2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI) | 978-1-6654-7997-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWECAI55315.2022.00048

Comparative Analysis of three outlier detection methods in univariate data sets

Xunqiang Gong[1,2], Fangze Zhang[1], Tieding Lu[1,2*] and Wei You[3]
[1]Faculty of Geomatics, East China University of Technology, Nanchang, PR China
[2]Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake, Ministry of Natural Resources, Nanchang, PR China
[3]Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, PR China
*Corresponding author's e-mail: tdlu@ecut.edu.cn

Abstract: **In the field of geomatics engineering and signal processing, observational data can bring a lot of information. However, the abnormal data (i.e. outliers) may be acquired due to human carelessness or limiting condition, causing interference to subsequent research. To solve this problem, outlier detection methods have been proposed to detect and remove outliers. Three common outlier detection methods (i.e. Z-score method, boxplot method and median absolute deviation method) in signal processing are introduced in this paper. A comparison of the three methods is conducted through experimental evaluation with two sets of experiments. The results of experiments show that the number of outliers detected with median absolute deviation method significantly outperform those of Z-score and boxplot methods. It shows that the median absolute deviation method can more effectively detect and remove outliers so that the more reliable and accuracy results can be obtained.**

Keywords: signal processing, outlier detection, median absolute deviation method, Z-score method, boxplot method

I. Introduction

Outlier detection of observation data is an important research hotspot in geomatics engineering and signal processing community because observation results provide fundamental sources for many engineering and socioeconomic applications. However, the survey data may contain outliers, which affect the accuracy of the research results, and make subsequent engineering and social applications cause deviations in actual use. For this reason, several approaches for dealing with outliers have been developed in the recent decades in order to take into consideration the possible outliers in observations. A common strategy is to design complex models that are naturally robust to the outliers. In this direction, robust estimation methods are developed in several ways to make them robust to outliers [1-3]. Although good results have been reported in many researches in the framework of robust estimation, most of the existing approaches can be only effective when data sets are contaminated by small proportion of outliers.

Another method to deal with outliers is to build a model for detecting outliers through probability and statistics in machine learning or deep learning, and then estimate the parameters based on refined observations [4-5]. In particular, the Z-score method, boxplot method and median absolute deviation method were introduced frequently to detect outliers of observations in univariate data sets [6-8]. However, the three mentioned above methods have not been compared and analyzed systematically so far. In order to assess the performance of the three outlier detection methods in this paper, a comparison is conducted through experimental evaluation with simulation data sets (with different numbers and magnitudes of outliers) and two sets of real-life data.

This paper is organized as follows: In Section 2, three methods for detecting outliers is reviewed. In Section 3, Experimental design with simulation and real-life data set are described. The description of the results and analyses are given in Section 4. Finally, some conclusions are made in Section 5.

II. Basic outlier detection methods in univariate data sets

In order to compare the performance of outlier detection method in univariate data sets, it would be pertinent to first conduct three basic outlier detection methods.

A. The Z-score method

If $\{x_1, x_2, \cdots, x_n\}$ is a set of observations, we will denote its sample mean by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

where $\bar{x}$ is the mean of the series. More generally, the n univariate observations x_i are independent and identically distributed with the Gaussian distribution.

The standard deviation is a measure that summarizes the amount by which every value within a dataset varies from the mean. Effectively it indicates how tightly the values in the dataset are bunched around the mean value. The standard deviation of a sample is known as S and is calculated using

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} \tag{2}$$

where $n-1$ is the number of degrees of freedom. The sample mean and the standard deviation are simple and easy to compute. One method that can be used to detect outliers is the Z-score method [6,9-10] using the mean and standard deviation. The decision criterion becomes

$$Z_i = \left|\frac{x_i - \bar{x}}{S}\right| > 2.5 \tag{3}$$

The basic idea of this method is that x_i is outlier if Z_i exceeds 2.5 [11].

B. The boxplot method

One simple way commonly employed to identify outliers is based on the concept of the boxplot and involves the use of

978-1-6654-7997-4/22/$31.00 ©2022 IEEE
DOI 10.1109/IWECAI55315.2022.00048

fences [12-13]. This method has come into common usage. The interquartile range (IQR) is defined as the difference between the third and first quartiles, that is

$$IQR = Q_3 - Q_1 \tag{4}$$

where the third quartile Q_3 is the value in the data set that holds 25% of the values above it. The first quartile Q_1 is the value in the data set that holds 25% of the values below it. In this plot, a box is drawn from the first quartile Q_1 to the third quartile Q_3 of the data. Specifically, the fences F_1 and F_3, are usually defined as

$$F_1 = Q_1 - 1.5\ \ IQR,\ F_3 = Q_3 + 1.5\ \ IQR \tag{5}$$

Observations outside the interval $[F_1, F_3]$ are traditionally marked as outliers. As is well-known, if the data come from the normal distribution, the former interval contains 99.3% of the values [7].

