UNIVERSITY OF MALAYA
The Leader in Research & Innovation

Integrating a stream distance measure into a spatial outlier detection algorithm for locating pollution in Klang River Basin

Nur Fatihah Binti Mohd Ali
Institute of Mathematical Scieces
Universiti Malaya
5th October 2022

9th MALAYSIA STATISTICS CONFERENCE 2022
Dealing with Uncertainties: Unearthing Measures for Recovery
Institut Latihan Statistik Malaysia
5th October 2022

Introduction

- In our country, Malaysia, we do depend on river water for various purposes.
- However, river pollution is a serious issue especially in the urban areas.
- The surface water pollution may threaten human health and the ecological system.
- River pollution causes multiple unscheduled interruptions in water supply.

Introduction

- Locating the source of river pollution is the hardest job.
- In statistics, spatial outliers represent locations that are significantly different from their neighborhoods, even though they may not be significantly different from the entire population (Shekhar et al.,2003).
- It helps in finding local instabilities in objects (Surya, 2014).
- Spatial outliers are basically the observed water quality parameters at any monitoring station that are significantly different from the corresponding readings at its neighbors.

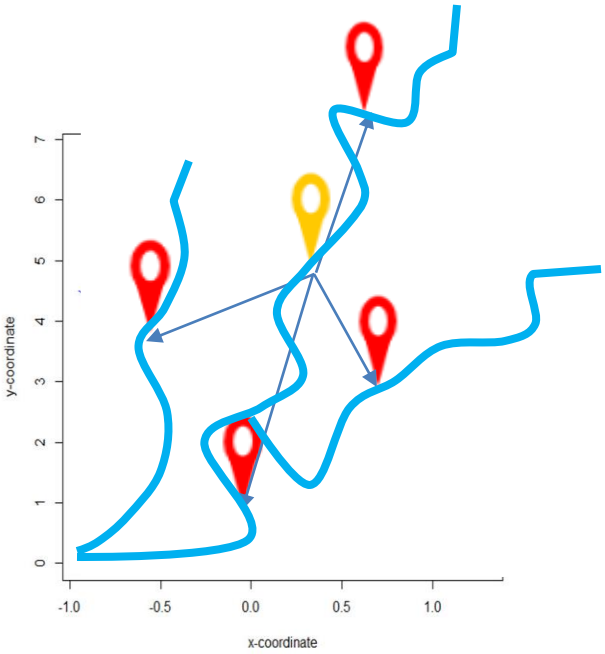


Fig. 1. Example of Spatial Outlier

Objective

- To develop a spatial outlier detection method for locating pollution in the river.

- To identify pollution sources in the Klang River basin.

Data

- Available from Department of Environmental Malaysia (DOE)
- Data is collected from January, 2019 until December, 2019

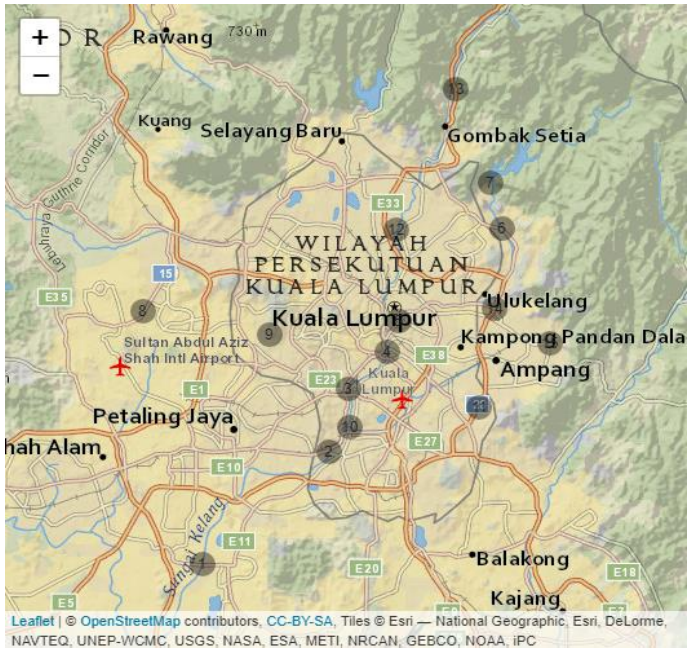
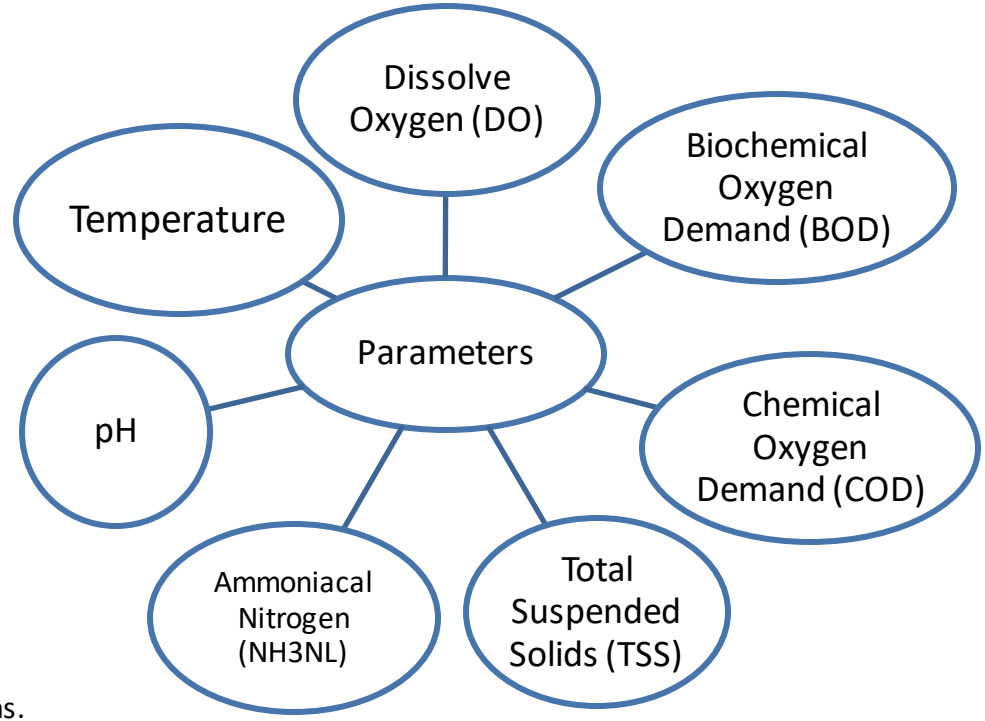


