ComTech: Computer, Mathematics and Engineering Applications, 13(2), December 2022, 111−121 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v13i2.7821 E-ISSN: 2476-907X

An Improved Weighted Median Algorithm for Spatial Outliers Detection

Zerlita Fahdha Pusdiktasari[1]*, Rahma Fitriani[2], and Eni Sumarminingsih[3]

[1-3]Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Brawijaya
Jln. Veteran, Jawa Timur 65145, Indonesia
[1]zerlitafahdha@gmail.com; [2]rahmafitriani@ub.ac.id; [3]eni_stat@ub.ac.id

Received: 19th October 2021/ **Revised:** 4th January 2022/ **Accepted:** 5th January 2022

How to Cite: Pusdiktasari, Z. F., Fitriani, R., & Sumarminingsih, E. (2022). An Improved Weighted Median Algorithm for Spatial Outliers Detection. *ComTech: Computer, Mathematics and Engineering Applications, 13*(2), 111−121. https://doi.org/10.21512/comtech.v13i2.7821

***Abstract* -** A spatial outlier is an object that significantly deviates from its surrounding neighbors. The median algorithm is one of the spatial outlier methods, which is robust. However, it assumes that all spatial objects have the same characteristics. Meanwhile, the Average Difference Algorithm (AvgDiff) has accommodated the differences in spatial characteristics, but it does not use statistical tests to determine the status of an object, whether it is an outlier or not. The research developed an improved version of the median algorithm and AvgDiff, called the Weighted Median Algorithm (WMA). WMA combined the advantages of the two methods. From the median algorithm, WMA adopted median and statistical test concepts. Meanwhile, from AvgDiff, WMA adopted the concept of using differences in objects' spatial characteristics as weights. A combination of the two advantages was innovated by calculating WMA's neighborhood score using a weighted median. Then, a simulation was conducted to analyze the accuracy of the method. The result confirms that when objects have heterogeneous spatial characteristics, WMA performs better than the median algorithm. The accuracy of WMA is not much higher than AvgDiff, but the use of WMA can prevent a serious false detection problem. The methods can be applied to an incidence rate of Covid-19 data in East Java.

Keywords: Weighted Median Algorithm (WMA), spatial outliers, average difference algorithm

I. INTRODUCTION

Outliers are objects in a data set with extreme values which deviate from other objects. Such deviations raise suspicions that the objects come from different mechanisms from the rest of the data. In a spatial context, outliers are referred to as spatial outliers. Spatial outliers are objects whose nonspatial attribute values differ significantly from their surrounding objects (nearest neighbors) (Shekhar, Lu, & Zhang, 2001). The spatial outlier detection method pays attention to the spatial correlation and spatial arrangement or position between spatial objects. Spatial outlier detection can lead to the discovery of special patterns, meaningful insights, and important implicit information. The relevant methods have been applied to many cases, such as identifying air pollutant networks (Araki, Shimadera, Yamamoto, & Kondo, 2017; Van Zoest, Stein, & Hoek, 2018; Wu, Tang, & Wang, 2018), leaking water from supply pipelines (Helwig, Guggenberger, Elmore, & Uetrecht, 2019), water pollutant (Shukla & Lalitha, 2021), potential mineralization (Nguyen, Vu, Trinh, & Nguyen, 2016), hot spots of soil pollution (Fu, Zhao, Zhang, Wu, & Tunney, 2016; Tepanosyan, Sahakyan, Zhang, & Saghatelyan, 2019; Xiao, Wang, Hou, & Erten, 2020), traffic density caused by unexpected or temporary incidents like traffic accidents and celebration (Djenouri & Zimek, 2018; Pu, Wang, Liu, & Zhang, 2019; Tang & Ngan, 2016), anomalous system behavior of Wireless Sensor Networks (WSN) (Ayadi, Ghorbel, Obeid, & Abid, 2017; Bosman, Iacca, Tejada, Wörtche, & Liotta, 2017), anomalous teen birth rate (Khan et al., 2017), and others.

In addition, the detection of spatial outliers can also be applied to the cases of Covid-19, which is recently a problem in all countries of the world (Baba, Midi, & Abd Rahman, 2021; Xia, An, Li, & Zhang, 2022). The outlier, in this case, is an area that behaves unusually based on the incidence rate of Covid-19. An area is considered normal (behaves naturally) if the area has almost the same high incidence rate as the surrounding areas. Areas that become outliers indicate factors that influence the number of cases, apart from the factor of transmission or spread from surrounding areas. For the rest of the research, the term 'outliers' is referred to as 'spatial outliers'.

*Corresponding Author

In spatial data, an object which is the center of attention is called a central object. The other objects surrounding a central object are defined as the nearest neighbors. The condition of the neighboring objects is summarized into a neighborhood score. Then, spatial outlier detection is performed by comparing the nonspatial attribute values of the central object with its neighborhood scores. A comparison score measures the comparison. If a central object extremely differs from its neighbors, the comparison score will be extremely large. In that case, the central object is considered an outlier (Lu, Chen, & Kou, 2003; Shekhar et al., 2001). Therefore, the neighborhood score must be carefully formed because it can describe the neighborhood's condition accurately.

The problem that often arises in the formation of spatial outlier detection algorithms is how to define the neighborhood score that well represents the condition of the neighbors. Previous studies such as spatial statistics by Shekhar et al. (2001), iterative z and iterative r by Lu et al. (2003), and improved z score by Aggarwal, Gupta, Singh, Sharma, and Sharma (2019) to calculate the neighborhood score. Others utilize the median, such as the median algorithm by Chen, Lu, Kou, and Chen (2008). Each method has its advantages and disadvantages. The use of mean is the simplest method, but it can mistakenly detect a central-normal object as a spatial outlier due to the true outlier in its neighborhood. This condition is known as the swamping effect (Baba et al., 2021; Kolbaşi & Ünsal, 2019; Wang & Serfling, 2018). The neighbor scores are pulled towards outlier values which cause the comparison score to be higher than it should be. The swamping effect is less likely to happen when the median is used. It is due to the robust nature of the median (Sajana & Sajesh, 2018). However, both mean and median-based methods still assume that all spatial objects have the same spatial characteristics (homogeneous). In many fields of application, the area of the objects, the distance between objects, or the length of the shared border between the objects tend to be different (Taha, Onsi, El Din, & Hegazy, 2019).

