Pak. J. Statist.
2025 Vol. 41(1), 1-16

MODIFIED SPATIAL OUTLIERS DETECTION ALGORITHMS USING ROBUST MEDIAN-BASED MEASURES: A COMPARATIVE STUDY

Zetty Izzati Zulki Alwani[1], Rossita Mohamad Yunus[2§]
and **Adriana Irawati Nur Ibrahim[3]**
Institute of Mathematical Sciences, Faculty of Science
Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[1] zettyziza@gmail.com
[2] rossita@um.edu.my
[3] adrianaibrahim@um.edu.my
[§] Corresponding author

ABSTRACT

Some neighbourhood-based approach algorithms may have solved the swamping problem that spatial outlier identification techniques such as the Z algorithm and Moran's scatter plot failed to do. Nevertheless, there is a lack of thorough investigation into the algorithms' ability to identify outliers through simulation. We proposed modified versions of the Weighted Z and Average Difference Algorithms (Kou et al., 2006) by employing robust median-based measurements such as median and MAD. The PM10 concentration data obtained from 32 monitoring stations in Peninsular Malaysia were applied to the proposed and original algorithms. The results showed that the top four outliers in each algorithm's list were the same three stations, but in varying orders. We then assessed the detection performance of the proposed methods using simulated spatial datasets. The performance criteria were based on the average number of true outliers detected in N replications divided by the number of true outliers. We calculated the total number of cases with correct and false detections throughout 10,000 replications. The simulation analysis showed that the proposed methods were comparatively better than the original algorithms in terms of high correct detection and low false detection, regardless of sample size and nearest neighbors. This study demonstrates that adopting robust median-based measures in the procedure improves detection accuracy.

KEYWORDS

Spatial outlier, outlier detection, PM10 level, simulation, neighbours.

2020 Mathematics Subject Classification: 62G32, 62P12

1. INTRODUCTION

Outlier detection can serve as a tool for data cleaning and identifying interesting, uncovered properties of the data. The traditional outlier detection method does not take into account spatial relationships among input variables, while spatial patterns often indicate spatial continuity and autocorrelation with nearby objects (Chen et al. 2008); therefore, the traditional methods are plausible only for detecting the outliers in a non-

spatial manner. Unlike regular outliers, spatial outliers do not necessarily deviate from the overall data set. The identification of spatial outliers is based on two features: (i) the spatial attribute, which specifies the spatial relationship, such as locations and boundaries; and (ii) the non-spatial attribute, which describes spatial objects. A spatial outlier is a spatially referenced object whose non-spatial attribute values diverge significantly from those of other spatially referenced objects in its spatial neighbourhoods (Cressie, 1993; Sharma et al., 2022; Shukla and Lalitha, 2023).

Since the beginning of the 21st century, spatial outlier detection has drawn the attention of many researchers. Spatial outliers arise when an observation deviates from its close neighbours. Several different outlier detection methods were developed based on the principle of distinguishing spatial outliers from the rest of the data. Examples of this method are the Moran scatterplot (Anselin, 1995), scatterplot (Haining and Haining, 1993), and the Z algorithm (Shekhar et al., 2001). A scatterplot distinguishes the attribute value from the average of the attribute values over the neighborhood. Moran scatterplots distinguish between the normalised attribute value and the neighbourhood average of normalised attribute values. The Z-algorithm is developed based on normalising the difference between a point attribute and its neighbourhood attribute average value. However, it was pinpointed in a few works using synthetic datasets (Chen et al., 2008; Lu et al., 2003) that these methods had a drawback, that is, the regular points were falsely detected as spatial outliers (swamping) due to the presence of neighbouring points with very high/low attribute values. Lu and his research group (Chen et al., 2008; Lu et al., 2003; Cheng et al., 2019) had resolved this issue by introducing a few neighbourhood-based approach algorithms to detect spatial outliers effectively.

A common approach to accessing the spatial outlier in a real spatial dataset is drawing the top m ranks of spatial outliers using various outlier detection algorithms (Cheng et al.; 2019, Xu et al., 2018). In the study of detecting spatial outliers in actual data, although several of these algorithms identify the top m spatial outliers, the order of spatial outliers in the top m rank varies across them (Lu et al., 2003, Kou et al., 2006). This result demonstrates that the performance of these algorithms is not identical. However, assessing which algorithm performs better in detecting spatial outliers seems implausible when using a real dataset. While using a real spatial dataset, the actual spatial outlier may exist but is unknown. In other words, we are uncertain which method is superior in detecting spatial outliers because the actual one is unknown in a real dataset. To overcome this issue, we usually generate simulated spatial data that follow specific model assumptions, with spatial outliers identified in advance in the data. Then, we perform a comparative study from simulated datasets using some performance criteria. Although there are some studies that use the strategy for comparing the performance of outlier detection algorithms based on simulations (for example, Cai et al., 2021; Cai and Kwan, 2022), many works on outlier detection in the literature ignore it (Lu et al., 2003; Kou et al., 2006; Aggarwal et al., 2019; Peralta et al., 2023; Tian et al., 2023).

This paper is divided into five sections. The following section describes the algorithm of the proposed methods. Section 3 provides an illustrative example using synthetic spatial data of the outlier detection algorithms. In Section 4, we apply the algorithms to identify spatial outliers in the Malaysian PM10 concentration data. In Section 5, we compare the performance of several outlier detection methods using a Monte Carlo simulation of

generated spatial data with the spatial outliers known in advance in the data. The last chapter discusses the results and findings of the analysis and gives the conclusion.