C. The median absolute deviation method

If $\{x_1, x_2, \cdots, x_n\}$ is a set of observations, we will denote its sample median by

$$\underset{i=1,\cdots,n}{median}(x_i) \tag{6}$$

which is simple the middle ordered observation when n is odd. When n is even, the median is the average of the ordered observations with ranks $\left({}^{n}/_{2}\right)$ and $\left({}^{n}/_{2}+1\right)$. The median is a measure of central tendency and insensitive to the presence of outliers.

The median absolute deviation (MAD) was first promoted by [1]. Its breakdown value is also about 50% and its influence function is bounded. The MAD is defined as follows [8,14]

$$MAD = b \underset{i=1,\cdots,n}{median} \left| x_i - \underset{j=1,\cdots,n}{median}(x_j) \right| \tag{7}$$

where b is a constant ($b = 1.4826$, usually). The sample median and the MAD are simple and easy to compute. Then, we must define the rejection criterion of a value. By default, we suggest a threshold of 2.5 as a reasonable choice [15]. The decision criterion becomes

$$C = \frac{\left| x_i - \underset{j=1,\cdots,n}{median}(x_j) \right|}{MAD} > 2.5 \tag{8}$$

III. Experimental Design

In order to evaluate the three outlier detection methods in univariate data sets, two sets of simulation data and two sets of real-life data are used in this evaluation.

A. Design of Simulation Experiments

In the simulation experiments, 100 sets of simulated univariate observation data with a true value of 10 are used in each experiment. Firstly, random errors are added to observations, and these random errors conform to the normal distribution $N(0, \sigma^2 I)$, where $\sigma = 0.1$ and I is the identity matrix. The simulated univariate observation data with random errors is shown in Fig 1.

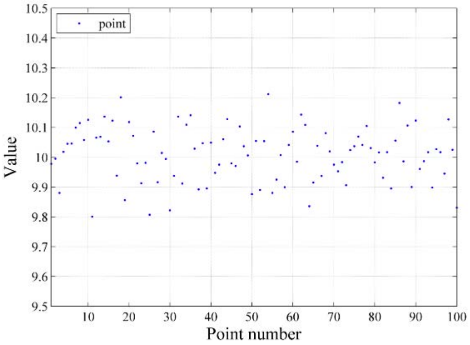

Fig 1. Simulated univariate observation data with random errors

There are two sets of experiments to compare the capabilities of detecting outliers of these three methods. In the first set of experiment, the magnitude of the outliers is fixed but the number of outliers is varied systematically from 0 to 50. The number of the outliers is fixed but the magnitude of outliers is systematically varied from 3σ to 13σ in the second set of experiment. In order to improve the reliability of numbers of outliers detected, the average value of 1000 repeated simulations is taken as the final value in each experiment. It should be noted that the results obtained with different magnitudes of outliers in the first set of experiment are very consistent, therefore the results of magnitude of outliers is 8σ (median of 3σ to 13σ) are presented respectively in this paper. In the second set of experiment, the results of number of outliers is 25 (i.e. 25%) are represented for a similar reason.

B. Design of Real-life Experiments

The first set of real-life data is based on the Nigerian's inflation rate from [15]. The data are over a thirty three years period ranging from 1981 to 2013 listed in Table 1, which was obtained from the Central Bank of Nigeria yearly statistical bulletin.

The second set of real-life data was used for the experiment, which is adopted from [11]. The data including five observations are shown in Table 2. All the three methods tested in the simulation experiments are also evaluated.

