Sains Malaysiana 51(3)(2022): 911-927
http://doi.org/10.17576/jsm-2022-5103-24

Improved Spatial Outlier Detection Method Within a River Network

(Kaedah Pengesanan Pencilan Reruang DiPerbaik dalam Suatu Jaringan Sungai)

NUR FATIHAH MOHD ALI[1], ROSSITA MOHAMAD YUNUS[1,*], IBRAHIM MOHAMED[1] & FARIDAH OTHMAN[2]

[1]*Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Federal Territory, Malaysia*
[2]*Department of Civil Engineering, Faculty of Engineering, Universiti Malaya, 50603 Kuala Lumpur, Federal Territory, Malaysia*

Received: 1 February 2021/Accepted: 13 August 2021

ABSTRACT

A spatial outlier refers to the observation whose non-spatial attribute values are significantly different from those of its neighbors. Such observations can also be found in water quality data at monitoring stations within a river network. However, existing spatial outlier detection procedures based on distance measures such as the Euclidean distance between monitoring stations do not take into account the river network topology. In general, water quality levels in lower streams will be affected by the flow from the upper streams. Similarly, the water quality at some tributaries may have little influence on the other tributaries. Hence, a method for identifying spatial outliers in a river network, taking into account the effect of river flow connectivity on the determination of the neighbors of the monitoring stations, is proposed. While the robust Mahalalobis distance is used in both methods, the proposed method uses river distance instead of the Euclidean distance. The performance of the proposed method is shown to be superior using a synthetic river dataset through simulation. For illustration, we apply the proposed method on the water quality data from Sg. Klang Basin in 2016 provided by the Department of Environment, Malaysia. The finding provides a better identification of the water quality in some stations that significantly differ from their neighbouring stations. Such information is useful for the authorities in their planning of the environmental monitoring of water quality in the areas.

Keywords: Euclidean distance; river distance; robust multivariate; spatial outlier; water quality

ABSTRAK

Reruang terpencil merujuk kepada cerapan dengan nilai atribut reruang berbeza secara signifikan berbanding daripada nilai kejiranannya. Cerapan ini boleh dikesan daripada data kualiti air yang dikumpul di stesen-stesen dalam jaringan sungai. Walau bagaimanapun, kaedah semasa untuk mengenal pasti pencilan reruang menggunakan jarak yang diukur antara stesen seperti jarak Euclidean tidak mengambil kira aspek topologi jaringan sungai. Secara umumnya, aras kualiti air pada hilir jaringan sungai dipengaruhi oleh aliran daripada hulu sungai. Begitu juga, kualiti air pada sesuatu jaringan sungai mungkin mempengaruhi sedikit kualiti air pada jaringan sungai yang berbeza. Kaedah dalam mengenal pasti reruang terpencil dalam jaringan sungai dengan mengambil kira kesan terhadap hubung kait aliran sungai bagi menentukan kejiranan sesebuah stesen dicadangkan. Walaupun penganggar kukuh jarak Mahalanobis digunakan dalam kedua-dua kaedah, tetapi kaedah yang dicadangkan ini menggunakan jarak aliran sungai dan bukannya jarak Euclidean. Berpandukan kaedah simulasi set data sungai sintetik, prestasi kaedah yang diperkenalkan ini terbukti lebih baik. Sebagai ilustrasi, kaedah yang diperkenalkan ini diterapkan pada data kualiti air yang diperoleh daripada Sg. Klang pada tahun 2016 yang disediakan oleh Jabatan Alam Sekitar, Malaysia. Keputusan daripada hasil kajian dapat membantu mengenal pasti kualiti air di beberapa buah stesen yang jauh lebih baik daripada stesen berdekatan. Maklumat ini sangat berguna kepada pihak berwajib dalam merancang pemantauan kualiti air di kawasan sekitarnya.

Kata kunci: Jarak aliran sungai; jarak Euclidean; kualiti air; penganggar multivariat; reruang terpencil

INTRODUCTION

The occurrence of outlying observations in spatial data may lead to unexpected, exciting and implicit information. A spatial or local outlier is a term used to describe a spatial point different in non-spatial attributes from its neighbors (Cressie et al. 2006; Haslett 1992). Previous studies have

shown that the identification of spatial outliers in spatial data sets may lead to significant interpretations such as climate changes, tornadoes, and hurricanes (Kelleher & Braswell 2021). In medicine, a number of works can also be found in identifying abnormal shapes such as tumors and infected tissues in the medical images (Baur et al. 2021; Liu et al. 2017). In public health, an emergency call in an area may have a good response rate but differ significantly if compared to the surrounding areas. Therefore, these anomaly patterns should be considered to improve health care standards and patient survival rates (Azimi et al. 2021). Recently, spatial outlier detection helped to determine the unusual correlation between COVID-19 cases and crime rates in Chicago (Yang et al. 2021). The detection of spatial outliers needs to be performed locally in the neighborhood in order to accommodate the spatial dependence between the spatial objects. In general, a spatial outlier does not follow a common feature of spatial data, notably the spatial dependency with its neighboring points (Filzmoser et al. 2014).

In water quality data, the spatial point is essentially the river location while the non-spatial attributes of each spatial point explain the water quality characteristics in the area. In practice, we assume that a positive spatial autocorrelation occurs at adjacent river locations, which means a neighboring river location with high non-spatial values is surrounded by locations with high non-spatial values as well. However, as river water flows from the upper to lower stream, water quality at the lower stream may be affected by river water quality at the upper stream, but rarely otherwise. Similarly, there is no causal relationship between the water quality levels of any two different tributaries. Hydrologic pathways are routes along which water moves from when it is received as precipitation until it is delivered to the most downstream point in a river basin. Water quality degradation in upstream parts can negatively affect downstream throughout a watershed (Peters & Meybeck 2000). Hence, in this paper, we proposed a method to identify multivariate spatial outliers that take into account the river network topology.

The importance of studies involving water quality within a river network has been highlighted in the literature. For example, the level of dissolved oxygen in river water is more directly affected by upstream-downstream relations and the river network (Mainali & Chang 2021). According to the data studied by Lachhab et al. (2021), the physical and chemical changes to the streams from the dams are affecting the biological communities in the downstream watercourse. In addition, Hasib and Othman (2020) studied that the criteria for determining pollution sources are when the pollution sources should be upstream while the monitoring station is downstream of the river. Hence, the spatial outlier detection method has been adopted to identify the changes in the water and help the decision-makers evaluate the effects of specific water quality measures (Talagala et al. 2019; Zheng et al. 2017).

Graphical representations such as the variogram cloud (Cressie 1993) and the Moran scatterplot (Anselin 1995) are useful to identify spatial outliers for univariate data. In addition, several algorithms were proposed based on the idea of how close the non-spatial attributes of a spatial point to the point estimate of its neighboring non-spatial attributes using summary statistics, namely, mean, median, weighted mean and average difference (Chen et al. 2008; Kou 2006; Kou et al. 2006; Lu et al. 2003; Shekhar et al. 2003). It has also been shown that the weighted mean and weighted average difference algorithms that use weighted average with inverse distance as the weights performed better than the mean and median algorithms (Kou 2006; Kou et al. 2006; Peter 2011). For multivariate cases, these outlier detection algorithms were extended using the Mahalanobis distance to calculate the outlier scores of the multiple attributes, assuming that the scores follow the multivariate normal distribution. Likewise, Cai et al. (2009) also applied Mahalanobis distance on the multiple outlier scores after normalizing the non-spatial attributes. The location quotient (LQ) algorithm was recently proposed by Alok Kumar and Lalitha (2018), using the proportion of attributes in a neighborhood over the proportion of attributes in a larger reference neighborhood. The LQ algorithm performs better in the simulation study than the mean and median algorithms for multiple attributes data. With the increasing number of multiple attributes, outlier detection in multivariate data tends to suffer from the swamping and masking effects (Wang & Serfling 2018). Several robust estimation methods, such as the minimum covariance determinant (MCD) (Rousseeuw & Van Driessen 1999), have been recommended to deal with these problems (Sajesh & Srinivasan 2013; Wang & Serfling 2018). Filzmoser et al. (2014) introduced an exploratory tool to identify outliers in a local spatial neighborhood based on pairwise robust Mahalanobis distances between the observations. The robustness of the methods comes from both the robust mean and covariance estimates used in computing the Mahalanobis distances. The determination of the distribution of these pairwise distances results in measuring the local outlyingness of the observation. The local behavior of

the method has been regulated by changing the size of the neighborhood. Additional robustness for identifying local outliers is included by tolerating a small percentage of similar neighbors, which occurs just by chance. By increasing this percentage, the method can be used to find locally homogeneous regions. Despite these notable works on the methods of detecting spatial outlier, none of these methods considers river network topology in the formulation.