2. DETECTION ALGORITHMS

We want to propose a modified version of the Weighted Z algorithm (Kou et al., 2006), where we used the median and median absolute deviation (MAD) instead of the mean and standard deviation; we rename it as the Median Weighted Z algorithm. The Weighted Z assigns different weights for different neighbours in computing the departure from the central object, where Z is termed due to the normalised difference between a spatial object and the weighted average of its spatial neighbors. The Median Weighted Z works similarly to the Weighted Z; however, the normalization using median and median absolute deviation highlights the outlierness from the central object.

Following Kou et al. (2006), suppose that $X = \{x_1, x_2, \dots, x_n\}$ is a set of spatial points, with a single or multiple spatial attributes such as a location with specified latitude and longitude. Suppose $f(x_i)$ is a non-spatial attribute value of spatial point x_i, where $i = 1, 2, \dots, n$. For each point x_i, let $NN_k(x_i)$ be a set of k nearest neighbors of spatial point x_i and $g(x_i)$ is the summary statistic for the non-spatial attribute values of the nearest neighbors of spatial point x_i in $NN_k(x_i)$. To detect the outliers, the attribute value of spatial point x_i is compared with the attribute values of its neighbors. Let $h(x_i)$ be the comparison of the attribute value of point x_i ($f(x_i)$) with its neighborhood. The principle that is being used is comparing the non-spatial attribute value of spatial point x_i, $f(x_i)$ and the neighborhood function that summarizes all the non-spatial attribute values of each spatial point x_i, $g(x_i)$. The comparison can be the difference or ratio of the two attribute values. Let m denote as the number of outliers to be identified, where $m \leq n$, where n is the number of data points. The Median Weighted Z spatial outlier detection algorithm is given as follows:

Algorithm 1 (Median Weighted Z)

Step 1: Find the k nearest neighbors set $NN_k(x_i)$ for each x_i. Then compute $d(x_i, x_j)$ the Euclidean distance between spatial point x_i and its neighbor x_j, where x_j belongs to $NN_k(x_i)$ for $j = 1, 2, \dots, k$, and $j \neq i$.

Step 2: For each x_i, compute the weighted average $w_j = d(x_i, x_j)^{-1} / \sum_{j=1}^{k} d(x_i, x_j)^{-1}$ for each spatial point in $NN_k(x_i)$.

Step 3: Find neighborhood function $g(x_i) = \sum_{j=1}^{k} (w_j \, . f(x_j))$, where $f(x_j)$ is the non-spatial attribute for spatial point x_j.

Step 4: Compute the comparison function $h_i = h(x_i) = f(x_i) - g(x_i)$.

Step 5: Compute $y_i = |(h_i - \mu^\star)/\sigma^*|$, where $\mu^\star$ is the median and σ^* is the median absolute deviation (MAD) of set $h = \{h_1, h_2, \dots, h_n\}$.

Step 6: Sort the values of y_i in ascending order. Say the values $\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_m$.

The top m of set $\{\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_r\}$ are identified as outliers.

We also want to propose a modified version of the Average Difference Algorithm (Kou et al., 2006), named the Median Average Difference. The Average Difference is based on the weighted average of the absolute difference between the spatial object and each of its neighbours, instead of obtaining the average of all its neighbours before comparison. The Median Average Difference works similarly to the Average Difference in the first four steps of the algorithm. In Step 5, the normalisation of the average difference using the median and median absolute deviation is carried out. The Median Average Difference spatial outlier detection algorithm is given as follows:

Algorithm 2 (Median Average Difference)

Steps 1 & 2: Steps 1 & 2 are the same as in Algorithm 1.

Step 3: For each x_i, find the set $\ell = \{\ell_j\}, j = 1, 2, \dots, k$ such that $\ell_j = |f(x_i) - f(x_j)|$ the absolute difference between attribute value of x_i and x_j for all $j \neq i$.

Step 4: Compute $\delta_i = \sum_{j=1}^{k} (w_j \,.\, \ell_j)$ for each x_i.

Step 5: Compute set $u = \{u_1, u_2, \dots, u_n\}$. where $u_i = |(\delta_i - \mu^{\dagger})/\sigma^{\dagger}|$, where $\mu^{\dagger}$ is the median and $\sigma^{\dagger}$ is the median absolute deviation (MAD) of set $\delta = \{\delta_1, \delta_2, \dots, \delta_n\}$.

Step 6: Sort the values u_i in ascending order. Say the values are $\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_r$.

The top m of set $\{\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_r\}$ are identified as outliers.

Deciding the m value in the top m-rank:

We choose the value of m in the same manner as detecting outliers in non-spatial data, where the data points with scores that are not within 1.5 or 2 standard deviations of the mean or median will be identified as outliers. The choice of the cutoff may depend on the skewness of the distribution of the scores and is often proceeded on a case-by-case basis. We illustrate the use of this principle in deciding the value of m in the following example.

3. ILLUSTRATIVE EXAMPLE USING A SYNTHETIC DATASET

This section provides an illustrative example of the proposed and original algorithms used for detecting spatial outliers created in a synthetic dataset.

