CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 9(2) (2024), Pages 341-351
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparing Outlier Detection Methods: An Application on Indonesian Air Quality Data

Anwar Fitrianto[1,2*], Amalia Kholifatunnisa[1], Anang Kurnia[1]

[1]Department of Statistics, IPB University, Bogor, West Java, Indonesia
[2]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia

Email: anwarstat@gmail.com

ABSTRACT

There are many methods for detecting outliers, but only a few methods consider data distribution. This research compares the outlier detection method on univariate data with a skewed distribution. Outlier detection methods used in this research are the famous Tukey's boxplot which is good symmetrically distrusted data, adjusted boxplot, sequential fences, and adjusted sequential fences. The methods are implemented to identify areas of concern due to poor air quality during the Implementation of Micro-Community Activity Restrictions. The study used Indonesian air quality index data. Based on the study, the adjusted boxplot method performs best based on the number of outliers detected, error rate, accuracy, precision, specificity, sensitivity, and robustness. Adjusted boxplot and adjusted sequential fences can detect tails that contain outliers accurately because the skewness coefficient makes them more robust. Meanwhile, Tukey's boxplot and sequential fences are poor methods since they couldn't detect true outliers correctly. Based on the results, adjusted boxplot is the best method. Then, areas that need attention due to poor air quality include South Sumatera, South Sulawesi, West Java, Riau, North Sumatera, Jambi, Jakarta, and East Java.

Keywords: adjusted boxplot, adjusted sequential fences, outlier, sequential fences, Tukey's boxplot

INTRODUCTION

The presence of outliers can be due to errors when entering data, measurement and data collection errors, and the original nature of the data ([1], [2], [3], [4], [5]). But, sometimes, outliers can provide important information that no other observations can give, such as unusual circumstances. It can be excluded from the analysis if it is because of a measurement error or an error in preparing the measurement ([6], [7], [8]).

By far, there have been many methods for detecting outliers. Dovoedo and Chakraborti, [9], noted that many outlier detection methods work efficiently when the data has a symmetric distribution. Nevertheless, the methods fail to detect outliers in a skewed distribution. Tukey's boxplot ignores few outliers, meanwhile, sequential fences have a low probability of detecting outliers in a skewed distribution. Several studies have

Submitted: September 12, 2024 Reviewed: September 23, 2024 Accepted: October 23, 2024
DOI: http://dx.doi.org/10.18860/ca.v9i2.29434

been developed to see the effectiveness of various outlier detection methods to get the most accurate results. A comparison for outliers' detection methods in univariate data conducted by Charter, et al [10] who have shown that adjusted boxplot may be appropriate when the data has a skewed distribution. On the other side, Hubert and Vandervieren,[11] stated that Tukey's boxplot is a general method used to detect outliers, but the data skewness makes the results wrong. So, it needs further investigation into this difference. Besides, the development of outlier detection methods for skewed distribution was also carried out by Wong and Fitrianto ([12]) by involving the sequential fences into the adjusted sequential fences method.

This study uses actual data on air quality index to see the effectiveness of each outlier detection method in each province of Indonesia on June 08, 2021. Research related to detecting outliers in air quality index data has also been carried out by Putri and Sudarmilah,[13]. Based on fact, outlier detection on air quality index data is helpful as an automation measurement tool to notify the public about air quality in an area. This can increase public awareness regarding air quality issues, and the dangers of air pollution. Air pollution is one of the essential health hazards globally because it causes about 7 million premature deaths each year, and as many as 600,000 of these deaths are in children. Even air pollution is estimated to cost the global economy up to $2.9 trillion per year due to fossil fuel emissions and various severe environmental problems,[14]. Motor vehicles are the most significant contributor to pollution in Indonesia due to the rapid increase in motorized vehicles every year,[15]. However, the Covid-19 pandemic is a considerable factor affecting air quality during 2020. Worldwide lockdowns lead to a temporary reduction in fossil fuel consumption. It has a good impact in the form of a significant decrease in air pollution compared to previous years so that the air quality in the world tends to improve,[14].