The spatial characteristics affect the relationship of an object with its neighbors. Figure 1 shows the importance of accommodating spatial characteristics in calculating neighboring scores. If A is the central object, B, C, and D become the nearest neighbors of object A. The D is known as an outlier whose nonspatial attribute value is extremely different from A, B, and C. The mean and median methods will interpret these conditions as illustrated in Figure 2.

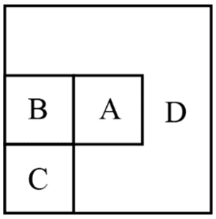

Figure 1 The Illustration of Spatial Configuration with Different Spatial Characteristics

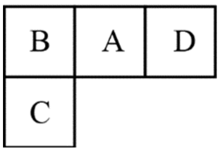

Figure 2 The Illustration of Spatial Configuration with Homogen Spatial Characteristics

All objects have the same characteristics. Based on the concept of mean and median, B or C will be chosen as a representation of the neighboring conditions of A. It causes A to be detected as a normal object because it is not significantly different from its neighbors (B and C). Meanwhile, the real condition shows that D has a larger area surrounding A and shares a longer common borderline with A than B and C with A. It indicates that D dominates the area around A and makes a greater contribution to A. Therefore, D must have a stronger relationship with A and should be chosen to represent the neighbor's condition of A. It is based on the first law of geography that everything is related to everything else, but nearby things are more related than distant things (Tobler, 1970). In this way, A is detected as an outlier because it is significantly different from its neighbors (D). From this illustration, it is necessary to give different weights to the objects with different spatial characteristics in the calculation of the neighborhood score, as in line with Colak, Memisoglu, Erbas, and Bediroglu (2018), Taha et al. (2019), and Ulak, Ozguven, Vanli, and Horner (2019).

According to Kou, Lu, and Chen (2006), outlier detection methods accommodate different characteristics. These methods are the weighted Z value approach (weighted Z) and the Average Difference Algorithm (AvgDiff). Both methods assign different weights for different neighbors in computing the Degree of Oulierness (DO) of the central object. The closer/stronger the spatial relationship between objects is, the greater the weight is given to the objects. For example, Fitriani, Pusdiktasari, and Diartho (2021) applied AvgDiff to identify the spatial outlying East Java regency/municipality in terms of economic growth. In their study, the use of AvgDiff was in accordance with the conditions of East Java regencies/municipalities with different characteristics. However, the methods still have several drawbacks. Weighted Z and AvgDiff have the potential for false detection when neighbors with extreme values have small weights. The non-robustness of the mean (average) used in those two methods is the cause. Furthermore, AvgDiff does not rely on a statistical test to determine the outliers. It is based on the rank of the DO, namely a concept of the top *m* outliers. Statistical tests cannot be used because the results of the difference between the value of the central object and its neighbors are in absolute values, so it does not follow the normal distribution. In the concept of top *m* outliers, a researcher is asked to determine the number of outliers (*m*). It is a hard-to-answer question because the researchers never know how many outliers exist in the data set (Su, 2011). Therefore, it is necessary to develop the outlier detection method further to improve the drawbacks of the median algorithm and AvgDiff. With the improvement, it expects a more accurate outlier detection method that can accommodate the real conditions in applied cases. It is important to do, given that outlier detection can be applied in wide cases. It is not only applied in cases related to geospatial on the earth's surface but can also be in objects that can be defined spatially. For example, outlier detection methods can be applied in the medical field to detect the position of cells that behave unusually, such as tumor/cancer cells, from other cells in the vicinity (Goovaerts, 2005; Ijaz, Attique, & Son, 2020; Prastawa, Bullitt, Ho, & Gerig, 2004). If the method results in false detection, it will certainly be a serious problem.

Based on the explanation, the research develops an improved version of the median algorithm and AvgDiff, called Weighted Median Algorithm (WMA). WMA combines the advantages of the two methods. From Median Algorithm, WMA adopts the median and statistical test concepts. From AvgDiff, WMA adopts the concept of using differences in objects' spatial characteristics as weights. The advantages of the two methods are combined by utilizing the weighted median in the calculation of neighborhood score $(g(x))$ and using a robust statistical test to determine the status of objects. Combining both advantages will improve the drawbacks of the median algorithm, which does not accommodate spatial characteristics differences. It will also improve the drawbacks of AvgDiff, which is not robust, and the determination of outliers that are only based on ranks. Two scenarios are used in the research. Scenario 1 is used to analyze the performance of WMA in detecting outliers in data with different spatial characteristics of objects. Then, scenario 2 analyzes the performance of WMA in detecting outliers without asking the researchers to determine the right *m*. The simulation is conducted 10.000 times to measure the accuracy of the methods.

II. METHODS

The research outlines the proposed algorithm, WMA. The algorithm is a systematic practical method to do a computation, in this case, detecting an outlier. The algorithm consists of inputs, effective step-by-step methods, and output. WMA is an improved version of the median algorithm by Chen et al. (2008) and AvgDiff by Kou et al. (2006). The improvement is made by adding different objects' spatial characteristics as weights to the median algorithm. It is done by changing the median into the weighted median in calculating the neighborhood score $(g(x))$. It also uses statistical tests in determining outliers as a substitute for top *m* outliers in AvgDiff. The inputs are as follows. It shows *S* as a set of spatial objects $\{s_1, s_2, \dots, s_n\}$, *k* as the number of the nearest neighbors of a central object. The neighbors are determined based on a spatial configuration, notated by $NN_k(i)$, and *X* as a set of attribute values $\{x_1, x_2, \dots, x_n\}$, in which x_1 is the attribute value of spatial object *i*.