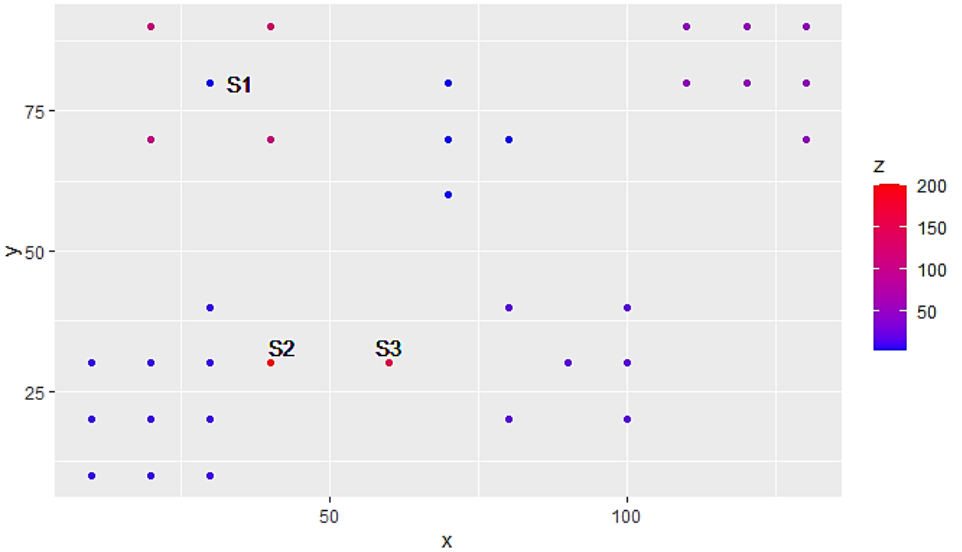

Figure 1: A Spatial Data Set. Objects are located in the X-Y Plane. The Colour Spectrum represents the Attribute Value of the Corresponding Object

As shown in Figure 1, all 34 object locations are in the $X - Y$ plane, with its associate attribute value described by a colour spectrum that ranges from 0 to 200. The expected number of spatial outliers is 3, and we have chosen k equals 3 in the algorithm. We can easily see that spatial points S1, S2, and S3 are spatial outliers because these objects' attribute values are significantly different from those in the neighbourhood.

The scores of all algorithms are depicted in Figure 2 for all spatial objects. We chose outliers in the top three because three objects scored more than 1.5 standard deviations of the mean for the Weighted Z, Median Weighted Z, and Median Average Difference algorithms.

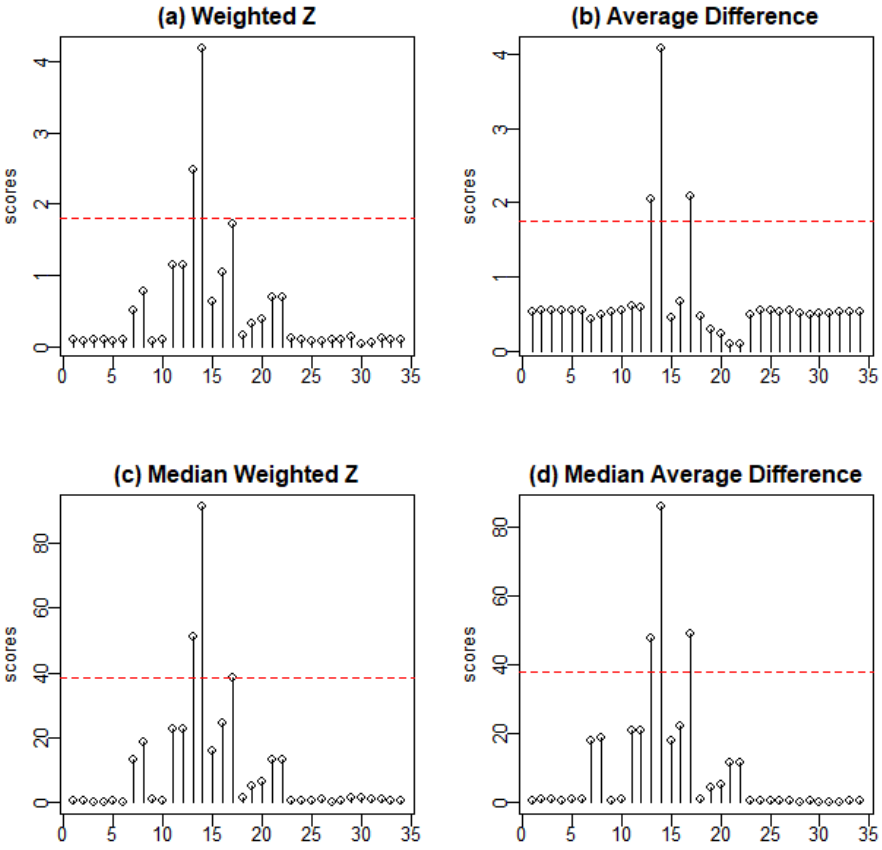

Figure 2: Scores of Spatial Points from the Four Algorithms. The Dotted Red Line represents a 1.5 Standard Deviation of the Mean

The top three spatial outlier candidates are the same for all algorithms. However, the rank of the outliers is different. The top three-rank spatial outliers for Weighted Z and Median Weighted Z are S2, S1, and S3, while the Average Difference and Median Average Difference give another different order of outliers: S2, S3, and S1.

4. APPLICATION ON THE REAL DATASET

One role of air quality monitoring is to provide information on the concentration of pollution in the environment. Detecting abnormal data (or outliers) in the network of air quality monitoring stations is crucial in monitoring air quality (Buelvas et al., 2023; Rollo et al., 2023). In Malaysia, particulate matter determines the air quality index and is recorded hourly across the country. Particulate matter sizes less than 10 microns (PM10) can go deep into the lungs and reduce lung function. In the context of spatial outlier detection, successful identification of a location with a concentration of PM10 level significantly different than its neighbourhood will embark on the research for discovering the indicator of such a difference.