Outlier detection in this kind of data can provide information to the public regarding air quality conditions in each province in an index. Air quality data has a predetermined threshold, we can compare the number of outliers based on the threshold with the results of outlier detection by different methods. Therefore, it is crucial to know the areas that still indicate poor air quality when implementing the Micro PPKM policy volume IX. If locations throughout Indonesia are within the limits of good air quality, then the policy suggests positive impacts. However, if there are still areas with poor air quality, the government must review the causes to resolve the problem appropriately. This research aims to compare the effectiveness of several outlier detection methods, namely Tukey's boxplot, adjusted boxplot, sequential fences, and adjusted sequential fences for univariate data with skewed distribution. The comparisons will be applied to the air quality index in each province of Indonesia to identify the areas of concern with poor air quality.

METHODS

Data

This research was carried out using data available in IQAir air quality platform on the website https://www.iqair.com. The air quality data comes from government and non-government air quality monitoring stations. Data from a few stations are used to anticipate the incidence of sensors at government stations having problems and reporting accurate data. Data processing and analysis were carried out using R software.

The data consists of the air quality index in 34 provinces in Indonesia on June 08, 2021. The reason for collecting data on June 08, 2021, is based on the Micro-Community Activity Restriction (PPKM), which has been running since January 2021. However, until eight series of the policy, it has not been fully implemented in 34 provinces of Indonesia. Until the government issued a regulation starting on June 01, 2021, the policy was simultaneously implemented in all provinces. It means that Micro PPKM Volume IX has been carried out for approximately one week. So that the implementation of the policy will be more stable in 34 provinces.

Analysis Procedures

The performance of outlier detection methods in this research is based on the number of outliers detected, the error rate, accuracy, precision, specificity, sensitivity values, and the strength of each observation against outliers. In general, the procedure for detecting outliers using the methods described in the previous section is as follows:

1. Doing data exploration
2. Defining true outliers on thresholds from data

   True outliers are outliers as a result from natural variation in a population. These outliers can provide valuable insights into the data distribution and the population's characteristics under study. True outliers can be identified based on thresholds and predefined categories of the air quality index. If the index is bigger, the air quality will be poor and unhealthy. In this research, observations with an air quality index of more than 100 were identified as outliers. The high air quality index is categorized as unhealthy that can interfere with health. So that there are four observations identified as true outliers and located on the right tail of the data distribution, which are 103, 105, 109, 117 from the provinces of North Sumatra, Jambi, Jakarta, and East Java.
3. Calculating and comparing the number of outliers detected by each method

   Procedures for comparing Tukey's boxplot, adjusted boxplot, sequential fences, and adjusted sequential fences methods are presented in a flow chart as follows (Figure 1):