Meanwhile, $weight$ is the value that measures the spatial relationship of the neighbors to the central object based on spatial characteristics information. Several characteristics can be used to define this relationship, such as the inverse distance, the length of the shared border, and the area. The concept gives larger weights to objects whose spatial relationship is stronger (Taha et al., 2019). The weights must satisfy the following condition

$$\sum_{j=1}^{k} weight_j = 1, j\epsilon NN_k(i) \tag{1}$$

Those notations are used to define the following steps for WMA. First, *k* nearest neighbors for each object *i* are defined using spatial relationships, such as Queen Contiguity, Rook Contiguity, and others. Then, spatial weights are calculated based on spatial information (length of the shared border ($border_j$)) between the central object *i* and each of its neighbors using Equation (2).

$$weight_j = \frac{border_j}{\sum_{j=1}^{k} border_j} \tag{2}$$

The weights are used to calculate the weighted median of the nearest neighbors. A weighted median is the 50% weighted percentile. The *k*-ordered nearest neighbors attribute values of $x_1, x_2, \dots, x_k$ with weights $weight_1, weight_2, \dots, weight_k$, x_i will be the weighted median if it satisfies the following condition (Edgeworth, 1887). If an ordered nearest neighbor (x_i) has a total weight of its previous neighbors of less than or equal to ½, and the total weights thereafter is less than equal to ½, the x_i is the weighted median.

$$\sum_{j=1}^{l-1} weight_j \leq \frac{1}{2} \quad and \quad \sum_{j=l+1}^{k} weight_j \leq \frac{1}{2} \tag{3}$$

The weighted median is called the neighborhood score $(g(x_i))$. This neighborhood score represents the condition of neighbors of a central object. It can be defined as follows. The neighborhood score $(g(x_i))$ is selected from the attribute values of weight ordered neighbors which satisfy the condition in Equation (3).

$$g(x_i) = \underset{j\epsilon NN_k(x_i)}{weighted\_med\{x_j, weight_j\}} \tag{4}$$

Then, the comparison score is calculated. The score measures the differences between a central object (x_i) and its neighborhood score $g(x_i)$. The comparison score indicates how big the difference of a central object and its neighbors. If the comparison score is extremely high, it shows that the central object has a very different characteristic from its surrounding and vice versa. The calculation uses Equation (5).

$$h_i = x_i - g(x_i) \tag{5}$$

Finally, the status of objects (normal/outlier) is defined using the statistical test in Equation (6). The μ^* and σ^* denote the robust mean and standard deviation, respectively. According to Chen et al. (2008) and Kolbaşi and Ünsal (2019), the median is used as μ^*, and Median Absolute Deviation (MAD) as σ^* of H set $\{h_1, h_2, \dots, h_n\}$.

$$out(\alpha, \mu^*, \sigma^*) = \left\{ h_i : \frac{|h_i - \mu^*|}{\sigma^*} > 3 \right\} \tag{6}$$

In the research, simulations are conducted. Generated data with predetermined outliers are used so their existence can be traced. The situation is hardly met when real data are used. The reason for using generated data is for the ease of evaluating the performance of the outlier detection methods (Ernst & Haesbroeck, 2017).

Figure 3 shows the spatial configuration of the objects. Numbers in Figure 3 show the ordered code for the objects. The values (attributes) for these spatial objects are generated based on the spatial lag model with a high positive autocorrelation value. It ensures that all spatial objects are normal objects. Normal object in spatial data has attribute values that tend to be similar or do not extremely differ from the attribute values of their neighboring objects. The spatial lag model is used because it is assumed that there is a relationship between the attribute value of an object and the attribute value of its surrounding objects. High positive spatial autocorrelation is used because it is in line with normal object definitions. A parameter (ρ) is used to measure the degree of spatial autocorrelation. The high value of ρ (close to 1) indicates strong positive spatial autocorrelation, so clusters of nearby spatial objects with similar attribute values will be formed.

The model used to generate the attribute values of the 36 spatial objects is shown in Equation (7). It has ρ = 0,9, indicating strong positive spatial autocorrelation. Strong autocorrelation shows that all the objects have strong relationships and affect one another. That is why they tend to have the same characteristics. In this condition, it can be easily controlled that there is no outlier besides the defined or generated ones

$$\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}((\boldsymbol{I} - \rho\boldsymbol{W})^{-1})') \tag{7}$$

Then, the spatial outlier can be created by selecting a particular normal object and replacing its attribute value with a more extreme value. The spatial outlier attribute value is generated as follows. For extremely high spatial outlier, it is shown in Equation (8). Meanwhile, the extremely low spatial outlier is in Equation (9).

$$X_{OS} \sim N(\mu_{NN} + 6\sigma, \sigma^2) \tag{8}$$

$$X_{OS} \sim N(\mu_{NN} - 6\sigma, \sigma^2) \tag{9}$$

Extremely high spatial outliers are the type of spatial objects with very high attribute values compared to their nearest neighbors. Meanwhile, the spatial outlier of the extremely low type is a spatial object with a very low attribute value compared to its nearest neighbors.

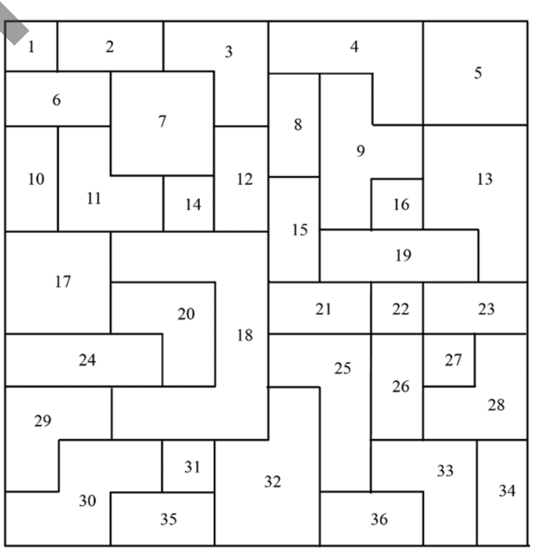

Figure 3 Spatial Configuration of Spatial Objects for Data Generation

The simulation is done with two scenarios. The first scenario (scenario 1) is applied to the median algorithm and WMA. The scenario is set so there are outliers in a data set with different spatial characteristics among the objects. Figure 4 illustrates a situation where the objects have various areas and lengths of shared borders.