The dataset consists of the concentration of PM10 of 32 monitoring stations in Peninsular Malaysia recorded on the 1st of January 2017, at 2 am. The data are provided by the Department of Environment Malaysia and available on the Open Data Malaysia website. The location of each station is determined by the latitude and longitude of the stations. In this section, we aim to identify spatial outliers in the PM10 monitoring stations using the outlier detection algorithms studied in this paper. In the analysis, each station was treated as a spatial object, and the number of neighbours for each station was chosen based on the Euclidean distance.

Figure 3 shows the location of 32 monitoring stations on the map of Peninsular Malaysia. The stations are labelled by numbers. The colour indicates the level of PM10 concentration. Blue spectrum colours indicate a low level of PM10 concentration, while pink spectrum colors indicate a high level of PM10 concentration. In this analysis, we choose the number of neighbours $k = 4$, which means we compare the level of PM10 of a station with its 4 nearest neighbours only in each algorithm.

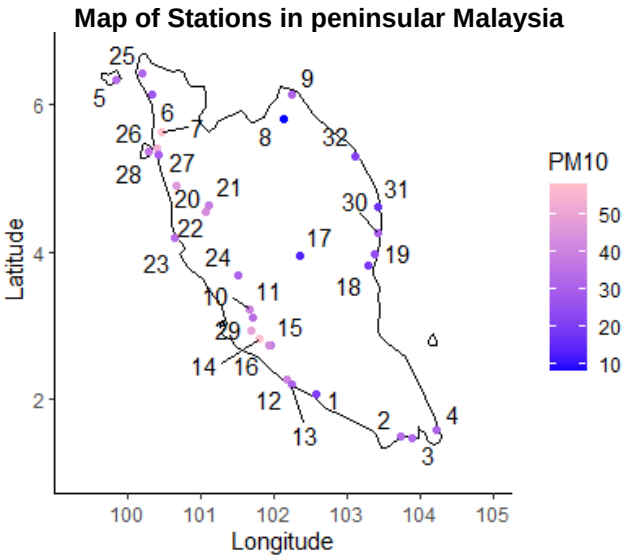

Figure 3: PM10 Levels for each Station (Date: 1/1/2017, Time: 2 am)

Table 1 shows the top four candidates detected as outliers for each method. All methods have chosen Station 8 as the spatial outlier with the top position in the ranking. Station 8 is in the northeast of Peninsular Malaysia. The four nearest neighbours of Station 8 are Stations 9, 21, 22, and 32 with respective attribute values of 31, 38, 43, and 20 $\mu g/m^3$. The attribute value of Station 8 is 8 $\mu g/m^3$, and it is significantly smaller than its neighbour attribute values. Thus, Station 8 is a spatial outlier.

The level of PM10 for Station 7 is 58 $\mu g/m^3$, while for its four nearest neighbours Station 6, Station 26, Station 27 and Station 28 are 27, 53, 29 and 37 $\mu g/m^3$, respectively. Hence, Station 7 is a spatial outlier because its attribute value is a bit larger than its neighbours' attributes. Although all methods have chosen Station 7 as a spatial outlier, the

rank order of Station 7 in the top 4 candidates is not the same for these methods. It is chosen as the second highest for the Weighted Z and Median Weighted Z algorithms, but it is chosen in the fourth position by the Average Difference and Median Average Difference.

Table 1
Top Four Outlier Candidates in the Dataset Detected by each Algorithm

Methods	Top Four Outlier Candidates
Weighted Z	1. Station 8: Sekolah Menengah Tanah Merah, Tanah Merah, Kelantan 2. Station 7: Sekolah Menengah Kebangsaan Tunku Ismail, Kedah 3. Station 14: Sekolah Kebangsaan Taman Semarak (Fasa 2), Negeri Sembilan 4. Station 27: Kolej Vokasional Seberang Perai at Pulau Pinang
Median Weighted Z	1. Station 8: Sekolah Menengah Tanah Merah, Tanah Merah, Kelantan 2. Station 7: Sekolah Menengah Kebangsaan Tunku Ismail, Sungai Petani, Kedah 3. Station 27: Kolej Vokasional Seberang Perai at Pulau Pinang 4. Station 14: Sekolah Kebangsaan Taman Semarak (Fasa 2), Negeri Sembilan
Average Difference	1. Station 8: Sekolah Menengah Tanah Merah, Tanah Merah, Kelantan 2. Station 27: Kolej Vokasional Seberang Perai at Pulau Pinang 3. Station 9: Sekolah Menengah Kebangsaan Tanjong Chat, Kota Bharu, Kelantan 4. Station 7: Sekolah Menengah Kebangsaan Tunku Ismail, Sungai Petani, Kedah
Median Average Difference	1. Station 8: Sekolah Menengah Tanah Merah, Tanah Merah, Kelantan 2. Station 27: Kolej Vokasional Seberang Perai at Pulau Pinang 3. Station 9: Sekolah Menengah Kebangsaan Tanjong Chat, Kota Bharu, Kelantan 4. Station 7: Sekolah Menengah Kebangsaan Tunku Ismail, Sungai Petani, Kedah

Station 27 is a spatial outlier because Station 27's attribute value is 29 $\mu g/m^3$, quite a lower PM10 level than its neighbours, Stations 7, 20, 26, and 28, with attribute values of 58, 47, 53, and 37 $\mu g/m^3$, respectively. The rank order of Station 27 is fourth in the rank for Weighted Z, third in the rank for Median Weighted Z, and second top for Average Difference and Median Average Difference algorithms.