Fig. 2. Map of Klang river with 15 monitoring stations.

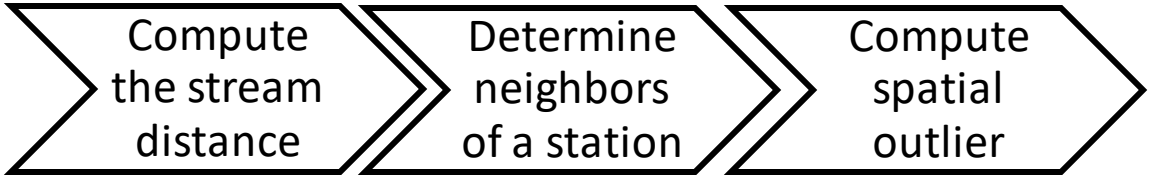


Methodology

For multivariate data, spatial outlier will be identified by Equation 1 (Ali et al. 2022);

$$\chi^2_{p;\alpha(i)}\left(\mathrm{MD}^2\left(W_{z_i}\right)\right) = \mathrm{MD}^2\left(W_{z_i}, W_{z_j}\right) \text{ for } i = 1, \ldots, n. \quad (1)$$

RHS : Pairwise robust Mahalanobis distance
LHS : Non-central Chi-square distribution
p : Degree of freedom
$\alpha(i)$: Degree of isolation
W_{z_i} : Water quality parameters at station z_i
W_{z_j} : Water quality parameters at station z_j



Methodology : Stream Distance

Stream distance is the shortest distance between two locations computed along stream network.

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

Connected station: $z_1 \rightarrow z_2$

Unconnected station : $z_2 \nrightarrow z_3$

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

For real data, the stream distances between the stations can be computed through SSN package in R software.