In the research, the length of the shared border is used as the weight. The unit of length is explained through some illustrations in Figure 4. It shows the subsets of the spatial configuration depicted in Figure 3. Figure 4(a) shows objects 1 and 2 sharing 1 unit border length. Then, Figure 4(b) shows objects 10 and 11 sharing 2 units of border length. Meanwhile, Figure 4(c) shows objects 9 and 16 sharing 2 units of border length. The method used to define a neighboring object in the research is Rook Contiguity.

In this scenario, objects 18 and 7 in Figure 3 are defined as outliers. In outlier detection methods that do not accommodate spatial characteristics, object 18 has the same contribution as objects 17 and 24 to the central object (object 20). However, object 18 has a really long shared border with object 20 compared to objects 17 and 24 to object 20. It indicates that object 18 dominates the neighborhood of object 20, which makes it a good choice to represent the neighbors. Hence, it is good to choose it to represent the neighbors. These characteristics make object 20 an outlier because it has an extremely different value from its neighbors. The simulation gives a 'TRUE' result if the method can correctly detect objects 18, 7, and 20 as outliers. The simulation is conducted 10.000 times to calculate the accuracy of the methods. Each simulation generates normal object values first using Equation (7) and outlier values for objects 18 and 7 using Equations (8) and (9).

Scenario 2 is applied to AvgDiff and WMA. The scenario is formed with the same spatial configuration (Figure 3) and weights as scenario 1 but is focused on analyzing the performance of WMA in detecting outliers without determining the exact *m*. Each simulation generates normal object values first using Equation (7), defines two objects randomly as outliers, and generates outlier values for the two selected objects using Equations (8) and (9). The simulation will give a 'TRUE' result if the method correctly detects the two objects defined before as outliers. The simulation is conducted 10.000 times to calculate the accuracy of the method. According to its algorithm, the determination of outliers in AvgDiff uses top *m* outliers. Based on this concept, with 36 objects, AvgDiff detects 2 ($m = 5\% \times n = 1.8 \approx 2$) objects with the highest DO as outliers. Meanwhile, in WMA, the determination of outliers is based on statistical tests using Equation (6).

The accuracy is calculated using Equation (10). The higher the accuracy of the method is, the more precise the method is in detecting spatial outliers that the researchers have predetermined. The improvement of spatial outlier detection methods is shown in Figure 5. The base of the spatial outlier detection

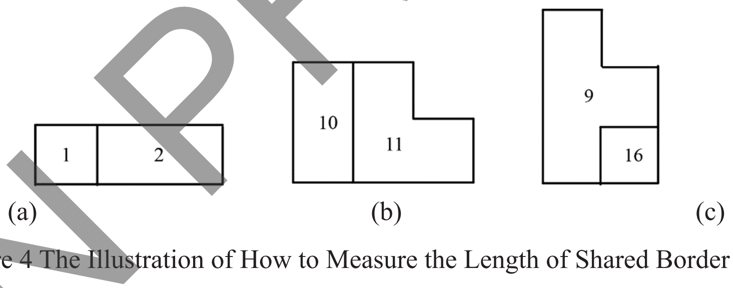

Figure 4 The Illustration of How to Measure the Length of Shared Border

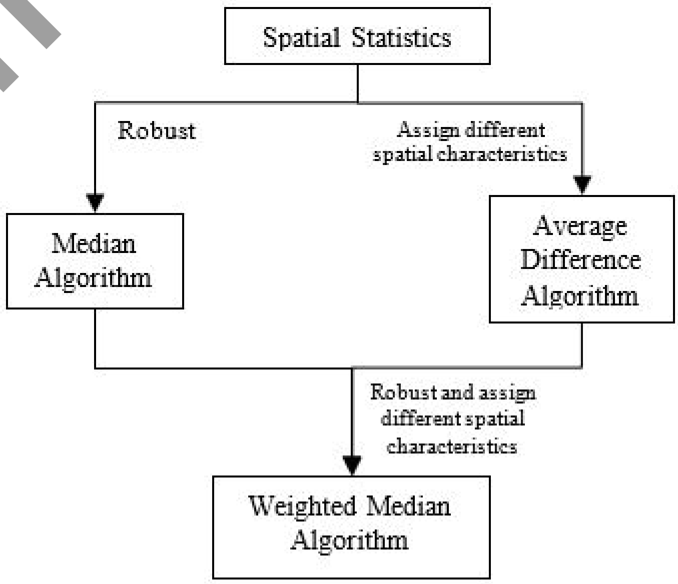
Spatial Statistics
Robust
Assign different spatial characteristics
Median Algorithm
Average Difference Algorithm
Robust and assign different spatial characteristics
Weighted Median Algorithm

Figure 5 The Improved Spatial Outlier Detection Methods Flowchart

methods improvement in this research is Spatial Statistics. The method then improved to be a robust outlier detection method in Median Algorithm, but it does not assign different spatial characteristics. Spatial Statistics is also improved by assigning different spatial characteristics in Average Difference Algorithm, but it is not robust. Therefore, in the research the advantages of those two methods are combined by improving a Weighted Median Algorithm. The method is robust and assign different spatial characteristics.

$$accuracy = \frac{\#TRUE}{10.000} \times 100\% \tag{10}$$

III. RESULTS AND DISCUSSIONS

The data are generated for 36 normal objects before being adjusted to the settings in scenarios 1 and 2. The data generation is done according to Equation (7). Each generation produces 36 attribute values for 36 objects. Moran's I test is conducted to ensure that the 36 objects are normal objects. Table 1 shows the results of Moran's I test for a one-time generation.

Table 1 The Result of Moran's I Test in the Generated Data

Moran's I Statistic	0,44202157
p-value	1,587e-06

Based on the results of Moran's I test in Table 1, the *p*-value is not greater than α = 5%, leading to the rejection of the null hypothesis. This rejection indicates that the generated attribute values have a strong spatial autocorrelation. Objects with strong spatial autocorrelation show that they have attribute values that tend to be the same as the attribute values of their nearest neighbors. This statement is also the definition of a normal object. Thus, it can be guaranteed that all objects with generated attribute values are normal objects.