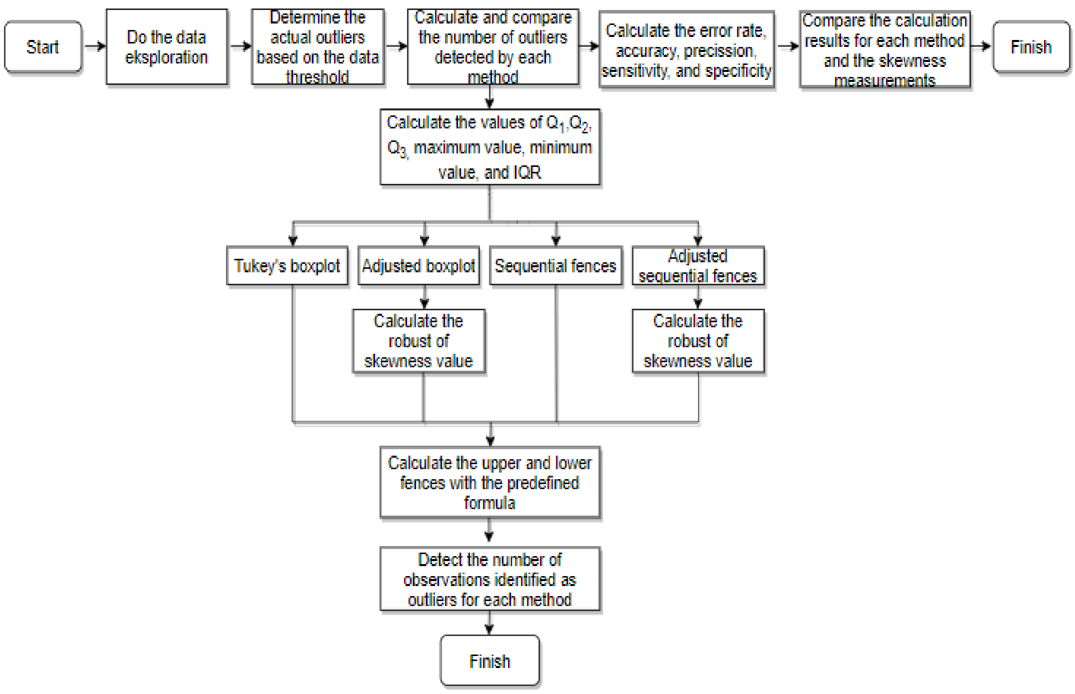

Figure 1. Flow chart of outlier detection procedure using Tukey's boxplot, adjusted boxplot, sequential fences, and adjusted sequential fences methods

Adjusted Boxplot

Hubert and Vandervieren, [11] have developed an adjusted boxplot method. There is an addition of measure skewness with medcouple for the adjusted boxplot method. According to Brys *et al.*, [16]for univariate samples with *n* sample size, $\{x_i, \dots, x_n\}$ sorted to be $x_1 \leq x_2 \leq \cdots \leq x_n$, median value or m_n is defined as follows.

$$m_n = \begin{cases} \frac{(x_{n/2} + x_{(n/2)+1})}{2} & , if\ n\ is\ even \\ x_{(n+1)/2} & , if\ n\ is\ odd \end{cases} \quad (1)$$

The medcouple ($\boldsymbol{MC_n}$)is defined as:

$$MC_n = med_{x_{i \leq m_n \leq x_j}} h(x_i, x_j) \quad (2)$$

when $x_i$is smaller or the same observations than median and x_j is the same or larger observations than the median, so $x_i \leq m_n \leq x_j$. If the value of $x_i \neq x_j$, the kernel function (h) is as follows:

$$h(x_i, x_j) = \frac{(x_j - m_n)(m_n - x_i)}{x_j - x_i} \quad (3)$$

There is a special case when $x_i = x_j = m_n$, the kernel function is defined when $m_1 < \cdots < m_n$ denotes the index of the observation which are tied to the median, so $x_{ml} = m_n$for all *l* = 1,2,...,*k* with *l* is the index number tied to the median and *k* is the highest index number of*l*. Then, [16] formulated it as follows:

$$h\left(x_{m_i}, x_{m_j}\right) = \begin{cases} -1 & if i + j - 1 < k \\ 0 & if i + j - 1 = k \\ +1 & if i + j - 1 > k \end{cases} \quad (4)$$

According to Chen, [17], the medcouple always lies between -1 and 1 because there is a denominator, $(x_j - x_i)$ from Eq. (3) as a comparator from $(x_j - m_n) - (m_n - x_i)$ to makes standardization. The medcouple has a positive value if the data is skewed to the right. Whereas, the medcouple becomes negative when the data is skewed to the left. When $x_j - m_n = m_n - x_i$, the medcouple is zero and has a symmetric distribution.

Hubert and Vandervieren, [11] incorporate the medcouple into the definition of the whiskers to adjusted the boxplot method and introducing the functions h_l(MC) and h_u(MC) to the outlier boundary. The boundaries are:

$$[Q_1 - \mathrm{h}_l(MC)(IQR);\ Q_3 + \mathrm{h}_u(MC)(IQR)]$$

Then, Hubert and Vandervieren, [11]compare three different models studied, namely the linear model, the quadratic model, and the exponential model, to get a suitable model for the adjusted boxplot method.

(1) Linear model

$$h_l(MC) = 1.5 + a(MC)$$
$$h_u(MC) = 1.5 + b(MC)$$

(2) Quadratic model

$$h_l(MC) = 1.5 + a_1(MC) + b_2(MC^2)$$
$$h_u(MC) = 1.5 + b_1(MC) + b_2(MC^2)$$

(3) Exponential model

$$h_l(MC) = 1.5(e^{a(MC)})$$
$$h_u(MC) = 1.5(e^{b(MC)})$$

when the value of *a*, a_1, a_2, *b*, b_1, $b_2 \in$ R.