The nearest neighbor concept is important in spatial outlier detection as neighboring points' characteristics will impact the spatial outlier identifications. Euclidean distance is the most common distance metric used for the choice of neighbors in developing algorithms for spatial outlier detection (Filzmoser et al. 2014; Shekhar et al. 2003). Some works in modeling river data used Euclidean distance to describe the spatial autocorrelation for geostatistical water quality studies (Cressie et al. 2006; Peterson & Urquhart 2006; Tortorelli & Pickup 2006). However, such an approach is not suitable for river network data sets (Money et al. 2009a). Alternatively, the river distance, which is the shortest distance along the river, is proposed when studying the spatial autocorrelation among river monitoring sites. The river distance (Cressie et al. 2006; Money et al. 2009b) is also known as the hydrologic distance (Peterson et al. 2006) or stream distance (Ver Hoef et al. 2006). Both the river distance and the river flow direction between river points is considered as essential features to develop the correlation matrix structure for river data that represents autocorrelation amongst river points (Anselin 1995; Cressie et al. 2006; de Fouquet & Bernard-Michel 2006; Jat 2017; Money et al. 2011; Ver Hoef & Peterson 2010; Ver Hoef et al. 2006).

This paper aims to develop an improved spatial outlier detection method for multivariate spatial river data. We considered a pairwise robust Mahalanobis distance between river locations and a new k nearest neighbor method that incorporates both the river distance and flow of connectivity between river points to identify neighbors of a given location. In the next section, we discussed the river network concept and illustrated an example of multivariate spatial outliers for river data, and introduced the proposed method and its properties. Next, we presented the simulation results to investigate the performance of the proposed method. Then, we applied the method to an actual Malaysian water quality dataset, followed by a discussion and a conclusion.

Materials and Methods

Spatial outlier detection can be used to identify stations where the values of non-spatial attributes measured are different from its neighboring station. To improve the spatial outlier detection in a river network, a river flow distance measure is considered instead of the Euclidean distance. The determination of the neighboring stations using Euclidean distance might increase the error when detecting spatial outliers. Thus, identifying spatial outliers in a river network requires a river flow distance measure to determine the pairs of neighboring stations. Then, the Mahalanobis distance and the pairwise Mahalanobis distance with river flow distance are used to identify the spatial outlier in the river network.

River Network and Flow Direction

Notation for a river network used in Cressie et al. (2006) and Ver Hoef et al. (2006) is closely followed in this paper. The river network S in the Euclidean space denotes the collection of a finite number of river segments, each segment is depicted as a straight line as illustrated in Figure 1. Each segment or straight line is connected at the junctions of the river and whose union constitutes the river network (Peiman et al. 2015). The whole network has a single most-downstream segment that splits up into other segments as going upstream. Upstream endpoints are called sources, and the downstream endpoint is the outlet of the mouth of the river network. Any station or point on the river network can be connected by a continuous line to the lowest station in that network.

For a station or point $z_i \in S$ on the i^{th} segment, we let $D_{z_i} \subseteq S$ denote the index set of river segments downstream of z_i, including the i^{th} segment. In Figure 1, two stations z_1 and z_2 are said to be flow connected on a river network, written as $z_1 \rightarrow z_2$ when the river segments downstream are connected and are denoted as $D_{z_1} \cap D_{z_2} = D_{z_1}$ or D_{z_2}. In contrast, two stations z_2 and z_3 are flow unconnected, written as $z_2 \nrightarrow z_3$ when the river segments downstream are not connected, $D_{z_2} \cap D_{z_3} \neq D_{z_2}$ or D_{z_3}. If z_j is upstream of z_i, that is, $D_{z_i} \subset D_{z_j}$, then we denoted the set of segments between z_j and z_i, inclusive of the j^{th} but exclusive of the i^{th} segment, by $B_{z_i, z_j} = D_{z_j} \backslash D_{z_i}$. If z_j is downstream of z_i, then $B_{z_i, z_j} = D_{z_i} \backslash D_{z_j}$. In the case that z_i and z_j are on the same segment, that is, $D_{z_i} = D_{z_j}$, we have $B_{z_i, z_j} = \emptyset$. The river network and flow direction methodology will generate simulated data and calculate the river distances in the next section.

River Distance

The distance from a station to the lowest station in the river network is considered the river's arc length along the curve paths. We approximated the arc length by a

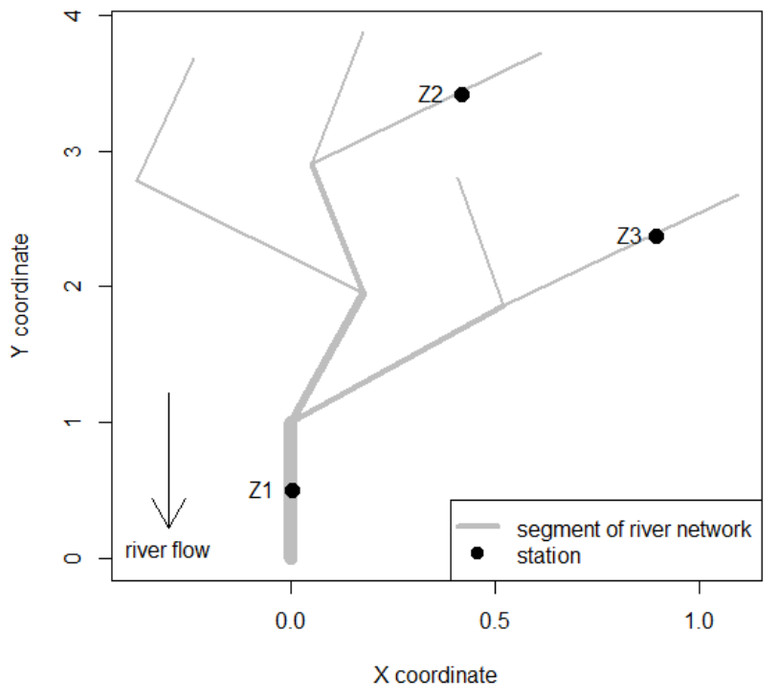

FIGURE 1. River network with three stations z_1, z_2 and z_3

succession of a significant enough number of chords. In other words, the river distance between two arbitrary stations z_i and z_j on the river network S is calculated by dividing the curve between the two stations into r segments, say z_{i1}, ..., $z_{ir\text{-}1}$ where $z_{i0} = z_i$ and $z_{ir\text{-}1} = z_j$. The distances between the successive stations are calculated using the Euclidean distance

$$d(z_{ir}, z_{ir-1}) = \sqrt{(X[z_{ir}] - X[z_{ir-1}])^2 + (Y[z_{ir}] - Y[z_{ir-1}])^2},$$

where X and Y are the X-coordinate and Y-coordinate corresponding to the k^{th} point z_{ik}. Then, the sum of these shortest distances between two stations z_i and z_j along the route gives the river distance between the stations. Following Rouquette et al. (2013), the river distance can be simply written as

$$d(z_i, z_j) \equiv \begin{cases} |z_i - z_j| & \text{if } z_i \text{ and } z_j \text{ are flow connected,} \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

after considering the directionality of the river network.