Station 14 is detected as a spatial outlier by Weighted Z and Median Weighted Z, but not detected by the other two methods in the top 4 rank. The attribute value of Station 14 is 58 $\mu g/m^3$, while its neighbours, Stations 11, 15, 16, and 29, are 34, 33, 42, and 49 $\mu g/m^3$, respectively. Station 14 could be an outlier because its attribute value is larger than its neighbors. Station 9 is detected as a spatial outlier by the Average Difference and Median Average Difference. The attribute value of Station 9 is 31 $\mu g/m^3$, while its neighbours' Stations 7, 8, 21, and 32 are 58, 8, 38, and 20 $\mu g/m^3$, respectively. The detection of Station 9 as a spatial outlier seems to be interesting. The data show a wide range of attribute values for the neighbours of Station 9 that is from 8 to 59 $\mu g/m^3$, while

the Station 9 attribute value is included in this range. There is a significant variation in PM10 levels for this neighbourhood.

Therefore, we conclude that Stations 8, 7, and 27 are spatial outliers because they are selected in the top 4 by all the algorithms. In this section, we have used all four algorithms to identify spatial outliers. In the next section, we compare the four algorithms in the sense of which detection method is better.

5. SIMULATION STUDIES

In this section, we simulated spatial datasets to compare the performance of four spatial outlier detection methods, namely the Weighted Z, Average Difference, and the modified version of the two algorithms, which are the Median Weighted Z and the Median Average Difference.

To do the simulation, first we generated spatial data. We followed the procedure suggested by Dorman et al. (2007) to generate the spatial data. First, we randomly generated the spatial attribute, identified by the latitude and the longitude, each from a uniform distribution $U(0,100)$. Then, we computed the distance matrix between spatial points $\mathrm{D} = (d_{ij})$, where d_{ij} is the Euclidean distance between cells i and j, for $i = 1,2,\dots,n$, $j = 1,2,\dots,n$, and $i \neq j$. For $i = j$, $d_{ij} = 0$. Next, we found the correlation matrix $\Omega = (w_{ij})$, where $w_{ij} = \exp(-p\, d_{ij})$ with p is a parameter that determines the correlation with an inter-cell error. If the value of p increases, then the strength of autocorrelation is also increasing. However, $p = 0$ indicates that there is no correlation. Thus, in this paper, we set $p = 0.0001$ to make the data more homogenous. Then, we computed the spatial non-attribute $z = \mathrm{W}^{\mathrm{T}}\lambda$ where λ was drawn from the standard normal distribution, and the weight matrix W was calculated by the Cholesky decomposition of $\Omega = \mathrm{W}^{T}\mathrm{W}$.

For the simulation, we used four different sample sizes denoted as n, where $n = 40, 60, 80,$ and 100. To create spatial data with 5% spatial outliers, we replaced 5% of each sample data size n generated using the Dorman et al. (2007) procedure with spatial outliers. We applied two steps. The first step is to choose the spatial attributes, and the second step is to replace the data with a spatial outlier, which is a value that deviates from the values in its neighborhood. For the first step, we arranged the non-spatial attribute value of the simulated dataset in ascending order to ensure that the locations of outliers are distributed evenly across the entire spatial attribute. Let β be the number of the outliers corresponding with 5 percent of outliers in the data. So, β is fixed for each sample size n. For the second step, after choosing β points, we replaced the values of the chosen β points with a value that is between the minimum and maximum of the original generated data and different than its neighborhood (Singh and Lalitha, 2018). The condition of choosing a value between the maximum and minimum of the value of generated data is important to ensure that no global outliers are falsely detected as spatial (local) outliers. Then, we used the spatial data with known outliers in advance to study the performance of all spatial outlier detection methods. To obtain the top m outlier ranking for all the algorithms, a plausible value of m is β in this simulation study.

We generated the dataset with 10,000 replications. To compare the performance of the algorithms, we used the proportion of means of detecting true outliers as the performance criteria. The proportion of the mean of detecting true outliers is defined as the average of the number of true outliers detected in 10,000 replications divided by the number of true outliers β. The closer the proportion of the mean to one, the better the performance of the algorithm.

For each algorithm and sample size $n = 40$, and selected values of nearest neighbor k, the spatial data with spatial outliers were generated 10,000 times, then, the proportions of the mean of detecting true outliers were calculated and presented in Figure 4.

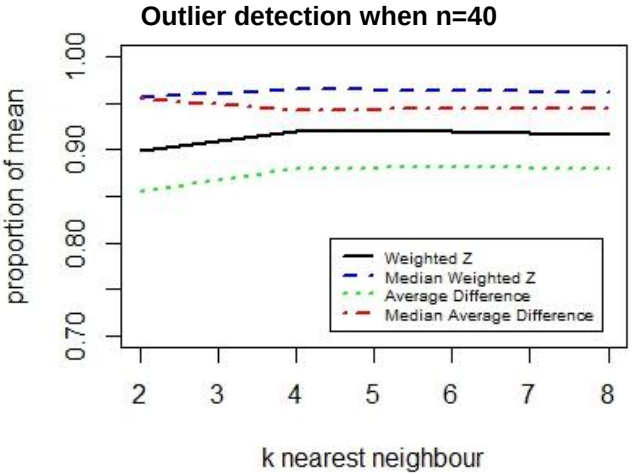

Figure 4: Performance of all Methods when $n = 40$ as the Number of Nearest Neighbors Increases

It is depicted in Figure 4 that both the Median Weighted Z and Median Average Difference methods are better than the other two methods because they have a higher proportion of the mean of detecting true outliers. The same situation was observed when $n = 60, 80,$ and 100. The number of nearest neighbours that gives the higher proportion of the mean of detecting true outliers ranges from $k = 4$ to 6.