From 36 normal objects, objects 18 and 7 are selected as spatial outliers. This selection is based on the characteristics of objects 18 and 7, which emphasize the importance of weights. Object 18 has 13 nearest neighbors, and object 20 is one of them. Object 20 shares 5 units of length common border with object 18. Meanwhile, it only shares 1 unit with object 17 and 2 units with object 24. Object 20 has values that are similar to objects 17 and 24. However, the area around object 20 is dominated by object 18 with extreme values. So, object 18 represents the neighbors of object 20 well. This condition makes object 20 an outlier because object 20 differs from its neighbors (object 18). Although the generated outliers are objects 18 and 7, object 20 is also an outlier.

Object 7 is selected as an outlier to show that the neighbors with the greatest weights do not always dominate. Object 3 is one of the neighbors of object 7. Both share a common border of 2 units. Other objects that are neighbors of object 3 only share 1 unit, namely objects 2, 4, 8, and 12. However, object 7, with extreme value, is not considered dominating. The reason is that objects 2, 4, 8, and 12 with similar values have 4 of 6 units border of object 3. So, those objects are a better representation of the neighbors of object 3. Because object 3 has similar values to the values of its neighbors, it will not be considered an outlier.

This scenario ensures that the true outliers are objects 18, 7, and 20. Therefore, a good outlier detection method can detect these three objects as outliers. The results of running the median algorithm and WMA for one-time simulation are presented in Table 2. The simulation is conducted one time so that the DO object can be analyzed. Although there are 36 objects, only 10 objects with the highest DO are presented in Table 2 for simplification.

Table 2 Comparison of the Results of Median Algorithm and WMA Based on Scenario 1

Rank	Object Index	DO of Median Algorithm	Object Index	DO of WMA
1	18	6,65746525	18	6,90930272
2	7	5,46865086	20	6,84920951
3	12	2,14256891	7	4,97271731
4	29	1,58279921	12	2,03841279
5	9	1,56876925	11	1,68248516
6	24	1,52384865	10	1,66903583
7	2	1,51474089	9	1,65483052
8	10	1,44017602	3	1,54885561
9	3	1,30381040	29	1,48769293
10	17	1,13747811	2	1,42073491

The results of outlier detection in Table 2 show that objects with DO values more than 3 in the median algorithm are objects 18 and 7. These two objects are indeed the two predetermined outliers. However, it is unable to detect object 20 as an outlier. The object is not even included in the top 10 objects with the highest potential as outliers. On the contrary, WMA detects those objects 18, 20, and 7 as having more than 3 DO values, which indicates those objects as outliers. This result indicates that WMA performs better than the median algorithm in detecting outliers in data with heterogeneous characteristics of objects. Simulations are conducted 10.000 times to measure the accuracy of the methods.

The determination of object status (normal/ outlier) is only based on its rank (top *m* outlier) and does not use statistical tests. It is due to the use of absolute differences in its algorithm, resulting in values that do not follow a normal distribution. This concept is computationally advantageous because the calculations become faster. However, this method requires the researchers' knowledge regarding the number of outliers (*m*) in the data. This information is rarely unknown.

WMA is developed to overcome this weakness of AvgDiff. The improvement is made while keeping the advantages of AvgDiff in accommodating the spatial characteristics to the calculation of DO. Table 3 presents the results of running the two methods based on scenario 2 for a one-time simulation. In this one-time simulation, objects 18 and 32 are the randomly selected outliers. However, in line with the explanation in scenario 1, object 20 is also an outlier because it is the neighbor of object 18.

Based on Table 3, AvgDiff successfully detects the 3 predetermined outliers as the objects with the highest DO. However, based on the concept of top *m* outliers in AvgDiff, *m* is calculated by the formula $m = 5\% \times n$. With $n = 36$, it is $m = 5\% \times 36 = 1{,}8 \approx 2$. If the researchers do not have a priori information about the number of outliers in the data, AvgDiff will only identify 2 objects with the highest DO as outliers, while, in fact, there are 3 outliers. The method fails to detect object 20 as an outlier.

WMA can identify the number of outliers and which objects are the outliers. WMA detects objects 18, 32, and 20 as the 3 objects with the highest DO without prior determination about the number of outliers in the data. From the results in Table 3, the DO values of objects 18, 32, and 20, which exceed 3, indicate that WMA successfully detects those predetermined objects as outliers. It confirms that WMA performs better than AvgDiff. Both produce the same detection results, but WMA can automatically determine outliers. Meanwhile, AvgDiff requires a predetermined correct number of outliers (*m*). Based on Table 4, which shows the simulation results, the WMA overperforms the median algorithm with an accuracy of 80,45%. Meanwhile, the accuracy of the median algorithm is 0,81%.

Table 3 Comparison of AvgDiff and WMA Based on Scenario 2

Rank	Object Index	DO of AvgDiff	Object Index	DO of WMA
1	18	8,7768072	20	9,65291968
2	32	6,8198899	18	9,59576428
3	20	5,8187378	32	5,43849870
4	14	4,0480307	16	2,67597554
5	31	3,5675264	14	2,43632067
6	15	3,0593516	15	2,40146974
7	25	3,0189807	33	2,36647726
8	36	2,9466585	6	1,77197493
9	17	2,8080744	17	1,49408098
10	12	2,7121336	31	1,26290926

Table 4 The Accuracy of the Methods Based on Two Simulation Scenarios

Scenario	Methods	
	Median Algorithm	**WMA**
1	0,81%	80,45%
	AvgDiff	**WMA**
2	87,25%	84,36%

Simulations are conducted 10.000 times to measure the accuracy of both methods. Based on the results in Table 4, AvgDiff's accuracy is 87,25% while WMA is 84,36%. Although the accuracy of WMA is not higher than AvgDiff, WMA can avoid a more serious problem when researchers cannot determine the exact *m*.

The WMA is the improved version of the median algorithm, which is weighted based on the spatial characteristics of the objects. When WMA is applied to objects with homogeneous characteristics, the same weight for all objects will produce DO, which is exactly the same as the DO of the median algorithm. The median algorithm and WMA are applied to data with homogeneous spatial characteristics objects to show these conditions. The detection results can be seen in Table 5.