SPATIAL NEIGHBORHOOD DETERMINATION

The spatial neighborhood for the river data is determined using the nearest neighbor algorithm based on the river distance and flow of the water along the river stream network. The neighbors for a point must lie on the same stream, from the upper stream to the downstream of the river network. For the river network in Figure 4, there are twelve sources of river flow, which are the most upstream of the river network and a single most downstream. A station located at the most downstream is connected to all stations located at the upper stream. Meanwhile, a station at the uppermost stream is connected only to the stations on the corresponding streamflow. Thus, the k neighbors for each station are determined according to their respective groups of the river flow streams. A matrix of the river network distance between every station and every other station is therefore obtained. Then the neighbor of each station is sorted according to the ascending order of the river distance values. A matrix of all sorted neighbors of each station is called a neighborhood matrix.

ILLUSTRATION OF SPATIAL OUTLIERS IN RIVER NETWORK

In this section, we used an artificial data set to illustrate the global and local outliers. We simulated $z = 50$ stations with two sets of geographical coordinates on the river network and two sets non-spatial attributes. The plot

in Figure 2(a) shows the bivariate data, which follow a normal distribution. The ellipse corresponds to values of the chi square of the robust Mahalanobis distance based on FAST-MCD. All the stations outside the ellipse are identified as a global outlier. Figure 2(b) shows the spatial *X*- and *Y*- coordinates of the stations on a river network. There are three selected stations with filled symbols. We chose $k = 6$ nearest neighbors along the river streams for each selected station, and these stations are drawn with the corresponding open symbols. Similar symbols are used in Figure 2(a); thus, we can see the relation between the values in the variable space and the coordinate space. The filled square corresponds to the global outlier since its neighbors' values are located outside the ellipse on the variable space. The filled triangle is a global outlier but not its neighbors. Meanwhile, the filled circle is in the ellipse of the variable space, but its neighbors are very different. Therefore, the filled square is identified as a global outlier; the filled triangle is a global and local outlier, while the filled circle is identified as a local outlier.

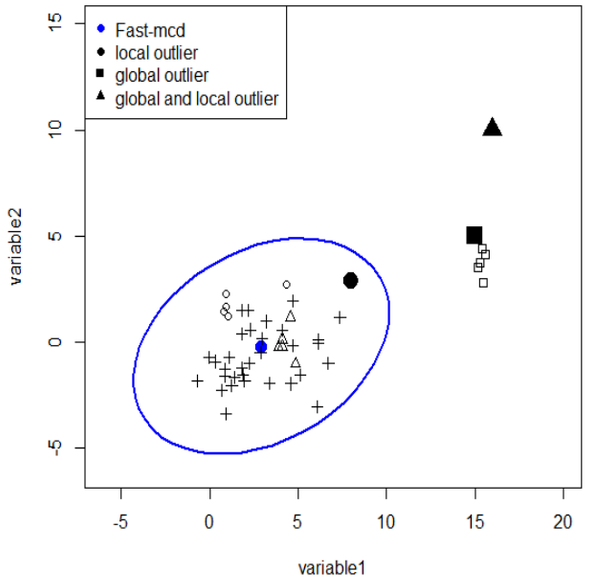

FIGURE 2a. Plot of the bivariate data

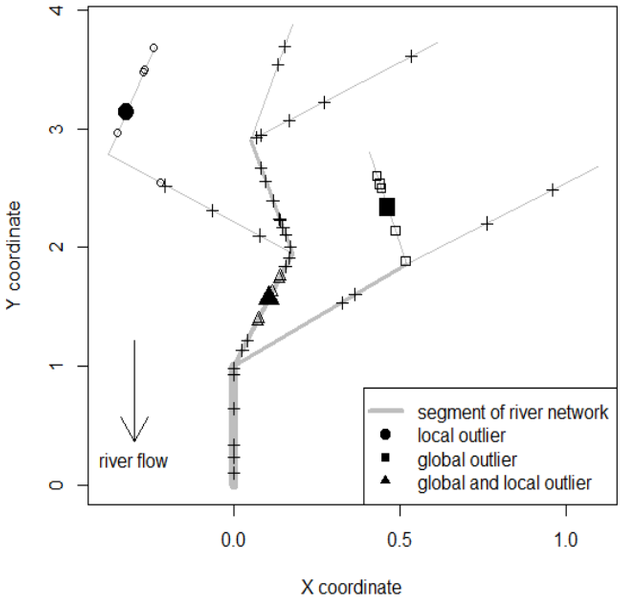

FIGURE 2b. Plot of spatial data

THE PROPOSED METHOD

We defined z_i for $i = 1, 2, \ldots n$ as the spatial point or station on the river network *S*. For each z_i, observations are consisting of non-spatial attributes $x_1, x_2, \ldots, x_p$. In river water quality study, z_i correspond to the i^{th} monitoring station while x_j takes the water quality parameter such as dissolving oxygen and biochemical oxygen demand which will be explained in the application section. Thus, we had a function $f_{z_i}: W_{z_i} \rightarrow R^p$, where R^p denotes the *p* dimensional variable space such that, attribute function W_{z_i} represents the attribute values of stations z_i. For simplicity, we used the notation $W_{z_1}, W_{z_2}, \ldots, W_{z_n}$ for the sample values on the stations and considered the samples following a normal distribution $N_p(\mu, \Sigma)$ with $\mu \in R^p$ and Σ is a $p \times p$ symmetric positive definite matrix. The global outliers in multivariate data are detected using the robust Mahalanobis distance given by

$$MD_{\mu,\Sigma}(W_{z_i}) = \sqrt{(W_{z_i} - \mu)^t \Sigma^{-1} (W_{z_i} - \mu)}, \qquad (2)$$

where the center μ and covariance Σ are estimated from the data. The robust minimum covariance determinant (FAST-MCD) introduced by Rousseeuw and Van Driessen (1999) is considered for estimating the mean, μ and covariance, Σ, to deal with the influence of outlying observations. The global outliers are determined when the values of the robust distance are more significant than the cut-off value. Here the cut-off value is the square root of 97.5% quantile of the chi square distribution

with p degrees of freedom, $\sqrt{\chi^2_{p;0.975}}$. However, the distance measured in (2) does not account for the spatial dependence among the stations as it only identifies the observations that differ from the majority of the data. Thus, the local outlier detection method adopted from Filzmoser et al. (2014) was constructed based on pairwise robust Mahalanobis distance between the observations at two stations, z_i and z_j, given by

$$MD^2_{\Sigma}(W_{z_i}, W_{z_j}) = (W_{z_i} - W_{z_j})^t \Sigma^{-1} (W_{z_i} - W_{z_j}). \quad (3)$$

The robust covariance FAST-MCD estimate, Σ, is then plugged into (3) and $z_j \subset D_{z_i}$. Now, we want to choose a subset of D_{z_i}, which can be considered as the neighbors of z_i. It can be achieved by considering the k nearest neighbor method but using the river network distance as described by (1). All the neighbors are then sorted to the ascending order of the distance values. Once a spatial point is identified to be different from most of its k neighbors, then it is a potential local outlier. A local outlier is then determined by the degree of isolation of a spatial point from a fraction of its neighbors denoted as $\alpha(i)$-quantile given by

$$\chi^2_{p;\alpha(i)}\left(MD^2(W_{z_i})\right) = MD^2\left(W_{z_i}, W_{z_{(\lfloor n(i)\cdot\beta \rfloor)}}\right) \text{ for } i = 1, \ldots, n. \quad (4)$$

The pairwise squared Mahalanobis distance on the right-hand side in (4) is a non-central chi square distribution with p degree of freedom. The non-centrality parameter of the squared Mahalanobis distance is represented on the left-hand side in (4). The neighbors of a spatial point z_i are denoted as $W_{z_{(\lfloor n(i)\cdot\beta \rfloor)}}$ where $n(i)$ is the number of neighbors, k; while β denotes a fraction of neighbors. The cut-off point is determined by β-value. If $\alpha(i)$ is significantly larger than β, observation z_i considered as a local outlier. In this study, a local outlier was detected when $\alpha(i)$ is greater than 10% (Filzmoser et al. 2014).