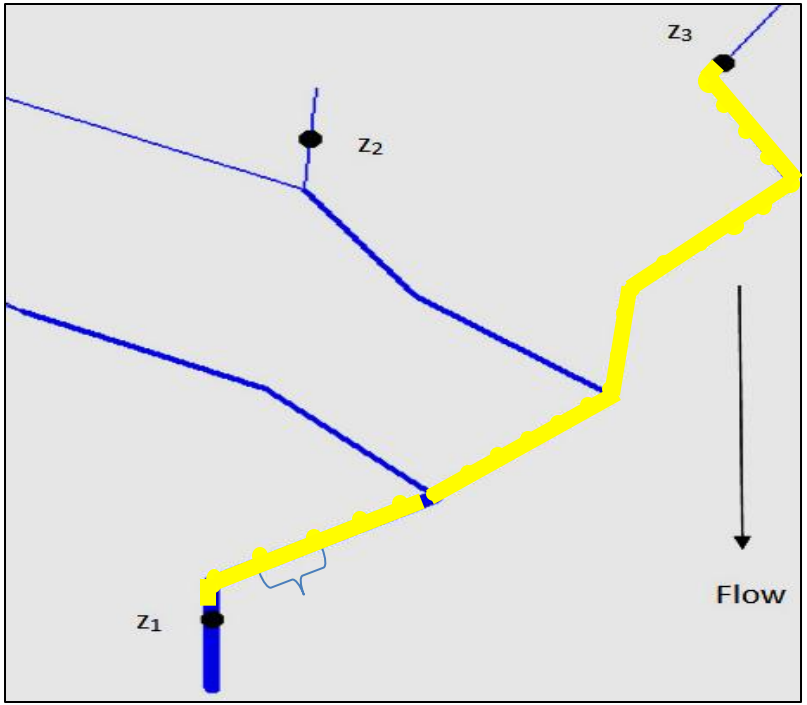


Fig. 3. River network

Methodology

However, considering only the next nearest neighbor of a station will be biased, as by chance, only the first nearest neighbor could be closed, but a second neighbour is far away.

Thus, we modified the Eq. 1 as shown in Eq. 3:

$$\chi^2_{p;\alpha(i)}\left(\mathrm{MD}^2\left(W_{z_i}\right)\right) = \mathrm{MD}^2\left(W_{z_i}, W_{z_{(\lceil k\cdot\beta\rceil)}}\right) \text{ for } i = 1, \ldots, n. \qquad (3)$$

k : Number of neighbors
β : A fraction of neighbors
$\alpha(i)$: Degree of isolation

$\alpha(i) \geq \beta \rightarrow Z_i$ is considered as potential spatial outlier

Results

Table 1. Descriptive statistics of the water quality parameters.

	DO	BOD	COD	TSS	pH	NH3NL	Temp
Min	2.635	3.833	12.83	9.16	7.116	0.063	25.92
1st Quartile	3.541	5.000	16.58	33.08	7.382	1.038	28.86
Median	4.798	7.333	22.50	38.66	7.410	4.475	29.26
Mean	5.097	8.358	25.46	64.52	7.439	4.019	29.16
3rd Quartile	6.470	9.683	29.60	71.00	7.486	6.126	29.64
Max	8.780	24.333	57.00	228.50	7.759	9.522	30.83

Table 2. DOE Water Quslity Index Classification

PARAMETER	UNIT	CLASS				
		I	II	III	IV	V
Ammoniacal Nitrogen	mg/l	< 0.1	0.1 - 0.3	0.3 - 0.9	0.9 - 2.7	> 2.7
Biochemical Oxygen Demand	mg/l	< 1	1 - 3	3 - 6	6 - 12	> 12
Chemical Oxygen Demand	mg/l	< 10	10 - 25	25 - 50	50 - 100	> 100
Dissolved Oxygen	mg/l	> 7	5 - 7	3 - 5	1 - 3	< 1
pH	-	> 7	6 - 7	5 - 6	< 5	> 5
Total Suspenbed Solid	mg/l	< 25	25 - 50	50 - 150	150 - 300	> 300
Water Quality Index (WQI)	-	< 92.7	76.5 - 92.7	51.9 - 76.5	31.0 - 51.9	> 31.0

Result

Result from *SSN* package in R

Table 3. The neighboring stations

Station	Neighboring Stations
1	10, 2, 9,3,4,5,11,12,13,7,6,14,15
2	3,4,5,1,11,12,13,7,6,14,15
3	4,2,5,12,1,13,7,6,14,15
4	5,3,2,12,13,1,7,6,14,15
5	4,3,2,1,7,6,14,15
6	7,14,15,5,4,3,2,1
7	6,14,15,5,4,3,2,1
8	NA
9	10,1
10	1,9
11	2,1
12	4,13,3,2,1
13	12,4,3,2,1
14	15,6,7,5,4,3,2,1
15	14,6,7,5,4,3,2,1