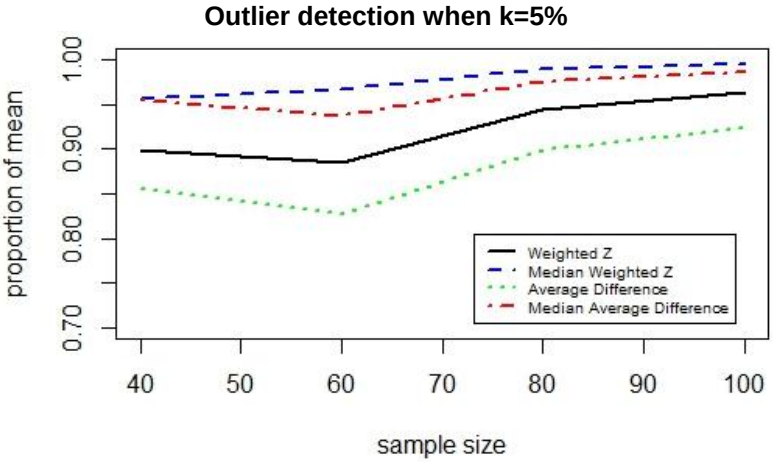

Figure 5: Performance of all Methods when $100(k/n) = 5\%$ as the Number of Sample Size Increases

Figure 5 shows the proportions of the mean of detecting true outliers for each algorithm, as a function of sample size n, where the percentage of nearest neighbor is $100(k/n) = 5\%$. The proportion of the mean of detecting true outliers is increasing steadily as n increases for the two proposed methods. The Median Weighted Z and Median Average Difference methods have better performance compared to the Weighted Z and Average Difference methods in detecting true outliers.

To investigate whether the methods could have solved the issue of falsely choosing some good points as spatial outliers and not detecting the outliers correctly, we consider three cases of detection: Case 1: Detect all true outliers; Case 2: Not detect any true outlier; and Case 3: Detect some true outliers and some good points.

We may detect all true outliers in a simulated dataset using an algorithm. We may fail to do so or can identify some true outliers only while falsely detecting the good points as outliers in the top m. For the simulation of 10,000 times, we then calculate the number of times we observed Case 1, Case 2, and Case 3.

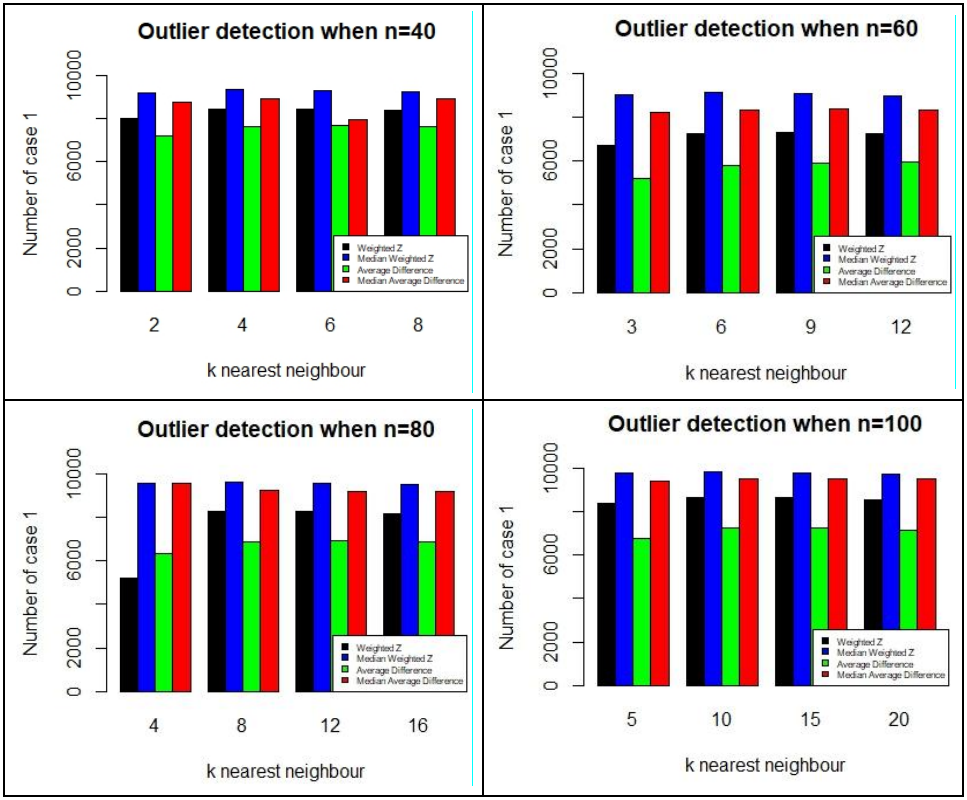

Figure 6: Case 1- Detect all True Outliers

Figure 6 shows the number of Case 1 for each algorithm as a function of nearest neighbour k when the percentage of outliers is 5%. The number of Case 1 for the Median Weighted Z is greater than the Weighted Z. More Case 1 was detected using the Median Average Difference than the Average Difference. Both the modified methods have a higher percentage of detecting all true outliers than the other two, regardless of the sample size and the number of nearest neighbours.