The proposed algorithm of the detection of spatial outliers is presented as follows. Given a spatial data set $z = \{z_1, z_2, \ldots, z_n\}$,

1. Set the attribute function (non-spatial attribute), a number of k nearest neighbors, $n(i)$ and a fraction of neighbors, β.

2. Calculate the robust covariance matrix Σ_p for the non-spatial attributes.

3. Compute $MD_{\mu,\Sigma}(W_{z_i}) = \sqrt{(W_{z_i} - \mu)^t \Sigma^{-1} (W_{z_i} - \mu)}$. If $MD \geq \sqrt{\chi^2_{p;0.975}}$, then z_i is a global outlier.

4. Calculate the distance between two pairs of observation z_i and z_j, where $i \neq j$. If the pairs of points are flow-unconnected, $z_i \nrightarrow z_j$, then the distance between these two points are equal to zero.

5. Determine and sort the k nearest neighbor for each spatial point. The first nearest neighbor has the smallest distance value with the candidate of spatial outlier and they are flow-connected.

6. Then, compute the degree of outlier for each spatial point by using $\chi^2_{p;\alpha(i)}\left(MD^2(W_{z_i})\right) = MD^2\left(W_{z_i}, W_{z_{(\lfloor n(i)\cdot\beta \rfloor)}}\right)$ for $i = 1, \ldots, n$.

7. Sort the value of the degree of isolation. If the degree of isolation is significantly larger than 10%, then we classify the observation as a local and global outlier when $MD \geq \sqrt{\chi^2_{p;0.975}}$, (step 3). Otherwise, the observation is classified as a local outlier only.

The algorithm of the river distance and the Euclidean distance above are summarized by a flowchart as presented in Figure 3.

SIMULATION STUDY

The performance of the proposed method was studied via simulation. For this purpose, we generated a synthetic river data set using an *R* package known as *SSN* (Ver Hoef et al. 2014) to resemble the actual river network. We then compared the proposed method with the method found in Filzmoser et al. (2014) in terms of their capability to detect outliers in the flow-connected river data.

DATA SIMULATION

A synthetic river data set is constructed by a spatial stream network package, a built-in *R* package known as *SSN* (Ver Hoef et al. 2014). The construction of the data set involves two steps. Firstly, we created a river network and secondly, we simulated auto-correlated variables on the river network. A river network is generated randomly like tree structures. For the formulated data set, such as in Figure 4, we used the iterative *TreeLayout* function, producing a more realistic network and not

creating any self-interactions (Ver Hoef & Peterson 2010). The stations' locations are generated and determined using *binomialDesign* function. Here, we randomly generated $z = 30$ stations on the constructed river network, with each of them consisting of coordinate values, *X*- and *Y*- coordinate, and distance values from a station to station at the lowest downstream of the river network. All of this information is known as spatial attributes. For the second step, we simulated the non-spatial attributes, W_{z_i}, on each station. Gaussian normal data sets are created for this simulation based on the correlation models: the *Exponential.tailup* function and the *Exponential.taildown* function (see details in Ver Hoef & Peterson (2010)).

SIMULATION FOR PERFORMANCE MEASURE

In this section, we use a synthetic data set as described before to investigate the performance of the proposed method. First, we randomly chose 4 stations on the river network and set a point as the global outliers and the rest as the local outliers. The synthetic data set properties are shown in Table 1.

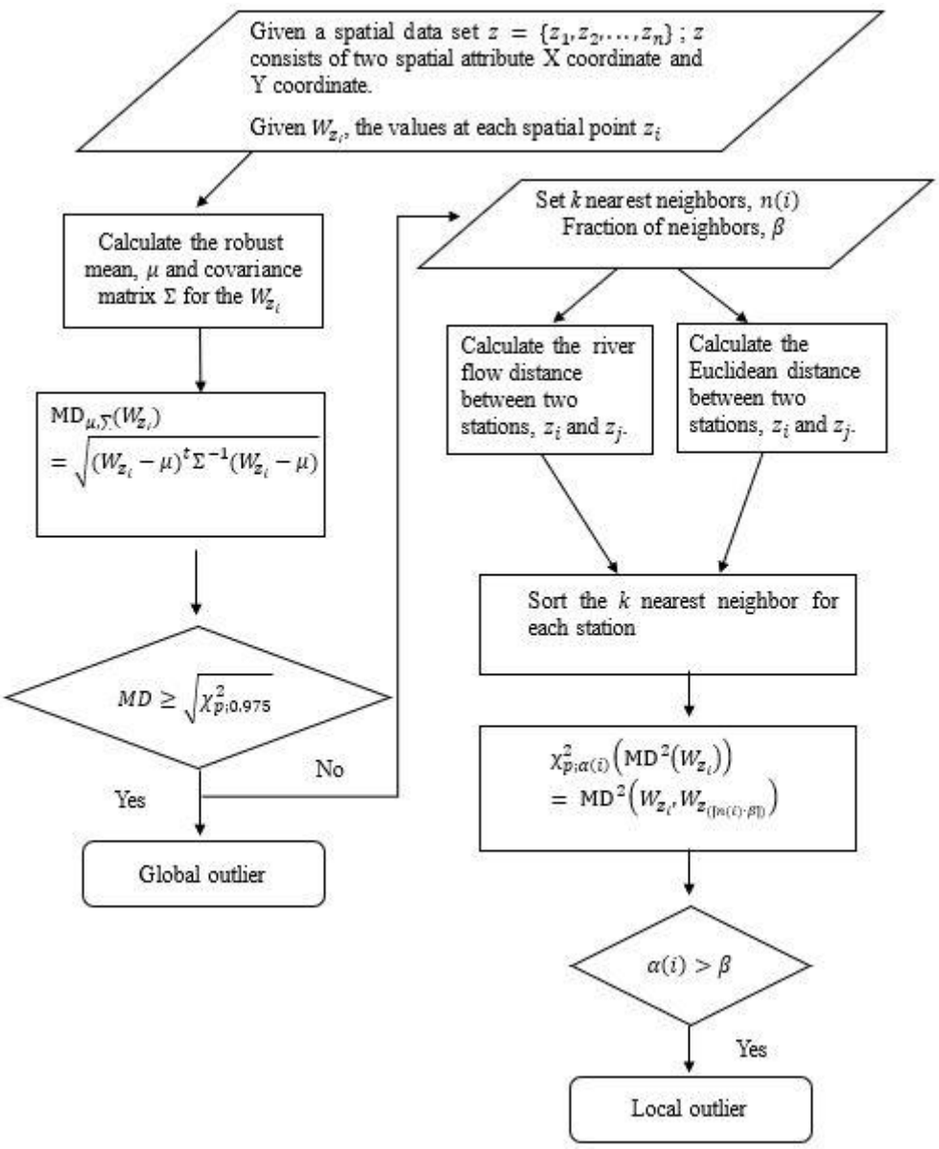

FIGURE 3. Flowchart of methodology of the spatial outlier detection

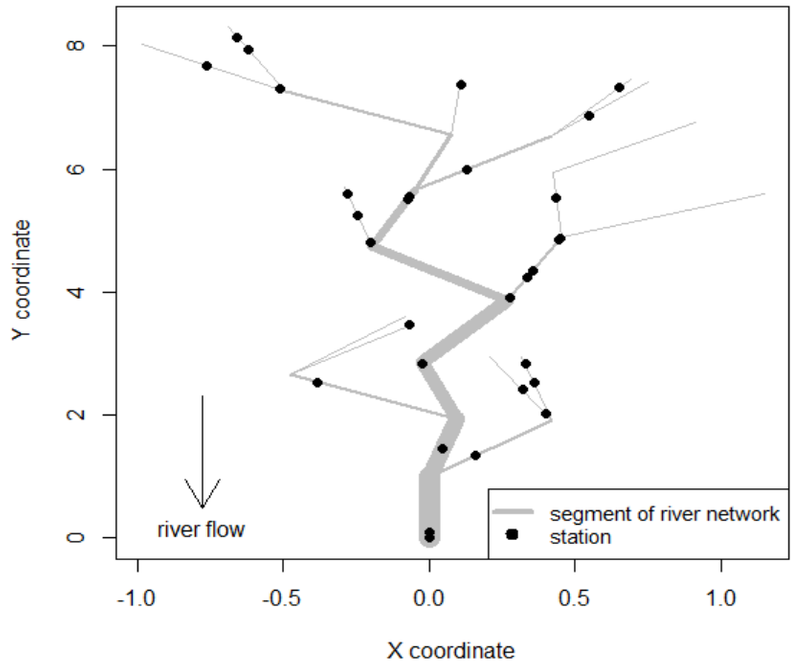

FIGURE 4. A synthetic river network generated with 30 stations

Second, we calculate the degree of isolation for each station using (4). We then rank the station according to the degree of isolation. Station with degree of isolation greater than the cut-off point β is identified as a local outlier. Third, we summarized the results in a confusion matrix as shown in Table 2. True positive rate (TPR) is the ratio between true positive and the summation of true positive and false negative, while false positive rate (FPR) is the ratio between false positive and the summation of false positive and true negative. The overall performance