Based on the result, the lower whiskers fail determined accurately by the linear model and the upper whiskers show less accuracy using the quadratic model than the exponential model. Therefore, the exponential model is appropriate for the upper and lower whiskers. So, the model will use to define the adjusted boxplot method. If the medcouple value > 0, the observations beyond the interval

$$[Q_1 - 1.5(e^{-4(MC)})(IQR);\ Q_3 + \ 1.5\ (e^{3(MC)}(IQR)]$$

will be identified as potential outliers. For the medcouple value < 0, the interval as follows:

$$[Q_1 - 1.5(e^{-3(MC)})\ (IQR);\ Q_3 + 1.5\ (e^{4(MC)}(IQR)]$$

Sequential Fences

Schwertman *et al.*, [18] proposed a common and simple fences method to detect outliers. Nevertheless, this method has one significant deficiency because it uses a single criterion for all sample sizes. It means in the large samples, only extremely unusual observations are identified as outliers because the fence is built far from the center and the possibility of real outliers will be neglected in small samples. Based on this problem, Schwertman and de Silva, [19] developed the new sequential fences procedure using Poisson distribution because of the possible misclassification of outliers and adjustment for various sample sizes. The IQR expectation value adjusts for various samples marker the standard deviation so that σ = $E(\frac{\boldsymbol{IQR}}{\boldsymbol{k_n}})$ with k_n is the appropriate constant from mathematic for various samples,. Then, Schwertman *et al.* [18] proposed the fences for the normal distribution as follows.

$$F_n = \ q_2 \pm \frac{IQR}{k_n} Z_\alpha \tag{5}$$

where q_i is the quartile with *i* is index quartile, IQR = $q_3 - q_1$, k_n the appropriate constant from mathematic for various samples.

Then, [19] explained that the fences were too large and some observations were misclassified as outliers, so modification was needed. Additionally, the least-squares quadratic equation for the approximating *t* distribution as marker degree of freedom (*df*) for sample sizes between 20 and 100 is

$$df \ = \ 7.6809524 \ + \ 0.5294157n - 0.00237n^2 \tag{6}$$

with n is the sample size in $20 \leq n \leq 100$. So, Schwertman and de Silva, [19] developed a formula by replacing Z with $t_{df,\alpha}$ on Eq. (5) as:

$$F_n = q_2 \pm \frac{IQR}{k_n} t_{df,\alpha} \tag{7}$$

The probability of detecting multiple outliers using m follows the Poisson model with the probability of $1 - \gamma$. For example, identifying the first outlier with m = 1 and γ = 0.05 means that 0.95 probability no observations beyond the fence detected as outliers and only 0.05 outliers probability that observations beyond the fence. Specifically, the sequential fences method for various samples by [19] has the following formula:

$$F_n = q_2 \pm \frac{t_{df,\alpha_{nm}}}{k_n} (IQR) \tag{8}$$

where *df* is calculated by Eq. (6), with m is the number of fences to identify observations that are suspected as outliers, n is the number of sample size, k_n is the constant value that is the adjustment to the expected value of the IQR with the standard deviation for various sample sizes.

Adjusted Sequential Fences

Wong and Fitrianto, [12] developed a method with adjustment sequential fences, especially for the data set that has a skewed distribution. The development of this method requires a combination of robust skewness. Bowley coefficient is denoted as follows.

$$\zeta = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

[12] built the sequential fences on the skewed distribution by incorporating the Bowley coefficient of skewness ζ into the sequential fences. The coefficient of IQR is substituted by functions like ζ, such that $f_l(\zeta)$ and $f_u(\zeta)$. The fences become

$$[Q_2 - f_l(\zeta)IQR, Q_2 + f_u(\zeta)IQR]$$

when $f_l(0) = f_u(0)$, it's the same as the standard sequential fences at symmetric distribution. Then, in asymmetric distributions, $f_l(\zeta)$ and $f_u(\zeta)$ use to adjust the fences.

RESULTS AND DISCUSSION

The Results of identifying Outliers for Each Method

The following is the result of detecting outliers using Tukey's boxplot, adjusted boxplot, sequential fences, and adjusted sequential fences methods:

Tukey's Boxplot

Outlier detection using Tukey's boxplot method shows failure to detect outliers (Figure 2). The formed fences included the lower inner fence is -35 and the upper inner fence is 125. Based on the inner fence, the fence formed was far from being observed to detect no outliers. It shows that Tukey's boxplot method is not robust against outliers and cannot work optimally on skewed data. Consequently, unusual observations are considered regular observations.