Figure 7 shows that fewer number of Case 2 were observed by the modified methods compared to the other two when $n = 40$, On the contrary, they perform differently in observing Case 2 when $n = 60$. However, the number of Case 2 is insignificant because it is under 10 cases out of 10,000. No Case 2 was detected by all the algorithms when $n = 80$ and 100. Hence, we concluded that the modified versions have a reasonably good performance with little chance of failure in not detecting all the true outliers.

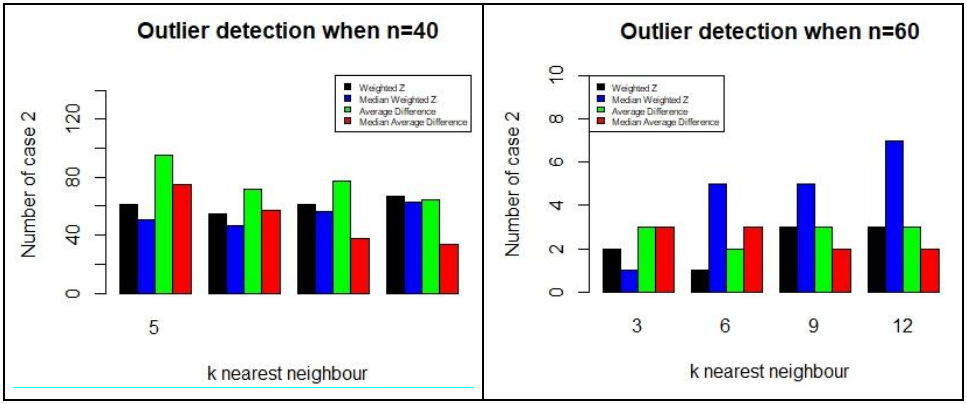

Figure 7: Case 2- Not Detect any True Outlier

In Figure 8, we observe that both the Weighted Z and Average algorithms execute more Case 3 than their modified versions. This result means that by using both the modified methods, we tend to have a lower percentage of cases of falsely detecting good points as outliers than the other two original versions, regardless of the sample size and the nearest neighbours.

We conclude from the simulation study that the Median Weighted Z and Median Average Difference are better than the Weighted Z and Average Difference methods for any value of n and k. Although the Weighted Z and Average Difference can detect more than roughly 80% of true outliers, the Median Weighted Z and Median Average Difference consistently outperform, with a higher percentage of cases in correct detection of outliers and a smaller percentage in falsely detecting good points as outliers.

6. CONCLUSION

We have modified two spatial outlier detection methods using robust median and median absolute deviation in the algorithms. We have assessed the performance of the proposed methods in outlier detection using actual and simulated data and have compared them to the two original methods. All these methods have chosen the true spatial outliers in the top three candidates for the synthetic data. However, the ranking of the spatial outliers in the top three differs for both original and proposed methods. We have successfully applied all of the algorithms to detect outliers in the PM10 concentration data. Then, we discovered that all algorithms have identified the same three monitoring stations as spatial outlier candidates in the top four ranks, but in varying orders. The difference in order indicates that the performance of the algorithms is dissimilar.

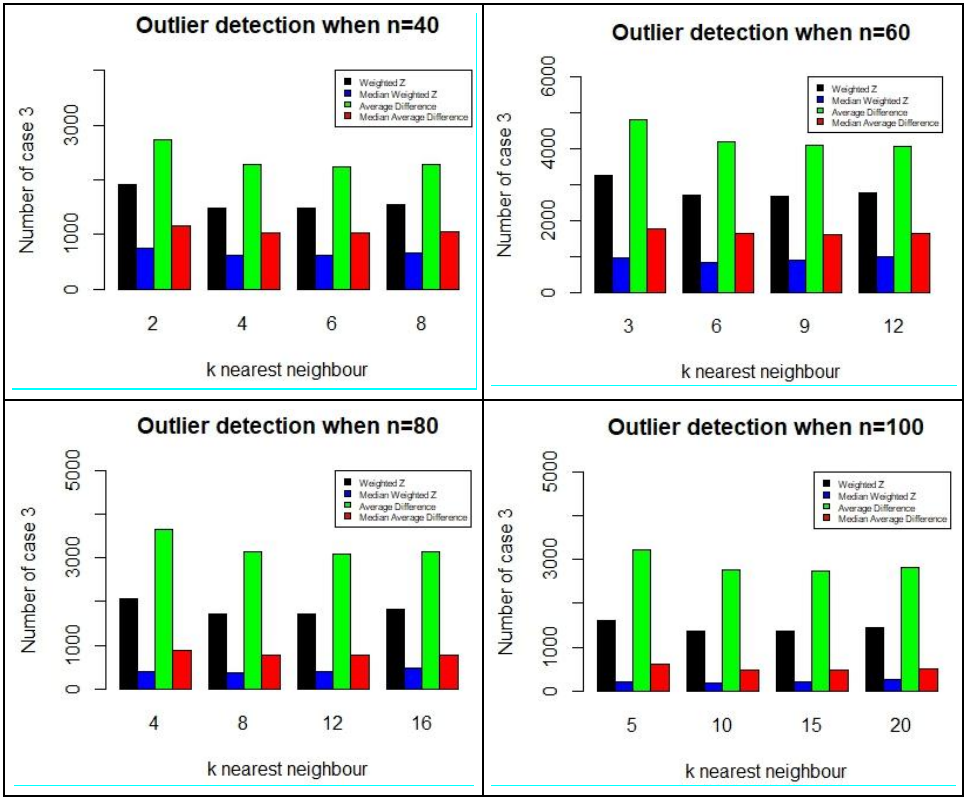

Figure 8: Case 3- Detect some Outliers and some False Outliers

If we choose a different value of the top m rank, then we would obtain different spatial outliers for the chosen m. In practice, we may choose the value of m in the same manner as detecting outliers in non-spatial data. As explained in Saleem et al. (2021), data points with scores that are not within 1.5 SD, 2 SD, or 3 SD of the mean or median can be identified as outliers. The choice of the cutoff point may be influenced by the skewness of the scores and requires further study or is often proceeded on a case-by-case basis. In this paper, we compare the efficiency of the algorithms once we have determined the value of m. In the simulation, we set the value of m to be the number of outliers that we have created in the spatial data for the ease of method comparison.