Table 1. Nigerian's inflation rate from [15]

Year	Inflation rate (%)	Year	Inflation rate (%)	Year	Inflation rate (%)	Year	Inflation rate (%)	Year	Inflation rate (%)
1981	20.9	1988	38.3	1995	72.8	2002	12.9	2009	13.9
1982	7.7	1989	40.9	1996	29.3	2003	14.0	2010	11.8
1983	23.2	1990	7.5	1997	8.5	2004	15.0	2011	10.3
1984	39.6	1991	13.0	1998	10.0	2005	17.9	2012	12.0
1985	5.5	1992	44.5	1999	6.6	2006	8.5	2013	8.0
1986	5.4	1993	57.2	2000	6.9	2007	5.4		
1987	10.2	1994	57.0	2001	18.9	2008	15.1		

TABLE 2. OBSERVATIONS FROM [11]

Observation number	1	2	3	4	5
Observation	6.27	6.34	6.25	63.1	6.28

IV. EXPERIMENTAL RESULTS AND ANALYSES

A. *Results and analyses of simulation experiments*

In this section, experimental analyses were conducted and the results by the three methods, i.e. Z-score method, boxplot method and MAD method, are compared with the number of outliers detected.

1) *Effect of the numbers of outliers in simulation experiments*

The numbers of outliers detected by using Z-score, boxplot and MAD methods are plotted in Fig 2. The number of outliers varies from 0 to 50. It is clear that when the number of outliers is 0, the numbers of outliers detected of these three methods are approximately 0 (i.e. their true value), but the results are slightly greater than the true value. It is because a few simulated biggish random errors are handled as outliers. When the numbers of outliers increase from 0 to 10, the numbers of outliers detected by using the three methods increase approximatively and linearly. It indicates that all three methods have a good performance for detecting outliers. As the numbers of outliers range from 10 to 35, the boxplot method and the MAD method have the values increased to a level of 34.212 and 34.967 respectively. However, the numbers of outliers detected with the Z-score method has a decrease from 8.733 to 0.002. This indicates that the Z-score method is not able to detect outliers properly when the number of outliers ranges from 10 to 35. Furthermore, the MAD method has a much higher ability than the other two methods to detect outliers in univariate data sets for improving the accuracies of observations when the numbers of outliers vary from 35 to 45. The numbers of outliers detected by using the MAD method have an increase from 1.553 to 40.166 when the numbers of outliers increase from 0 to 45, but the numbers of outliers detected by using the Z-score and boxplot methods do not increase continuously with the number of outliers. The results indicate that (i) the Z-score method for detecting outliers is based on the characteristics of a normal distribution and uses the mean and standard deviation as the measures, which are very sensitive to the presence of outliers [11]. Thus the Z-score method has a very poor performance when the observations contain greater than 10% outliers; (ii) The boxplot method does not use the extreme potential outliers in computing the measure of identifying outliers so that are not distorted by a few extreme values. But the boxplot method is still not suitable when the observations contain greater than 35% outliers; (iii) the median is a measure of central tendency and insensitive to the presence of outliers so that the MAD method is an effective method to detect outliers when the observations contain outliers. It should be noted that it is difficult to detect outliers by using MAD method when the number of outliers reaches 50 (i.e. 50%) because its breakdown value is about 50%.

2) *Effect of the magnitudes of outliers in simulation experiments*

The numbers of outliers detected at different outlier levels for all the three methods are shown in Fig 3. The magnitudes of outliers range from 3σ to 13σ. It is clear that (i) the number of outliers detected by using the Z-score method decreases slowly, i.e. from 1.523 to 0.031; (ii) the values of the boxplot and MAD methods follow a logarithmic growth trend from 3σ to 7σ, and the values are rather stable from 7σ to 13σ; (iii) the values of boxplot and MAD methods are much larger than that of the Z-score method, and the value of MAD method is slightly larger than that of the boxplot method. The results confirm the fact that when the numbers of outliers are more than 10 (i.e. 10%), the Z-score method is invalid. By contrast, the MAD method can detect more outliers when the magnitudes of outliers are smaller than 7σ. It must be emphasized that, with an increase in the magnitudes of the outliers (3σ to 7σ), the difference between the number of outliers detected for MAD method and that for boxplot method decreases slowly. It indicates that the boxplot method cannot detect outliers more effectively than the MAD method when the magnitude of outliers is small. On the other hand, the results indicate that the MAD method has much better performance over boxplot method, and superiority becomes more and more obvious with a decrease in the magnitude of outliers. It is interesting to see that the values for boxplot and MAD methods are smaller than its true value (i.e. 25) when the magnitudes of outliers are smaller than 7σ. This is because a few generated outliers are handled as biggish random errors. It is clear that when the magnitudes of outliers are bigger than 7σ, these two methods are able to obtain a value approach to its true value (i.e. 25).