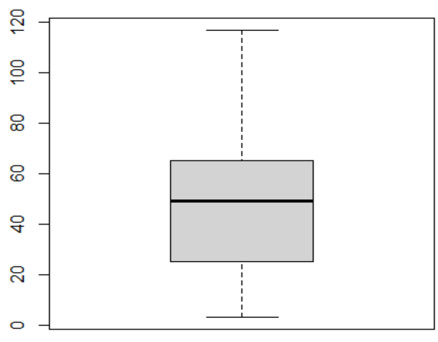

Figure 2. Tukey's boxplot of the air quality data

Adjusted Boxplot

The number of outliers detected using the adjusted boxplot method was four observations with the air quality index are 103, 105, 109, and 117 (Figure 3). These observations are indexes from the provinces of North Sumatra, Jambi, Jakarta, and East Java. The adjusted boxplot method forms a top fence is 101 and a lower fence is -65. Observations that exceed this value will be detected as outliers. This result is very satisfying because the four true outliers were detected correctly. It is also due to the effect of adjusting the method when used on skewed distribution data. The adjustment is in the form of adding the skewness coefficient called medcouple with the value is -0.1219512. It significantly affects the accuracy of this method in detecting outliers. The boundaries formed will be more robust and not affected by the outliers. So that all outliers were detected correctly.

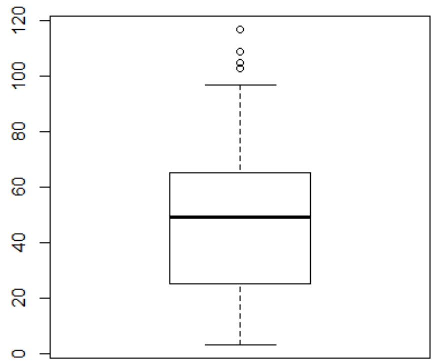

Figure 3. Adjusted boxplot of the air quality data

Sequential Fences

The sequential fences method can form several fences sequentially to detect outliers in the data with a significance level of 5%. In this case, the upper fence formed is 148.05969. While the value of the lower fence is -50.596857. After the first fence in the upper and lower fences was formed, the results show that no observations were identified as outliers. The formation of the upper and lower fences is stopped. If we look at Figure 8, the lower fence is not visible because its value is too small. So it can be concluded that no outlier was detected, and this method failed to detect outliers in the air quality index data.

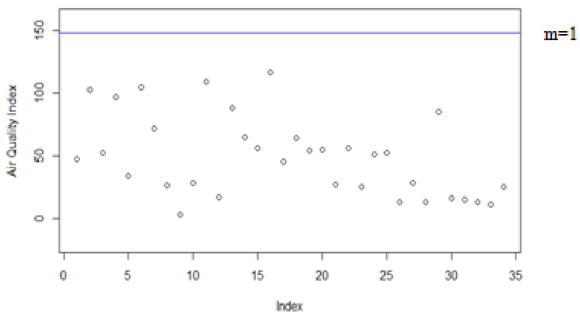

Figure 4. Sequential fences of the air quality data

Adjusted Sequential Fences

This outlier detection method has a skewness coefficient of -0.2. In addition, the adjusted sequential fences method forms three upper fences and one lower fence to detect outliers. The top three fences are 93.51039; 82.28991; 77.02420 and the one lower fences is -97.32036. The cause of this difference is that the number of outliers detected on the top fence increases until the third fence no longer detects outliers.