To compare the performance of the outlier detection algorithms, we have simulated spatial data with spatial outliers. Based on the simulation, the proposed Median Weighted Z and Median Average Difference outperformed the Weighted Z and Average Difference regardless of sample size and number of nearest neighbours. Using the Median Weighted Z and Median Average Difference algorithms, we tend to obtain a higher chance of detecting the true outliers and a lower chance of false detection of good points as outliers. This research found that adopting robust measures in the outlier detection procedures improved the detection accuracy of spatial outliers. The proposed methods for identifying outliers among spatial points can help with pollution control in environmental engineering.

ACKNOWLEDGMENT

Funding: This work was supported Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme FP083-2018A [FRGS/1/2018/STG06/UM/02/12], and Universiti Malaya Research Grant RF015B-2018. The third author was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme [FRGS/1/2019/STG06/UM/02/2].

Conflict of interest:

The authors declare that they have no conflict of interest.

Ethical approval:

This article does not contain any studies with animals or human participants by the author.

Data availability:

The data that support the findings of this study are provided by the Department of Environmental Malaysia and are openly available in Malaysia Open Data Portal at:
https://www.data.gov.my/data/ms_MY/dataset/?q=kualiti+udara&sort=title_string+asc

REFERENCE

1. Aggarwal, V., Gupta, V., Singh, P., Sharma, K. and Sharma, N. (2019). Detection of spatial outlier by using improved Z-score test. In 2019 *3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 788-790). doi 10.1109/ICOEI. 2019.8862582
2. Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, 27(2), 93-115.
3. Buelvas, J., Múnera, D., Tobón V, D.P., Aguirre, J. and Gaviria, N. (2023). Data quality in IoT-based air quality monitoring systems: a systematic mapping study. *Water, Air, & Soil Pollution*, 234(4), 248. https://doi.org/10.1007/s11270-023-06127-9
4. Cai, J. and Kwan, M.P. (2022). Detecting spatial flow outliers in the presence of spatial autocorrelation. *Computers, Environment and Urban Systems*, 96, 101833.
5. Cai, J., Deng, M., Guo, Y., Xie, Y. and Shekhar, S. (2021). Discovering regions of anomalous spatial co-locations. *International Journal of Geographical Information Science*, 35(5), 974-998.
6. Chen, D., Lu, C.T., Kou, Y. and Chen, F. (2008). On detecting spatial outliers. *Geoinformatica*, 12, 455-475.
7. Cheng, Z., Zou, C. and Dong, J. (2019). Outlier detection using isolation forest and local outlier factor. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, Chongqing, China (pp. 161-168).
8. Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Son, New York (NY).
9. F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W. and Kühn, I. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5), 609-628.

10. Haining, R.P. and Haining, R. (1993). *Spatial data analysis in the social and environmental sciences*. Cambridge University Press.
11. Kou, Y., Lu, C.T. and Chen, D. (2006). Spatial weighted outlier detection. In *Proceedings of the 2006 SIAM international conference on data mining* (pp. 614-618). Society for Industrial and Applied Mathematics.
12. Lu, C.T., Chen, D. and Kou, Y. (2003). Detecting spatial outliers with multiple attributes. In *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence* (pp. 122-128). IEEE.
13. Peralta, B., Soria, R., Nicolis, O., Ruggeri, F., Caro, L. and Bronfman, A. (2023). Outlier vehicle trajectory detection using deep autoencoders in Santiago, Chile. *Sensors*, 23(3), 1440.
14. Rollo, F., Bachechi, C. and Po, L. (2023). Anomaly detection and repairing for improving air quality monitoring. *Sensors*, 23(2), 640.
15. Saleem, S., Aslam, M. and Shaukat, M.R. (2021). A Review and Empirical Comparison of Univariate Outlier Detection Methods. *Pak. J. Statist.*, 37(4), 447-462.
16. Sharma, A., Jiang, Z. and Shekhar, S. (2022). Spatiotemporal data mining. In *Handbook of Spatial Analysis in the Social Sciences*, edited by S.J. Rey, R.S. Franklin, United Kingdom; Edward Elgar Publishing.
17. Shekhar, S., Lu, C.T. and Zhang, P. (2001). Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 371-376). San Francisco, California, United States.
18. Shukla, S. and Lalitha, S. (2023). Geographically Weighted Comedian method for spatial outlier detection. *Japanese Journal of Statistics and Data Science*, 6(1), 279-299.
19. Singh, A.K. and Lalitha, S. (2018). A novel spatial outlier detection technique. *Communications in Statistics-Theory and Methods*, 47(1), 247-257.
20. Tian, Z., Zhuo, M., Liu, L., Chen, J. and Zhou, S. (2023). Anomaly detection using spatial and temporal information in multivariate time series. *Scientific Reports*, 13(1), 4400.
21. Xu, X., Liu, H., Li, L. and Yao, M. (2018). A comparison of outlier detection techniques for high-dimensional data. *International Journal of Computational Intelligence Systems*, 11(1), 652-662.