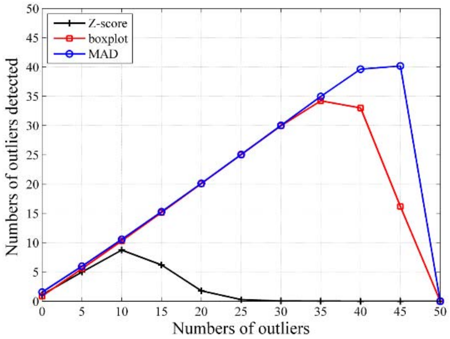

Fig 2. Numbers of outliers detected as an increase in the number of outliers in simulation experiments

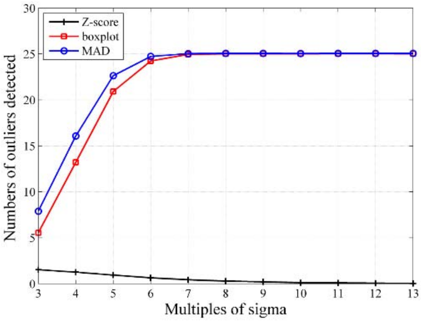

Fig 3. Magnitudes of outliers detected as an increase in the magnitude of outliers in simulation experiments

B. *Results and analyses of the real-life experiments*

1) *Results and analyses of the first set of real-life experiment*

According to the data in Table 1, the Z-scores of the observations by using Z-score method can be obtained directly from formulas (1)-(3). And the decision values by using MAD method could also be calculated from formulas (6)-(8). The results are shown in Table 3. From the results of Table 3, it can be found that z-score 3.011 by Z-score method is greater than 2.5, meaning that inflation rate 72.8% corresponding to year 1995 is considered as an outlier. However, the decision values 3.262, 3.1.3, 3.422, 3.863, 5.420, 5.396 and 7.334 by MAD method are greater than 2.5. This indicates that the inflation rates 39.6%, 38.3%, 40.9%, 44.5%, 57.2%, 57.0% and 72.8% corresponding to the years 1984, 1988, 1989, 1992, 1993, 1994 and 1995 are marked as outliers. At the same time, the interval by boxplot method is $[-18.750\%, 53.250\%]$ according to formulas (4) and (5). This implies that inflation rates 57.2%, 57.0% and 72.8% corresponding to the different years 1993, 1994, and 1995 are possible outliers from the data. The above results indicate that Z-score method and boxplot method can only identify one and three outliers respectively. But seven outliers (i.e. inflation rates of years 1984, 1988, 1989, 1992, 1993, 1994 and 1995) can be detected by using MAD method. Therefore, MAD method has a better capacity for detecting outliers.

2) *Results and analyses of the second set of real-life experiment*

The calculation for the second set of real-life data is similar to that for the first set of real-life data. The z-scores for Z-score method and decision values for MAD method are listed in Table 4. From the results of Table 4, it can be found that all z-scores of the observations for Z-score method are less than 2.5 meaning that the observations do not contain outliers. The results show that the efficiency of Z-score method is quite constrained for small data sets. Therefore, this method is very unsuitable to detect outliers in small data sets. As illustrated in Table 6, the decision value for MAD method of observation number 4 is greater than 2.5. This indicates that it is an outlier. The interval $[-36.430, 77.410]$ for boxplot method can be obtained from formulas (4) and (5). All observations are inside the interval, meaning that the observations do not contain outliers. From the above analysis, we can find that only MAD method can identify the obvious outlier. Therefore, MAD method is a better choice for outlier detection when the observations of small data sets contain outliers.

TABLE 3. Z-SCORES FOR Z-SCORE METHOD AND DECISION VALUES FOR MAD METHOD IN THE FIRST SET OF REAL-LIFE EXPERIMENTS

Year	Inflation rate (%)	Z-score	Decision value	Year	Inflation rate (%)	Z-score	Decision value
1981	20.9	0.036	0.969	1998	10.0	0.588	0.368
1982	7.7	0.720	0.650	1999	6.6	0.783	0.785
1983	23.2	0.168	1.251	2000	6.9	0.766	0.748
1984	39.6	1.108	3.262	2001	18.9	0.078	0.724
1985	5.5	0.846	0.920	2002	12.9	0.422	0.012
1986	5.4	0.852	0.932	2003	14.0	0.359	0.123
1987	10.2	0.577	0.343	2004	15.0	0.302	0.245
1988	38.3	1.034	3.103	2005	17.9	0.135	0.601
1989	40.9	1.183	3.422	2006	8.5	0.674	0.552
1990	7.5	0.731	0.675	2007	5.4	0.852	0.932
1991	13.0	0.416	0.000	2008	15.1	0.296	0.258
1992	44.5	1.389	3.863	2009	13.9	0.365	0.110
1993	57.2	2.117	5.420	2010	11.8	0.485	0.147
1994	57.0	2.105	5.396	2011	10.3	0.571	0.331
1995	72.8	3.011	7.334	2012	12.0	0.474	0.123
1996	29.3	0.518	1.999	2013	8.0	0.701	0.613
1997	8.5	0.674	0.552				