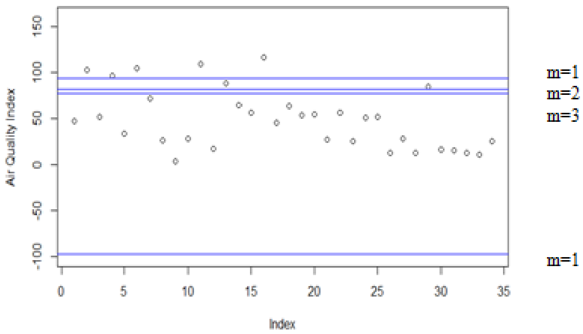

Figure 5. Adjusted sequential fences of the air quality data

As for the lower fence, it only forms a fence because no more outliers are detected. This method has succeeded in detecting seven outliers which are 85, 88, 97, 103, 105, 109, and 117 from South Sulawesi, West Java, Riau, North Sumatra, Jambi, Jakarta, and East Java. Four observations correspond to the true outliers, while the other observations are on the right tail. This method has been adjusted for data with skewed distribution and gets an additional skewness coefficient of -0.2. It causes the method to form a more robust boundary against outliers (Figure 5).

Outliers detection using the adjusted sequential fences method gives exciting results. As many as three provinces in Indonesia were detected as outliers like South Sulawesi, West Java, and Riau with blue color (Figure 6) even though this region was still in the moderate category. The high air quality index in these areas is an alarm for the government to pay attention and increase vigilance so that these areas do not fall into unhealthy air quality.

Figure 6. Indonesian map based on the result of adjusted sequential fences

Performance Evaluation

The performance of four outlier detection methods, namely Tukey's boxplot, adjusted boxplot, sequential fences, and adjusted sequential fences is evaluated. The evaluations are based on the number of outliers detected correctly according to the true outliers by looking at the error rate, accuracy, precision, sensitivity, specificity, and judging from the accuracy of detecting the tail side that contains outliers. The following are the results of the performance evaluation of each method.

Table 1. The performance evaluation of each method

Performances	Method			
	Tukey's Boxplot	**Adjusted Boxplot**	**Sequential Fences**	**Adjusted Sequential Fences**
Error Rate(%)	0	0	0	10
Accuracy (%)	88.235	100	88.235	91.176
Precision(%)	NaN	100	NaN	50
Sensitivity	0	1	0	1

Specificity	1	1	1	0.9
Skewness Coefficient	-	-0.122	-	-0.2
Outliers Tail Side	-	Right	-	Right

Based on the table above, each method has its advantages. Tukey's boxplot, Sequential Fences, and Adjusted Boxplot have very low error rate values of 0 and perfect specificity values of 1. Tukey's boxplot and sequential fences also show surprising results. They have the same performance values with the lowest accuracy of 88.235%, precision of NaN, poor sensitivity with 0 values, and very poor tail-side outlier results compared to other methods. The NaN values are caused because these methods cannot detect any outliers on air quality index data. Then, the adjusted sequential fence has a high accuracy of 91.176% after the adjusted boxplot, with a perfect accuracy value of 100%. Moreover, adjusted boxplot and adjusted sequential fences are the most sensitive method with a perfect value of 1. Overall, the adjusted boxplot method has the best error rate, accuracy, precision, sensitivity, and specificity values. It is due to the success of the adjusted boxplot method in identifying outliers according to true outliers.

It is also important to evaluate the accuracy of the detected outlier tail side. If the tail side of the outlier is detected incorrectly, there is a possibility that usual observations are classified as unusual observations. Such as detecting outliers in the air quality index data focused on the right side of the tail because we want to know observations classified as unhealthy air quality index. So there is a possibility that observations with a good air quality index can be classified as unusual observations because of errors in detecting the tail side of the outlier. According to the outlier detection results from each method, only the adjusted boxplot method and the adjusted sequential fences method successfully detected the outlier side based on the focus of this research. The skewness coefficient in both methods also supports this accuracy. Adjusted boxplot has skewness coefficient namely medcouple of -0.122 and adjusted sequential fences has skewness coefficient namely Bowley coefficient of -0.2. So, Tukey's boxplot and sequential fences methods are poor because they fail to detect outliers and outliers tail.

