-04-14/web/packages/TSA/TSA.pdf