TABLE 1. Synthetic data set properties

	Type				
	Station, z	Variables, W	Regular observations	Global outlier	Local outlier
Synthetic Data	30	2	26	1	3

can be depicted by the plot of the true positive rate against the false positive rate. The plot is also known as the Receiver Operating Characteristic (ROC) curve which represents the trade-off between the TPR and FPR. The area under the ROC curve, also known as AUC, has been widely used in measuring the performance of outlier detection methods, and the value will always be between 0 and 1. A larger AUC value indicates a better classification performance (Fawcett 2006).

SIMULATION RESULTS

We used the simulation procedure above to compare the performance of the spatial outlier detection methods. The experiment was repeated 500 times for different k =1,…, 30, β = 0.05, 0.1, 0.2 and sample size $nsim$=50, 100, 500. The results are reported in Table 3 and Figure 5.

From Table 3, the maximum and average values of AUC for the river distance method are greater than that for the Euclidean distance method. Moreover,

TABLE 2. Confusion matrix

		Predicted Label		Total
Positive		Negative		
Actual Label	Positive	True Positive (TP)	False Negative (FN)	TP+FN
	Negative	False Positive (FP)	True Negative (TN)	FP+TN
Total		TP+FP	FN+TN	TP+FP+ FN+TN

as we expect, the mean value of AUC increases with the increasing number of simulations from 50 to 100. However, the results for *nsim*=100 and 500 do not differ much. The overall results can be better seen when presented graphically as shown in Figure 5. It is observed that the proposed method performs better when $\beta = 0.05$ and $\beta = 0.1$. However, the results are almost the same for $\beta = 0.2$. We also note that the mean AUC value for the river distance method is higher for a greater k. Hence, we may conclude that the river distance method provides a good alternative when working with data within a river network. Other results of the performance study are available from the authors upon request.

TABLE 3. The performance of the outlier detection methods

	Methods	*AUC*					
		nsim = 50		*nsim* = 100		*nsim* = 500	
		Mean	sd	mean	sd	mean	sd
0.05	River distance method	0.38	±0.02	0.39	±0.02	0.41	±0.02
	Euclidean distance method	0.27	±0.09	0.29	±0.09	0.31	±0.08
0.1	River distance method	0.41	±0.09	0.41	±0.04	0.44	±0.03
	Euclidean distance method	0.35	±0.09	0.36	±0.09	0.37	±0.09
0.2	River distance method	0.25	±0.09	0.25	±0.07	0.25	±0.06
	Euclidean distance method	0.20	±0.07	0.22	±0.06	0.24	±0.06

APPLICATION TO REAL DATA

We demonstrate the proposed method of identifying spatial outliers using Sg. Klang data set for the year 2016. The data is obtained from the Department of Environment, Malaysia. Sg. Klang Basin is located within the states of Selangor and Kuala Lumpur in Malaysia. The river drains 1288 km^2 from the steep

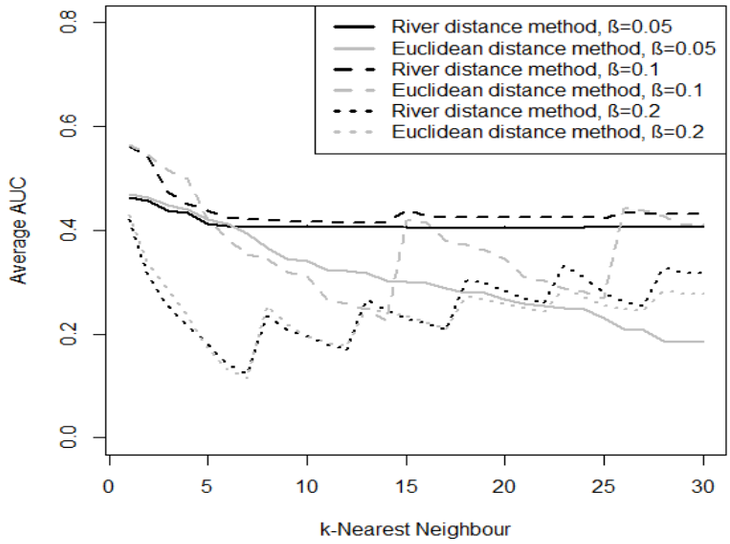

FIGURE 5. Average AUCs value versus the number of neighborhood for the river distance method and the Euclidean distance method for β = 0.05,0.1,0.2

mountain rainforests of the main centre of Peninsular Malaysia to the river mouth in Port Klang. There are 16 water quality stations located along the rivers, as shown in Figure 6. The basin consists of the main Sg. Klang and 11 tributaries, including Sg. Gombak, Sg. Kerayong, Sg. Penchala and Sg. Damansara. Stations 1 to 7 are located along Sg. Klang, Stations 9 to 10 are along Sg. Penchala while Stations 12 to 14 are along Sg. Gombak. Stations 8 and 11 are located along Sg. Damansara and Sg. Kerayong, respectively. The river flow connectivity between the stations are summarized in Table 4. It can be seen that Station 8 does not connect to any other station in the river flow direction. Besides, since Station 1 is located at the most river downstream, it connects

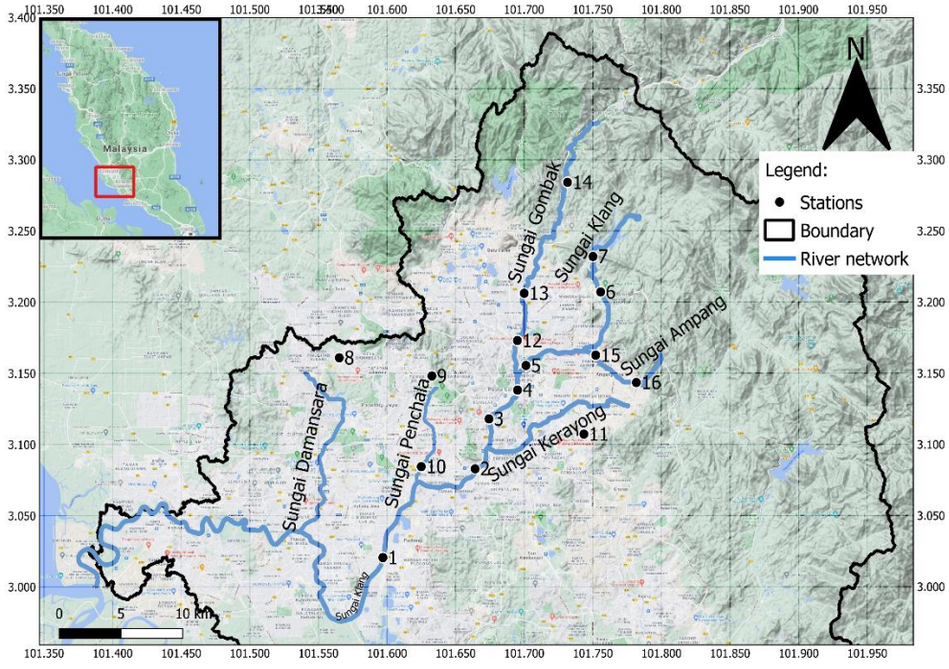

FIGURE 6. Location of stations of Sg. Klang Basin

to all stations upstream except Station 8 and has the highest connectivity percentage. The topography of Sg. Klang basin is given in Figure 6 with the higher ground indicated by darker color. We can see that Station 7 is located at the most upstream of Sg. Klang while Stations 8, 9, 11, 14 and 16 are the most upstream of their own tributaries. These stations are generally on the higher ground in Sg. Klang basin. Station 1 is located nearest to the mouth of Sg. Klang. Good water qualities are expected at Stations 7 and 14 since they are located near the water source.