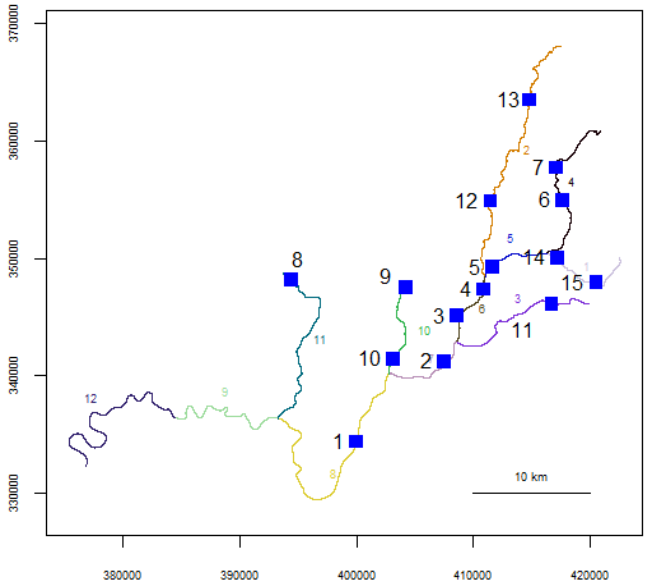


Fig. 4. The location of stations in river network

Results

Table 4. The degree of isolations and WQI of each station

Stations	River	Degree of isolation, $\alpha(i)$ (%)	WQI
1	Klang	13.35*	55
2	Klang	4.86	55
3	Klang	7.03	63
4	Klang	1.53	60
5	Klang	1.26	61
6	Klang	11.98*	89
7	Klang	3.11	76
8	Damansara	NA	89
9	Penchala	91.14*	86
10	Kerayong	39.02*	57
11	Kerayong	10.69*	42
12	Gombak	13.90*	60
13	Gombak	17.41*	87
14	Ampang	18.20*	58
15	Ampang	6.62	65

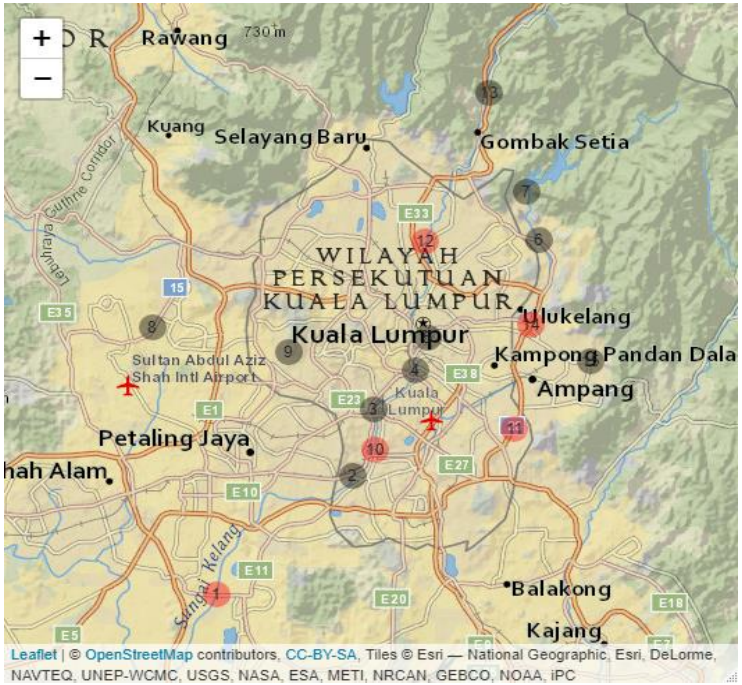


Fig. 5. The location of the stations. Red: Spatial outliers and moderate to bad water quality; Black: Regular observation

Discussion and Conclusion

- Integrating a stream distance into the detection method allows us to correctly identify spatial outliers within a river network.
- However, to find the reasons for the spatial outlier requires much more detailed studies.
- The method successfully detects the most polluted area, which is located at station 11 (Sg. Kerayong).
- Sg. Kerayong experienced bad water quality compared to its neighbors, and some effort should be made to preserve the good water quality in this area.
- More progress could be made by delving deeper into the data with higher dimensions.

References

Glińska-Lewczuk, K., Gołaś, I., Koc, J., Gotkowska-Płachta, A., Harnisz, M., & Rochwerger, A. (2016). The impact of urban areas on the water quality gradient along a lowland river. *Environmental monitoring and assessment*, *188*(11), 1-15.

Liu, Y., Li, H., Cui, G., & Cao, Y. (2020). Water quality attribution and simulation of non-point source pollution load flux in the Hulan River basin. *Scientific Reports*, *10*(1), 1-15.

Peter Filzmoser, Anne Ruiz-Gazen, and Christine Thomas-Agnan. Identification of local multivariate outliers. *Statistical Papers*, 55(1):29-47, 2014.

Xu, W., Gao, H., Liu, Y., & Li, L. (2017, July). An adaptive spatial outlier detection algorithm with no parameter for WSN. In *2017 20th International Conference on Information Fusion (Fusion)* (pp. 1-8). IEEE.

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

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

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

Ali, N. F. M., Yunus, R. M., Mohamed, I., & Othman, F. (2022). Improved Spatial Outlier Detection Method Within a River Network (Kaedah Pengesanan Pencilan Reruang DiPerbaik dalam Suatu Jaringan Sungai). *Sains Malaysiana*, *51*(3), 911-927.

Appelhans, T., Detsch, F., Reudenbach, C., & Woellauer, S. (2016, April). mapview-Interactive viewing of spatial data in R. In *EGU General Assembly Conference Abstracts* (pp. EPSC2016-1832).

Plant, R. E. (2018). Interactive maps with the leaflet and mapview packages.

THANK YOU