CONCLUSIONS

The study has shown that the adjusted boxplot performs better than the other methods based on the error rate, accuracy, precision, sensitivity, and specificity values. Meanwhile, adjusted boxplot and sequential fences have become the most sensitive methods and have higher accuracy than others. Concerning the accuracy of the tail side that contains outliers, the adjusted boxplot and adjusted sequential fences is the better method . The skewness coefficient also causes the accuracy in detecting the tail side of outliers because these methods are more robust to outliers. Both methods also have high sensitivity and precision than other methods

Unfortunately, it was found that Tukey's boxplot and sequential fences have poor accuracy, precision, sensitivity, and tail side outlier results than other methods. Tukey's boxplot and sequential fences methods are poor because they do not correctly detect true outliers and outliers on the tail side. Based on the results by each method, the areas of concern are South Sumatra, South Sulawesi, West Java, Riau, North Sumatra, Jambi, Jakarta, and East Java. Some sites such as South Sulawesi, West Java, Riau are moderate air quality. However, the regions are detected as an unusual observation by one of the methods used in this research because the observations have very high air quality index compared to other observations. The high air quality index in these areas alarms the government to pay attention and increase vigilance.

REFERENCES

[1] G. B. Begashaw and Y. B. Yohannes, "Review of outlier detection and identifying using robust regression model," *International Journal of Systems Science and applied mathematics*, vol. 5, no. 1, pp. 4–11, 2020.

[2] B. B. Alkan, C. Atakan, and N. Alkan, "A comparison of different procedures for principal component analysis in the presence of outliers," *J Appl Stat*, vol. 42, no. 8, pp. 1716–1722, 2015.

[3] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean J Anesthesiol*, vol. 70, no. 4, pp. 407–411, 2017.

[4] C. Leys, M. Delacre, Y. L. Mora, D. Lakens, and C. Ley, "How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration," *International Review of Social Psychology*, vol. 32, no. 1, pp. 1–10, 2019.

[5] J. W. Osborne and A. Overbay, "The power of outliers (and why researchers should always check for them)," *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, p. 6, 2019.

[6] N. L. Chaudhary and W. J. Lee, "Detecting and removing outliers in production data to enhance production forecasting," in *SPE Hydrocarbon Economics and Evaluation Symposium*, SPE, 2016, p. D021S005R004.

[7] V. F. Rofatto, M. T. Matsuoka, I. Klein, M. Roberto Veronez, and L. G. da Silveira Jr, "A Monte Carlo-based outlier diagnosis method for sensitivity analysis," *Remote Sens (Basel)*, vol. 12, no. 5, pp. 1–41, 2020.

[8] T. V Pollet and L. van der Meij, "To remove or not to remove: the impact of outlier handling on significance testing in testosterone data," *Adaptive Human Behavior and Physiology*, vol. 3, pp. 43–60, 2017.

[9] Y. H. Dovoedo and S. Chakraborti, "Boxplot-based phase I control charts for time between events," *Qual Reliab Eng Int*, vol. 28, no. 1, pp. 123–130, 2012.

[10] N. J. Carter, N. C. Schwertman, and T. L. Kiser, "A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry," *Stat Methodol*, vol. 6, no. 6, pp. 604–621, 2009.

[11] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Comput Stat Data Anal*, vol. 52, no. 12, pp. 5186–5201, 2008.

[12] H. S. Wong and A. Fitrianto, "Adjusted sequential fences for detecting uniovariate outliers in skewed distribution," *ASM Science Journal*, vol. 12, no. 5, pp. 107–115, 2019.

[13] D. A. P. Putri and E. Sudarmilah, "Comparative study for outlier detection in air quality data set," *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 11, pp. 584–592, 2019.

[14] IQAir, "2019 World Air Quality Report: Region & City PM2. 5 Ranking," *IQAir*, vol. 1, pp. 1–35, 2020.

[15] A. Masito, "Analisis risiko kualitas udara ambien (NO2 dan SO2) dan gangguan pernapasan pada masyarakat di Wilayah Kalianak Surabaya," *Jurnal Kesehatan Lingkungan*, vol. 10, no. 4, pp. 394–401, 2018.

[16] G. Brys, M. Hubert, and A. Struyf, "A robust measure of skewness," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 996–1017, 2004.

[17] D. Chen, "A Comparison of Alternative Bias-Corrections in the Bias-Corrected Bootstrap Test of Mediation," 2018.

[18] N. C. Schwertman, M. A. Owens, and R. Adnan, "A simple more general boxplot method for identifying outliers," *Comput Stat Data Anal*, vol. 47, no. 1, pp. 165–174, 2004.

[19] N. C. Schwertman and R. de Silva, "Identifying outliers with sequential fences," *Comput Stat Data Anal*, vol. 51, no. 8, pp. 3800–3810, 2007.