Parameters measured at each station include dissolving oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SS), ammoniacal nitrogen (NH_3NL), temperature, and pH. For example, DO levels in waterfalls are typically

TABLE 4. The percentage of the flow connectivity

The i^{th} Station, (z_i)	River	Percentage of connectivity to other stations	The i^{th} Station, (z_i)	River	Percentage of connectivity to other stations
1	Sg. Klang	0.93	9	Sg. Penchala	0.13
2	Sg. Klang	0.8	10	Sg. Penchala	0.13
3	Sg. Klang	0.73	11	Sg. Kerayong	0.13
4	Sg. Klang	0.73	12	Sg. Gombak	0.4
5	Sg. Klang	0.53	13	Sg. Gombak	0.4
6	Sg. Klang	0.4	14	Sg. Gombak	0.4
7	Sg. Klang	0.4	15	Sg. Ampang	0.4
8	Sg. Damansara	0	16	Sg. Ampang	0.4

higher than those in pools and slower-moving stretches. The process of respiration consumes oxygen in water by aquatic animals, decomposition, and various chemical reactions. Wastewater from sewage treatment plants often contains organic materials that are decomposed by microorganisms. The amount of oxygen consumed by these organisms in breaking down the waste is known as BOD. The COD measures the amount of oxygen required to oxidize the organic material present in water chemically. Thus, the BOD and COD measure the total amount of oxygen removed from water biologically or chemically in a specified time and at a specific temperature. The SS is a measure of suspended particulate matter produced by anthropogenic sources such as urban development, road building, land clearing, and agriculture (NOA 2020). Ammonium is an ionized form of ammonia. The measurement of ammonium indicates the potential to form ammonia or ammoniacal nitrogen pollutants in rivers when pH and temperature change (Ibrahim et al. 2015). The hydrolysis of organic nitrogen can form ammoniacal nitrogen and enter the river system directly from industrial or sewage effluent. These parameters are essential to assess the quality status of river water, known as the water quality index (WQI). The determination of WQI for each location also permits the categorical class based on the National Water Quality Standard (NWQS). The scores WQI ranges from 0 to 100, where the state of the river varies between polluted to clean, where 0 to 59 scores are polluted rivers, 60 to 80 scores are slightly polluted, and 81 to 100 scores are considered clean rivers. Table 5 provides overall values of summary statistics of the water quality parameters of Sg. Klang basin in 2016. We can see that mean values of DO range from 3.9 to 8.5 mg/L, indicating moderate to high DO levels in the river water. The maximum BOD and NH_3NL levels are considerably high, which are 18.9 mg/L and 8.6 mg/L, respectively, and might contribute to bad water quality. The mean COD level is 30.3 mg/L indicating that most stations are polluted at different degrees. In addition, some stations have significantly high TSS value which is more

than 300 mg/L. Overall, the WQI status of Sg. Klang data ranges between 39.8 and 65.8, indicating different water quality at stations in Sg. Klang basin.

In order to continue with the local outlier detection method, the neighbors are constructed based on the spatial location for each station. Again, the set of neighbors for

TABLE 5. Summary statistics of river water quality parameters

Variables	DO mg/L	BOD mg/L	COD mg/L	TSS mg/L	pH	NH_3NL mg/L	Temp (°C)	WQI
Min	3.9	5.5	16.1	11.21	7.0	0.1	25.9	39.8
Max	8.5	18.9	52.8	580.5	7.7	8.6	30.8	65.8
Mean	5.3	10.6	30.3	81.6	7.4	4.4	29.2	50.4
Std. dev	1.5	3.6	10.1	138.0	0.2	3.0	1.1	8.0

the monitoring station may differ for the river distance and Euclidean distance methods because the river distance method considers the effect of river flow within the river network topology. We then apply the two methods to the data. Three crucial results are observed. Firstly, the Mahalanobis distance for each monitoring station is computed and plotted in Figure 7. There are four monitoring stations with distances greater than $\chi^2_{p;0.975}$. They are Stations 1, 2, 8, and 14 and are classified as global outliers. Other stations are referred to as regular observations. Secondly, we identify local outliers among the regular observations based on the degree of isolation as tabulated in Table 6. We found different sets of local outliers for both methods, that station with the degree of isolation exceeding 10%. Station 11 has the highest degree of isolation in the river distance method, followed by Stations 10, 6, 9, 15, and 7. Meanwhile, for the Euclidean distance method, Station 9 has the highest degree of isolation, followed by Stations 6, 7, and 11. Thirdly, we also calculate the degree of isolation for the global outliers. The results are tabulated in the first few rows of Table 6. Both methods identify Station 14 as the only global and local outlier. We summarize the results in Table 7.

Table 8 tabulates the parameter values of water quality for the identified spatial outliers. The overall

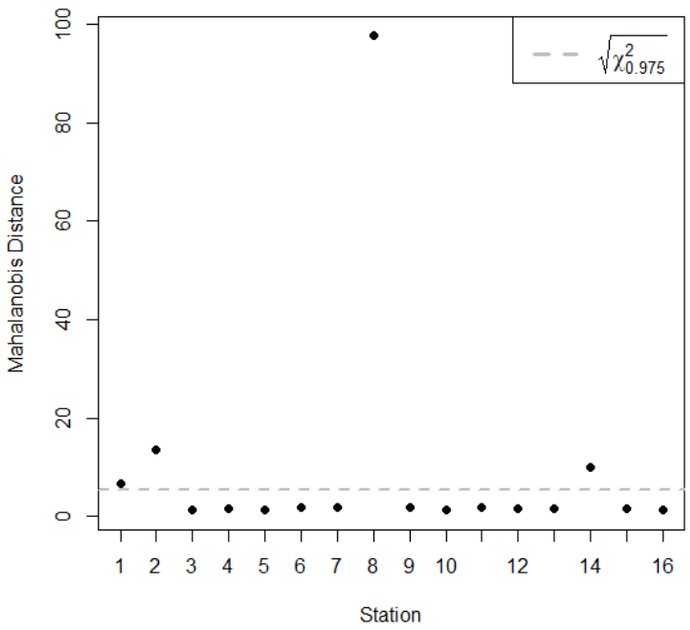

FIGURE 7. Plot of Mahalanobis distance for identifying global outliers

TABLE 6. Degree of isolation for global outliers and local observations

Station	Degree of isolation (%)	
Global outliers	River distance method	Euclidean distance method
1	6.54	6.54
2	0.00	0.00
8	0.00	0.00
14	14.84	28.70
Regular observations		
3	0.08	0.08
4	9.64	9.64
5	0.10	0.10
6	19.07	13.06
7	12.38	12.38
9	16.29	16.29
10	25.36	2.78
11	99.99	12.25
12	5.90	5.90
13	6.50	6.50
15	15.41	5.98
16	4.91	4.91

TABLE 7. List of outliers

Types	River distance method	Euclidean distance method
Global outlier	1, 2, 8, 14	1, 2, 8, 14
Local outlier	6, 7, 9, 11, 10, 15	6, 7, 9, 11
Global and local outlier	14	14

mean values for each parameter are given in the second row. We can see that some of the parameter values are much higher or lower than the overall means indicating the correct identification of these stations as spatial outliers. For example, Stations 1, 2, and 8 are situated downstream of the basin. We observed that Station 8 has an extremely high reading of SS but a shallow reading NH_3NL. As for Station 2, the SS reading is also very high, while Station 1 has the lowest reading of DO. On the other hand, Station 14, located upstream of the river, recorded the best water quality in the river basin.