TABLE 4. Z-SCORES FOR Z-SCORE METHOD AND DECISION VALUES FOR MAD METHOD IN THE SECOND SET OF REAL-LIFE EXPERIMENTS

Observation number	1	2	3	4	5
Observation	6.27	6.34	6.25	63.1	6.28
Z-score	0.448	0.445	0.449	1.789	0.447
Decision value	0.225	1.349	0.674	1277.485	0.000

V. CONCLUSIONS

Outlier has been recognized as an outlying observation which can result to false interpretation of the data set for statistical inference. In order to detect the possible outliers in observations, three common outlier detection methods are introduced. Comparative experimental evaluation has been carried to assess the performance of the three outlier detection methods. Both simulation data sets and real-life data sets have been used for experiments. In the simulation data sets, different numbers and different magnitudes of outliers were designed. Consequently, the effects of the numbers and magnitudes of outliers were investigated. The results show that MAD method has much better performance over Z-score and boxplot methods. In real-life data experiment, the MAD method was able to better detect the probable outliers because the number of outliers detected is much more than those of Z-score and boxplot methods. From these experimental results, it is therefore concluded that the MAD method can effectively detect the outliers compared with Z-score and boxplot methods so that the more pure observations can be obtained.

ACKNOWLEDGMENTS

This work was jointly supported by the National Natural Science Foundation of China (Grant Nos. 42101457, 42061077 and 41974013).

REFERENCES

[1] Hampel F.R. (1974) The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69(346): 383-393.

[2] Gong X., Li Z. (2017) A robust weighted total least-squares solution with Lagrange multipliers. Survey Review, 49(354): 176-185.

[3] Agostinelli C., Bianco A.M., Boente G. (2020) Robust estimation in single-index models when the errors have a unimodal density with unknown nuisance parameter. Annals of the Institute of Statistical Mathematics, 72(3): 855-893.

[4] Rahul K., Banyal R. (2021) Detection and correction of abnormal data with optimized dirty data: a new data cleaning model. International Journal of Information Technology and Decision Making, 20(02): 809-841.

[5] Fouzi H., Abdelkader D., Sun Y. (2018) Detecting abnormal ozone measurements with a deep learning-based strategy. IEEE Sensors Journal, 18(17): 7222-7232.

[6] Zhang Z., Cheng Y., Liu N. (2014) Comparison of the effect of mean-based method and Z-score for field normalization of citations at the level of Web of Science subject categories. Scientometrics, 101(3): 1679-1693.

[7] Li A., Feng M., Li Y., Liu Z. (2016) Application of outlier mining in insider identification based on boxplot method. Procedia Computer Science, 91: 245-251.

[8] Gong X., Shen L., Lu T. (2019) Refining training samples using median absolute deviation for supervised classification of remote sensing images. Journal of the Indian Society of Remote Sensing, 47(4): 647-659.

[9] Schaafsma A. (2018) A new method for correcting middle cerebral artery flow velocity for age by calculating Z-scores. Journal of Neuroscience Methods, 307: 1-7.

[10] Li X., Tripe D., Malone C., Smith D. (2020) Measuring systemic risk contribution: the leave-one-out z-score method. Finance Research Letters, (36): 101316.

[11] Rousseeuw P.J., Hubert M. (2011) Robust statistics for outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1): 73-79.

[12] Schwertman N.C., Owens M.A., Adnan R. (2004) A simple more general boxplot method for identifying outliers. Computational statistics and data analysis, 47(1): 165-174.

[13] Zhou X., Gu G. (2021) An algorithm of generating random number by wavelet denoising method and its application. Computational Statistics, (4): 1-18.

[14] Dickey J., Junek W.N., Borghetti B.J. (2020) Baznet: a deep neural network for confident three-component backazimuth prediction. Pure and Applied Geophysics, 178(3): 2459–2473.

[15] Nkechinyere E.M., Iheagwara A.I., Okenwe I. (2015) Comparison of different methods of outlier detection in univariate time series data. International Journal for Research in Mathematics and Mathematical Sciences, 1(2): 22-50.