The robustness of the method is also confirmed by Chen *et al*. (2008) by developing median algorithm to overcome the drawback of spatial statistics with masking and swamping effects. The research results confirm that median algorithm is a robust method which can overcome the effects of masking and swamping. The result is in line with the results of research by Su (2011), and the validity of the method is also confirmed by Wang, Wang, Hong, and Wan (2004). These three previous studies show the good performance of the median algorithm for homogeneous spatial characteristics objects. Thus, when the object has homogeneous characteristics, WMA is as good and robust as a median algorithm. However, when the objects have heterogeneous spatial characteristics, WMA performs better than a median algorithm.

On March 11th, 2020, the World Health Organization (WHO) declared Covid-19 as a global pandemic. In Indonesia, as of December 20th, 2021, this virus has caused more than 4,6 million confirmed infection cases and around 144.000 confirmed deaths (Mathieu et al., 2020) . East Java is one of the provinces with the highest number of cases of Covid-19. Although some cases are still found in several areas, the case rate is low enough, and the recovery rate has increased to 4,1 million (Satuan Tugas Penanganan COVID-19, 2021). However, the government should not be off guard. The wave of the spread of Covid-19 can happen again at any time. In addition, new variants of Covid-19 which are the result of mutations continuously appear, such as the Alpha, Beta, Delta, and Gamma I variants (Duong, 2021) to the newest one, Omicron (WHO, 2021) Therefore, studies related to this pandemic need to be continued. They can be used as a consideration in determining prevention and efforts to overcome the spread of Covid-19 in the future. If the detected outliers are extremely high (areas with high values surrounded by areas with a tendency to low values), the government is suggested to make a particular strategy to reduce the transmission rate or the number of Covid-19 cases in that area. If the detected outliers are extremely low (areas with low values surrounded by areas with high values), the outliers can be considered a pilot area in suppressing the number of Covid-19 cases to be applied in other areas.

The model is applied to determine the unusual behavior based on the number of confirmed Covid-19 cases. Areas with significantly different behavior are considered outliers. The data are the incidence rate (transmission) which is the ratio of the accumulated number of positive confirmed cases of Covid-19 to the population in the city/regency in East Java from March 2020 until December 2021. The data are taken from the covid19.go.id. Then, the neighboring method used is Queen Contiguity, considering the irregular shape of the object (areas). Spatial information as weight is Euclidean distance, based on the CDC statement (2020) that Covid-19 can spread through air contaminated with droplets and small airborne particles containing the virus. Table 6 shows the top ten objects (areas) that have the most risk of becoming spatial outliers.

Table 5 The Results of Median Algorithm and WMA in Detecting Outliers in Data with Homogeneous Characteristics Objects

Rank	Object Index	DO of AvgDiff	Object Index	DO of WMA
1	13	4,78593827	13	4,78593827
2	23	4,13793846	23	4,13793846
3	27	1,73384047	27	1,73384047
4	25	1,64925227	25	1,64925227
5	17	1,55829115	17	1,55829115
6	9	1,35445116	9	1,35445116
7	20	1,09954773	20	1,09954773
8	24	1,00967065	24	1,00967065
9	29	0,98937845	29	0,98937845
10	26	0,94638824	26	0,94638824

The difference in the detection results between the median algorithm and WMA is because the median algorithm assumes that all objects (areas) have the same characteristics as outlined in the concept of contiguity (neighborhood). Based on the East Java map in Figure 6, it can be seen that Mojokerto Regency and Pasuruan Regency (with the tendency of low values) dominate the neighboring Sidoarjo Regency. It causes the median algorithm to choose areas with low Covid-19 cases to represent neighboring Sidoarjo Regency, resulting in Sidoarjo Regency with high Covid-19 cases being detected as outliers. However, city/district areas in East Java have different spatial characteristics, and in Covid-19 cases, distance is important, not only contiguity. So, the median algorithm is unable to accommodate this case properly.

WMA only detects Surabaya City as an outlier. The WMA result states that Sidoarjo Regency is not an outlier considering differences in characteristics, especially the distance between regions as spatial information. The distance of Sidoarjo Regency, which is very close to Surabaya City, is indicated by a fairly large weighting value. This condition makes Surabaya City dominate the neighbor of Sidoarjo Regency. Hence, Sidoarjo Regency is not detected as an outlier because it has the same high value as its neighbors (Surabaya City). In other words, Sidoarjo Regency with a high Covid-19 incidence rate is normal because it is influenced by its neighbors, which also have a high Covid-19 incidence rate. With regard to the goodness of the method, WMA is more suitable to be used to detect outliers in the case of the incidence rate of Covid-19.

Apart from the differences in the three methods described, there is one thing in common. All three methods detect Surabaya City as an outlier. Some factors support this condition. Its geographical position, which is a coastal settlement, makes Surabaya have a high potential as a stopover and settlement for immigrants. In addition, its highly dense population and large port make Surabaya have a very big role in receiving and distributing industrial goods. Then, as a trade center, Surabaya City is the second largest metropolitan city after Jakarta. Malls and cafes in this city are the largest compared to other cities in East Java. These places are the main entertainment for the citizens of Surabaya City. Psychologically, it is not easy for its citizen to refrain from gathering and spending time outside their homes. So, their mobility is still high as it is not easy for them to stay home. It causes a high incidence rate as well. However, those complex characteristics of Surabaya City cannot be found in other cities around it. Although currently in Indonesia, the incidence rate of Covid-19 has decreased, this finding can be used as a consideration for the government in the future if a similar case occurs. The government is advised to pay attention and focus on prevention and action for Surabaya, as it is a center of mobility and the entrance and exit of East Java.