Moreover, we can observe that the stations identified as a local outlier by both methods have some parameter values that differ from their neighboring stations. Stations 6 and 7 are neighbors and separated from other stations, but with different readings of NH_3NL, which is much lower than their neighbors, resulting in different water quality classes. Hence, they are identified as local outliers. As for Station 11, its neighbors are Stations 1 and 2, which are global outliers. A similar argument stands for Station 9.

More importantly, Stations 10 and 15 are also identified as local outliers by the river distance method. The neighbors of Station 10 under the river network (Figure 8(a)) is only Stations 1 and 9 compared to the additional neighbors, Stations 2, 3, and 4 under the

TABLE 8. Mean parameter values for detected outliers

	DO mg/L	BOD mg/L	COD mg/L	SS mg/L	pH	NH3NL mg/L	TEMP (°C)	WQI
Overall mean	5.30	10.60	30.28	81.55	7.41	4.40	29.17	50.41
				Station (Global outliers)				
1	3.93	10.66	31.16	53.33	7.44	4.59	29.27	50.08
2	4.24	10.33	28.70	173.62	6.98	5.15	29.00	45.36
8	7.55	12.25	39.65	580.5	7.54	0.09	30.82	52.86
14	8.45	5.38	18.50	13.08	7.73	0.15	25.92	65.82
				Station (Local outliers)				
6	5.57	7.70	22.83	43.37	7.28	2.53	29.09	52.86
7	6.80	5.70	16.08	11.20	7.50	0.42	29.83	64.08
9	7.22	5.45	16.66	23.87	7.51	0.21	28.95	65.10
*10	4.09	15.54	44.79	45.50	7.29	8.26	29.26	43.20
11	4.11	18.87	52.79	31.45	7.36	8.63	30.79	39.84
*15	5.63	13.79	40.12	35.17	7.39	8.35	29.87	44.38

* outliers identified by river distance method

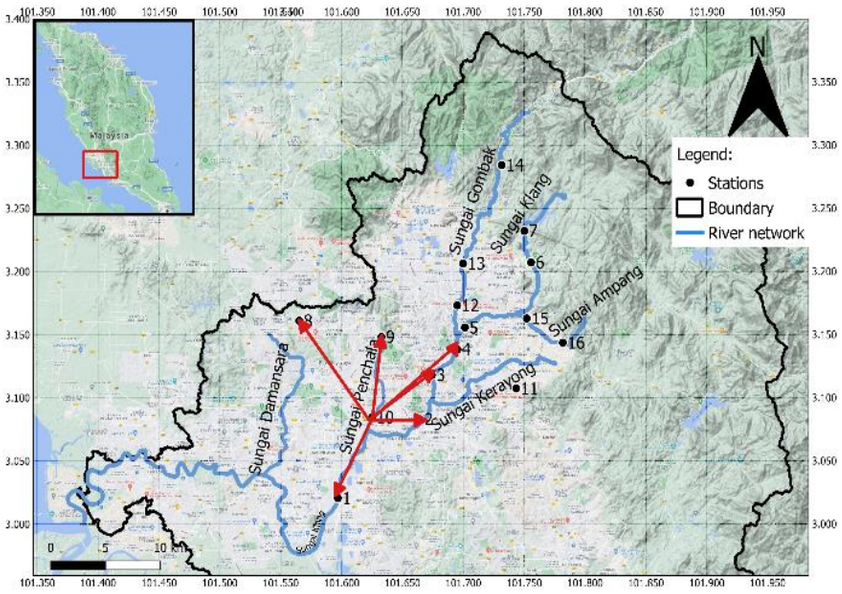

FIGURE 8a. Local outliers by Euclidean distance method

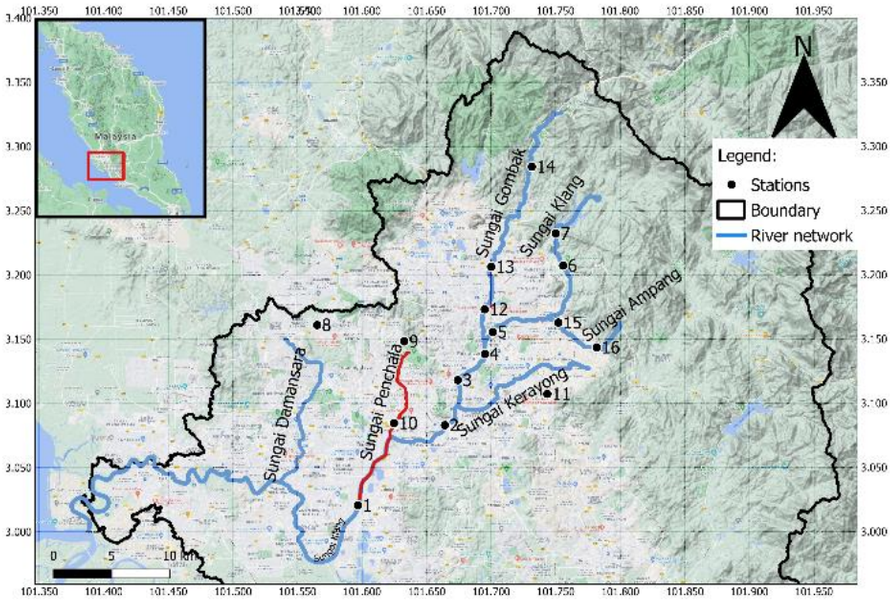

FIGURE 8b. Local outliers by the river distance method

Euclidean distance (Figure 8(b)). However, water quality at Station 10 does not differ much from Stations 2, 3, and 4, so it is not identified as a local outlier by the Euclidean distance method. However, the water quality of Station 10 differs from that of Stations 1 and 9. Hence identified as a local outlier by the river distance method. A similar argument also holds for Station 15.

Thus, the proposed method removes the false effect of the station on different river flow system when identifying the spatial outlier in the river network system. This is achieved by considering the neighbors which are flow-connected to each other only. Hence, the proposed method provides important advantages toward improving the accuracy of detecting spatial outlier in river network. Here, the results can now be used by the authority to monitor river water quality in Sg. Klang Basin. For example, the local authority can focus on Station 10, located at Sg. Penchala, by implementing steps that can improve the water quality in the area. Moreover, the concept proposed in this paper can be generalized to other real applications such as detecting abnormal reading in the gas flow in the piping system to avoid explosion and traffic flow data to avoid congestion.

Conclusion

In this paper, we have proposed an improved method of identifying spatial outliers in a river basin by considering the effect of river flow on the determination of neighbors of the monitoring stations. We have also analyzed the computational structures to determine the spatial neighborhoods in river network settings and detect the spatial outliers in the multivariate data. Here, we also evaluated the performance of the proposed method via simulation and compared the results of outlier detection between our approach and method in Filzmoser et al. (2014). Additionally, we provided experimental results from applying our proposed method on water quality data of Sg. Klang Basin to show its effectiveness and usefulness. Our approach highlighted the importance of the river flow distance to determine the degree of isolation of an observation. The pairwise robust Mahalanobis distance requires the correct definition of its neighborhood, especially to analyze the spatial outlier in the river data. Incorporating the river flow distance into the spatial outlier method provided an advantage in detecting the spatial outlier on the river network and decreased the error while determining the true outlier. The AUC values for the river distance method are slightly higher than the existing detection method from the simulation results. Thus, the proposed method is suitable for spatial attributes on the river network and multivariate data. Also, the truth concerning the spatial outlier observations

is unknown, and the outlier detection methods may be interpreted in the context (Ernst & Haesbroeck 2017). Thus, considering the river flow distance is the right decision to enhance spatial outlier detection in the river data. However, the performance of the proposed method may be further enhanced by increasing the local nature of spatial points to estimate the covariance matrix in the spatial outlier detection algorithms. In addition, the percentage of river flow connectivity between stations is another interesting factor that can be considered for future work. The proposed framework is useful to be applied in other real-life applications that use similar river network property.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2018/STG06/UM/02/12), Universiti Malaya Research Grant (RF015B-2018), and the Department of Environment, Malaysia.

REFERENCES

Alok Kumar, S. & Lalitha, S. 2018. A novel spatial outlier detection technique. *Communications in Statistics-Theory and Methods* 47(1): 247-257.

Anselin, L. 1995. Local Indicators of Spatial Association - LISA. *Geographical Analysis* 27(2): 93-115.

Azimi, A., Bagheri, N., Mostafavi, S.M., Furst, M.A., Hashtarkhani, S., Amin, F.H. & Kiani, B. 2021. Spatial-time analysis of cardiovascular emergency medical requests: Enlightening policy and practice. *BMC Public Health* 21(1): 1-12.

Baur, C., Denner, S., Wiestler, B., Navab, N. & Albarqouni, S. 2021. Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. *Medical Image Analysis* 69: 101952.

de Fouquet, C. & Bernard-Michel, C. 2006. Geostatistical models for concentrations or flow rates in streams. *Comptes Rendus Geoscience* 338(5): 307-318.

Cai, Q., He, H. & Hong Man. 2009. SOMSO: A self-organizing map approach for spatial outlier detection with multiple attributes. In *IEEE International Joint Conference on Neural Networks*. pp. 425-431.

Chen, D., Lu, C-T., Kou, Y. & Chen, F. 2008. On detecting spatial outliers. *Geoinformatica* 12(4): 455-475.

Cressie, N.A.C. 1993. *Spatial Statistics.* New York: John Wiley and Sons. Inc.

Cressie, N., Frey, J., Harch, B. & Smith, M. 2006. Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics* 11: 127-150.

Ernst, M. & Haesbroeck, G. 2017. Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery* 31(2): 371-399.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861-874.

Filzmoser, P., Ruiz-Gazen, A. & Thomas-Agnan, C. 2014. Identification of local multivariate outliers. *Statistical Papers* 55(1): 29-47.

Hasib, N.A. & Othman, Z. 2020. Assessing the relationship between pollution sources and water quality parameters of Sungai Langat Basin using association rule mining. *Sains Malaysiana* 49(10): 2345-2358.

Haslett, J. 1992. Spatial data analysis-challenges. *Journal of the Royal Statistical Society: Series D (The Statistician)* 41(3): 271-284.

Ibrahim Mohamed, Faridah Othman, Adriana IN Ibrahim, ME Alaa-Eldin & Rossita M Yunus. 2015. Assessment of water quality parameters using multivariate analysis for Klang River basin, Malaysia. *Environmental Monitoring and Assessment* 187(1): 4182.

Jat, P. 2017. Geostatistical estimation of water quality using river and flow covariance models. PhD Thesis. The University of North Carolina at Chapel Hill (Unpublished).

Kelleher, C. & Braswell, A. 2021. Introductory overview: Recommendations for approaching scientific visualization with large environmental datasets. *Environmental Modelling & Software* 143: 105113.

Kou, Y. 2006. Abnormal pattern recognition in spatial data. PhD thesis. Virginia Tech. (Unpublished).

Kou, Y., Lu, C-T. & Chen, D. 2016. Spatial weighted outlier detection. In *Proceedings of the 2006 SIAM International Conference on Data Mining.* SIAM, 2006. pp. 614-618.

Lachhab, A., Trent, M.M. & Motsko, J. 2021. Multimetric approach in the effects of small impoundments on stream water quality: Case study of Faylor and Walker Lakes on Middle Creek, Snyder County, PA. *Water and Environment Journal* 35(3): 1007-1017.

Laporan Kualiti Alam Sekeliling. 2019. {Enviro Knowledge Center. Technical report, Department of Environment Malaysia, Nov 2020. https://enviro2.doe.gov.my/ekmc/digital-content/laporan-kualiti-alam-sekeliling-2019/.

Liu, F., Su, W., Zhao, J. & Liang, X. 2017. On-line detection method for outliers of dynamic instability measurement data in geological exploration control process. *Sains Malaysiana* 46(11): 2205-2213.

Lu, C-T., Chen, D. & Kou, Y. 2003. Algorithms for spatial outlier detection. In *Third IEEE International Conference on Data Mining*. pp. 597-600.

Mainali, J. & Chang, H. 2021. Environmental and spatial factors affecting surface water quality in a Himalayan watershed, Central Nepal. *Environmental and Sustainability Indicators* 9: 100096.

Money, E.S., Sackett, D.K., Aday, D.D. & Serre, M.L. 2011. Using river distance and existing hydrography data can improve the geostatistical estimation of fish tissue mercury at unsampled locations. *Environmental Science & Technology* 45(18): 7746-7753.

Money, E., Carter, G.P. & Serre, M.L. 2009a. Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in New Jersey. *Water Research* 43(7): 1948-1958.

Money, E., Carter, G.P. & Serre, M.L. 2009b. Modern space/time geostatistics using river distances: Data integration of turbidity and *E. coli* measurements to assess fecal contamination along the Raritan River in New Jersey. *Environmental Science & Technology* 43(10): 3736-3742.

Peters, N.E. & Meybeck, M. 2000. Water quality degradation effects on freshwater availability: impacts of human activities. *Water International* 25(2): 185-193.

Peiman Asadi, Davison, A.C. & Engelke, S. 2015. Extremes on river networks. *The Annals of Applied Statistics* 9(4): 2023-2050.

Peter Chu Su. 2011. Statistical geocomputing: Spatial outlier detection in precision agriculture. Master's thesis. University of Waterloo (Unpublished).

Peterson, E.E. & Urquhart, N.S. 2006. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Maryland. *Environmental Monitoring and Assessment* 121(1-3): 615-638.

Peterson, E.E., Merton, A.A., Theobald, D.M. & Urquhart, N.S. 2006. Patterns of spatial autocorrelation in stream water chemistry. *Environmental Monitoring and Assessment* 121(1-3): 571-596.

Rouquette, J.R., Dallimer, M., Armsworth, P.R., Gaston, K.J., Maltby, L. & Warren, P.H. 2013. Species turnover and geographic distance in an urban river network. *Diversity and Distributions* 19(11): 1429-1439.

Rousseeuw, P.J. & Van Driessen, K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41(3): 212-223.

Sajesh, T.A. & Srinivasan, M.R. 2013. An overview of multiple outliers in multi-dimensional data. *Sri Lankan Journal of Applied Statistics* 14(2): 87-120.

Shekhar, S., Lu, C-T. & Zhang, P. 2003. A unified approach to detecting spatial outliers. *GeoInformatica* 7(2): 139-166.

Talagala, P.D., Hyndman, R.J., Leigh, C., Mengersen, K. & Smith-Miles, K. 2019. A feature-based procedure for detecting technical outliers in water-quality data from *in situ* sensors. *Water Resources Research* 55(11): 8547-8568.

Tortorelli, R.L. & Pickup, B.E. 2006. Phosphorus concentrations, loads, and yields in the Illinois river basin, Arkansas and Oklahoma. 2000-2004. Technical report.

Ver Hoef, J.M. & Peterson, E.E. 2010. A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association* 105(489): 6-18.

Ver Hoef, J.M., Peterson, E., Clifford, D. & Shah, R. 2014. SSN: An R package for spatial statistical modeling on stream networks. *Journal of Statistical Software* 56(3): 1-45.

Ver Hoef, J.M., Peterson, E. & Theobald, D. 2006. Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* 13(4): 449-464.

Wang, S. & Serfling, R. 2018. On masking and swamping robustness of leading nonparametric outlier identifiers for multivariate data. *Journal of Multivariate Analysis* 166: 32-49.

Yang, M., Chen, Z., Zhou, M., Liang, X. & Bai, Z. 2021. The impact of COVID-19 on crime: A spatial temporal analysis in Chicago. *ISPRS International Journal of Geo-Information* 10(3): 152.

Zheng, G., Brantley, S.L., Lauvaux, T. & Li, Z. 2017. Contextual spatial outlier detection with metric learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 2161-2170.

*Corresponding author; email: rossita@um.edu.my