Table 6 The Results of Outlier Detection on Covid-19 Incidence Rate in 38 Cities/Regencies in East Java

1	2	3	2	4	2	5
1	37	12,505	37	0,0439	37	8,124
2	15	3,842	15	0,0231	25	2,355
3	30	2,461	25	0,0209	15	2,241
4	9	2,279	30	0,0094	30	2,032
5	10	1,846	9	0,0069	9	1,704
6	7	1,6227	14	0,0067	24	1,222
7	8	1,3171	6	0,0058	10	1,174
8	34	1,2410	16	0,0054	8	1,167
9	38	1,1638	38	0,0051	38	1,109
10	2	1,0450	10	0,0051	6	1,068

Note: 1 = Rank; 2 = Object Index; 3 = DO of Median Algorithm; 4 = DO of AvgDiff; 5 = DO of WMA

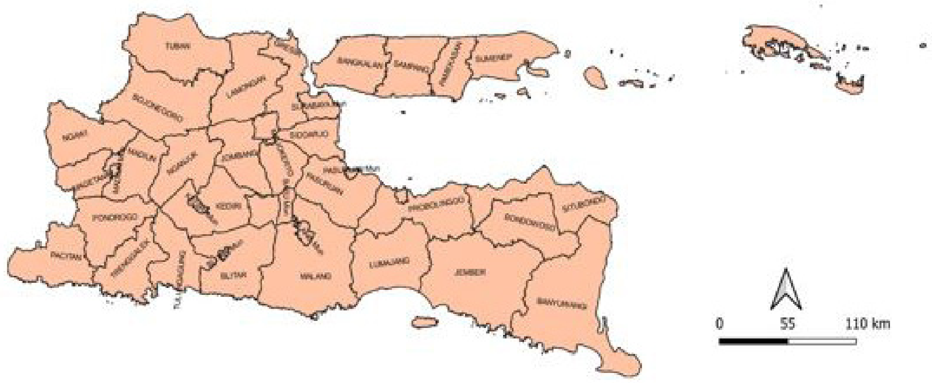

Figure 6 The Map of Cities/Regencies in East Java

IV. CONCLUSIONS

In the research, a method for detecting spatial outliers is developed, namely WMA. It is an improved version of the median algorithm and AvgDiff. WMA is confirmed to be as good and robust as the median algorithm when applied to objects with homogeneous spatial characteristics. When the objects have heterogeneous spatial characteristics, WMA performs better than a median algorithm. Even though the accuracy of WMA is not much higher than AvgDiff, the use of WMA can prevent a serious false detection problem when there is no prior information about the true number of outliers. With this concept, if it is applied to data on Covid-19 cases in cities/districts in East Java, it is possible to provide detection results. The Surabaya City and Sidoarjo Regency are a group of outliers with extreme values compared to other areas around them.

Currently, WMA focuses only on univariate nonspatial attributes with one information for the weights. In the future, researchers can improve it. So, it is suitable for data with multivariate nonspatial attributes and a combination of spatial information for the weights. WMA can also be improved to detect outliers in groups.

REFERENCES

Aggarwal, V., Gupta, V., Singh, P., Sharma, K., & 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). IEEE. https://doi.org/10.1109/ICOEI.2019.8862582

Araki, S., Shimadera, H., Yamamoto, K., & Kondo, A. (2017). Effect of spatial outliers on the regression modelling of air pollutant concentrations: A case study in Japan. *Atmospheric Environment*, *153*(March), 83–93. https://doi.org/10.1016/j.atmosenv.2016.12.057

Ayadi, A., Ghorbel, O., Obeid, A. M., & Abid, M. (2017). Outlier detection approaches for wireless sensor networks: A survey. *Computer Networks, 129*, 319–333. https://doi.org/10.1016/j.comnet.2017.10.007

Baba, A. M., Midi, H., & Abd Rahman, N. H. (2021). A spatial outlier detection method for big data based on adjacency weighted residuals and its application to COVID-19 data. *Economic Computation and Economic Cybernetics Studies and Research*, *55*(3), 87–102. https://doi.org/10.24818/18423264/55.3.21.06

Bosman, H. H., Iacca, G., Tejada, A., Wörtche, H. J., & Liotta, A. (2017). Spatial anomaly detection in sensor networks using neighborhood information. *Information Fusion*, *33*(January), 41–56. https://doi.org/10.1016/j.inffus.2016.04.007

Chen, D., Lu, C. T., Kou, Y., & Chen, F. (2008). On detecting spatial outliers. *GeoInformatica*, *12*, 455–475. https://doi.org/10.1007/s10707-007-0038-8

Colak, H. E., Memisoglu, T., Erbas, Y. S., & Bediroglu, S. (2018). Hot spot analysis based on network spatial weights to determine spatial statistics of traffic accidents in Rize, Turkey. *Arabian Journal of Geosciences, 11*, 1–11. https://doi.org/10.1007/s12517-018-3492-8

Djenouri, Y., & Zimek, A. (2018). Outlier detection in urban traffic data. In *WIMS '18: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics* (pp. 1–12). https://doi.org/10.1145/3227609.3227692

Duong, D. (2021). Alpha, Beta, Delta, Gamma: What's important to know about SARS-CoV-2 variants of concern? *CMAJ: Canadian Medical Association Journal*, *193*(27), E1059–E1060. https://doi.org/10.1503/cmaj.1095949

Edgeworth, F. Y. (1887). On observations relating to several quantities. *Hermathena*, *6*(13), 279–285.

Ernst, M., & Haesbroeck, G. (2017). Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery*, *31*, 371–399. https://doi.org/10.1007/s10618-016-0471-0

Fitriani, R., Pusdiktasari, Z. F., & Diartho, H. C. (2021). Growth interdependence in the presence of spatial outliers: Implementation of an average difference algorithm on East Java regional economic growth, 2011-2016. *Regional Statistics*, *11*(3), 119–132. https://doi.org/10.15196/RS110306

Fu, W., Zhao, K., Zhang, C., Wu, J., & Tunney, H. (2016). Outlier identification of soil phosphorus and its implication for spatial structure modeling. *Precision Agriculture*, *17*, 121–135. https://doi.org/10.1007/s11119-015-9411-z

Goovaerts, P. (2005). Detection of spatial clusters and outliers in cancer rates using geostatistical filters and spatial neutral models. In *Proceedings of the Fifth European Conference on Geostatistics for Environmental Applications* (pp. 149–160). https://doi.org/10.1007/3-540-26535-X_13

Helwig, Z. D., Guggenberger, J., Elmore, A. C., & Uetrecht, R. (2019). Development of a variogram procedure to identify spatial outliers using a supplemental digital elevation model. *Journal of Hydrology X*, *3*(April), 1–11. https://doi.org/10.1016/j.hydroa.2019.100029

Ijaz, M. F., Attique, M., & Son, Y. (2020). Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors*, *20*(10), 1–22. https://doi.org/10.3390/s20102809

Khan, D., Rossen, L. M., Hamilton, B. E., He, Y., Wei, R., & Dienes, E. (2017). Hot spots, cluster detection and spatial outlier analysis of teen birth rates in the U.S., 2003–2012. *Spatial and Spatio-Temporal Epidemiology*, *21*(June), 67–75. https://doi.org/10.1016/j.sste.2017.03.002

Kolbaşi, A., & Ünsal, A. (2019). A comparison of the outlier detecting methods: An application on Turkish foreign trade data. *Journal of Mathematics and Statistical Science*, *5*, 213–234.

Kou, Y., Lu, C. T., & 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.

Lu, C. T., Chen, D., & Kou, Y. (2003). Algorithms for spatial outlier detection. In *Third IEEE International Conference on Data Mining* (pp. 597–600). IEEE. https://doi.org/10.1109/ICDM.2003.1250986

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Gavrilov, D., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020). *Coronavirus (COVID-19) cases*. Retrieved from https://ourworldindata.org/covid-cases

Nguyen, T. T., Vu, D. T., Trinh, L. H., & Nguyen, T. L. H. (2016). Spatial cluster and outlier identification of geochemical association of elements: A case study in Juirui copper mining area. *Bulletin of the Mineral Research and Exploration, 153*(153), 159–167.

Prastawa, M., Bullitt, E., Ho, S., & Gerig, G. (2004). A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis*, *8*(3), 275–283. https://doi.org/10.1016/j.media.2004.06.007

Pu, J., Wang, Y., Liu, X., & Zhang, X. (2019). STLP-OD: Spatial and temporal label propagation for traffic outlier detection. *IEEE Access*, *7*, 63036–63044. https://doi.org/10.1109/ACCESS.2019.2916853

Sajana, O. K., & Sajesh, T. A. (2018). Detection of multidimensional outlier using multivariate spatial median. *Journal of Computer and Mathematical Sciences*, *9*(12), 1875–1881.

Satuan Tugas Penanganan COVID-19. (2021). *Penanganan COVID-19 2021: Kesembuhan melebihi 4,1 juta, kasus aktif tersisa 4 ribu dan vaksinasi melampaui 161 juta orang.* Retrieved from https://covid19.go.id/p/berita/penanganan-covid-19-2021-kesembuhan-melebihi-41-juta-kasus-aktif-tersisa-4-ribu-dan-vaksinasi-melampaui-161-juta-orang

Shekhar, S., Lu, C. T., & Zhang, P. (2001). *A unified approach to spatial outlier detection.* Retrieved from https://hdl.handle.net/11299/215495

Shukla, S., & Lalitha, S. (2021). Spatial analysis of water quality data using multivariate spatial outlier detection algorithms. *GANITA, 70*(2), 87–96.

Su, P. C. (2011). *Statistical geocomputing: Spatial outlier detection in precision agriculture* (Master's thesis). University of Waterloo.

Taha, A., Onsi, H. M., El Din, M. N., & Hegazy, O. M. (2019). A model for spatial outlier detection based on weighted neighborhood relationship. *arXiv Preprint,* 1–12. https://doi.org/10.48550/arXiv.1911.01867

Tang, J., & Ngan, H. Y. T. (2016). Traffic outlier detection by density-based bounded local outlier factors. *Information Technology in Industry, 4*(1), 6–18.

Tepanosyan, G., Sahakyan, L., Zhang, C., & Saghatelyan, A. (2019). The application of Local Moran's I to identify spatial clusters and hot spots of Pb, Mo and Ti in urban soils of Yerevan. *Applied Geochemistry*, *104*(May), 116–123. https://doi.org/10.1016/j.apgeochem.2019.03.022

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240. https://doi.org/10.2307/143141

Ulak, M. B., Ozguven, E. E., Vanli, O. A., & Horner, M. W. (2019). Exploring alternative spatial weights to detect crash hotspots. *Computers, Environment and Urban Systems, 78*(November), 1–9. https://doi.org/10.1016/j.compenvurbsys.2019.101398

Van Zoest, V. M., Stein, A., & Hoek, G. (2018). Outlier detection in urban air quality sensor networks. *Water, Air, & Soil Pollution, 229*, 1–13. https://doi.org/10.1007/s11270-018-3756-7

Wang, S., & Serfling, R. (2018). On masking and swamping robustness of leading nonparametric outlier identifiers for multivariate data. *Journal of Multivariate Analysis*, *166*(July), 32–49. https://doi.org/10.1016/j.jmva.2018.02.003

Wang, Z. Q., Wang, S. K., Hong, T., & Wan, X. H. (2004). A spatial outlier detection algorithm based multi-attributive correlation. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)* (pp. 1727–1732). IEEE. https://doi.org/10.1109/ICMLC.2004.1382054

WHO. (2021). *Informasi terbaru tentang Omicron*. Retrieved from https://www.who.int/indonesia/news/detail/30-11-2021-informasi-terbaru-tentang-omicron

Wu, H., Tang, X., & Wang, Z. (2018). Probabilistic automatic outlier detection for surface air quality measurements from the China national environmental monitoring network. *Advances in Atmospheric Sciences, 35*, 1522–1532. https://doi.org/10.1007/s00376-018-8067-9

Xia, H., An, W., Li, J., & Zhang, Z. (2022). Outlier knowledge management for extreme public health events: Understanding public opinions about COVID-19 based on microblog data. *Socio-Economic Planning Sciences*, *80*(March), 1–12. https://doi.org/10.1016/j.seps.2020.100941

Xiao, F., Wang, K., Hou, W., & Erten, O. (2020). Identifying geochemical anomaly through spatially anisotropic singularity mapping: A case study from silver-gold deposit in Pangxidong district, SE China. *Journal of Geochemical Exploration*, *210*(March), 1–20. https://doi.org/10.1016/j.gexplo.2019.106453