ASM Sc. J., 11, Special Issue 3, 2018 for SANREM, 40-47

Comparison of Spatial Pattern Analysis in Profiling Air Pollution Phenomena in Peninsular Malaysia

Siti Hajar Ya'acob[1*] and Abdul Hamid Mar Iman[1]

[1]*Sustainable Science Program, Faculty of Earth Science, Universiti Malaysia Kelantan,Jeli Campus, Locked Bag No.100, 17600 Jeli, Kelantan, Malaysia*

This research attempts to apply spatial statistics by employing GIS in providing evidence of underlying spatial contribution to potentially air pollutants concentration accumulation. We perform profiling of the air pollutants within 16 years observation (2000-2015) from 37 fixed monitoring stations in Peninsular Malaysia using ArcGIS software version 10.5. Kriging interpolation model present as the best model as compared to IDW based on the RMSE value that closest to 1. The RMSE value from kriging model for PM10, SO2 and O3 were 7.8096, 0.015 and 0.0028 respectively. Only SO2, NO2, and CO showed significant of Z-score from Getis-Ord general G, and Moran's I calculation ($p < 0.05$). The initial profiling was able to identify relevant patterns that show crucial spatial characteristics. Hence the profiled pattern is used in further analysis such as determination a hotspot, or cold spots that should be prioritized for future prediction of air pollution impact.

Keywords: Spatial autocorrelation, spatial pattern, air pollution, GIS

I. INTRODUCTION

Air pollution is an environmental problem that has its long history across the global. Rapid urban development alongside with the shift from fuelwood to coal and then to oil as fuel consumption plays a vital role in expansion dramatically the air pollution problems (Mosley, 2014). The air pollution problem leads to scholarly debate and public concern as they potentially will degrade human's health. Previous air pollution research, particularly in Malaysia, has a deal with various aspects, however, employing spatial concept using Geographical Information System (GIS) in dealing with air pollution phenomena is insufficient. Amongst the advantages of applying GIS in this field of study is it can portray the spatial correspondence of air pollutants dispersion and directly explore the potential exposure pathways in space and time (Yerramilli *et al.*, 2011). This study aims to get the best interpolation model that suitably used for air pollution database and to analyze primarily the spatial pattern of air pollution phenomena.

II. DATA AND METHODS

1. Study Area

There are 37 Continuous Air Quality Monitoring (CAQM) stations located

*Corresponding author's e-mail: hajar.y@umk.edu.my

throughout Peninsular Malaysia enforced by Department of Environment (DOE), Malaysia (Figure 1). In this research, all the 37 locations of CAQM stations were selected to represent air pollution exposure in Peninsular Malaysia. These stations are operated effectively until April 2017 by Alam Sekitar Malaysia Sdn. Bhd (ASMA), the subsidiary company of the DOE.

2. *Air pollutants and meteorological data sources*

This research uses a time-series approach to associate the data components starting from 2000 until the year 2015. A dataset comprising of air pollutants (PM10, O3, NO2, SO2, and CO) and meteorological parameters (wind speed, temperature, and humidity) at the 37 stations were collected from DOE prior spatially and statistically analyzed in ArcGIS version 10.5. All received data were initially in the form of hourly before being converted to a yearly average value. Hence, a total of 296 data sets (37 observations x 8 parameters) were prepared in a spreadsheet and spatially referenced using each location's longitude and latitude. The coordinates were projected in the World Grid System of 1984 (WGS84) during the process of converting the spreadsheet data into GIS.

3. *Assessment of different spatial interpolation model*

The collected dataset in this research is insufficient to estimate unknown concentration located in between the monitoring stations. Therefore, spatial interpolation was used in this research to predict the unknown value at other location than sampled point values. A comparison of two types of interpolation model has initially performed namely Inverse Distance Weighting (IDW) and Kriging. IDW is a deterministic method while Kriging is a geostatistical method. There were three sub-group of Kriging namely ordinary kriging (OK), simple kriging (SK) and universal kriging(UK). Either one of them are weighted average model hence they have the same underlying mathematical formulation as following (Wong *et al.*, 2004):

$$z(x_0) = \sum_{i=1}^{n} \lambda_i . z(x_i) 0 \text{ and } \sum_{i=1}^{n} \lambda_i = 1 \quad (1)$$

where z is air pollution concentration at an unsampled point xo, zi is a set of neighboring sampled values that sampled at location xi, λ represents the weight assigned to each neighboring values, and the sum of the weight is one. In IDW, the weight λ depends solely on the distance to the prediction location. This research uses the weight or power of 2 which is the default value in the ArcGIS. To select the best interpolation model, we then performed a cross-validation test.

4. *Spatial autocorrelation methods*

Two types of spatial autocorrelation were chosen to test the spatially clustered tendency (Scott and Janikas, 2010) of the selected air

pollutants namely Getis-Ord general G and global Moran's I. Both methods are an inferential statistic; hence the null hypothesis is the values associated with features are randomly distributed. A confident level of 95% was selected and p-value < 0.05 was considered as statistically significant clustering. General G statistic of overall spatial association is as following equation:

$$G = \frac{\sum_{i=1}^{n}\sum_{i=j}^{n} w_i x_i x_j}{\sum_{i=i}^{n}\sum_{j=1}^{n} x_i x_j}, \ \forall_{j \neq i} \quad (2)$$

where X_i and X_j are concentrations of pollutants for features i and j, $W_{i,j}$ is the spatial weight between i and j, n is the total number of features in the dataset and □j ≠ i indicates that feature i and j cannot be the same features. Also, the second type of spatial autocorrelation, global Moran's I was calculated using formula as follow:

$$I_t = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x)(x_j - x)}{\sum_{i=1}^{n}(x_i - x)} \quad (3)$$

$$S_0 = \sum_{i=i}^{n}\sum_{j=1}^{n} w_{ij} \quad (4)$$

where $(X_i - X)$ is the deviation of feature i from its mean of concentration, W_{ij} is the spatial weight between the feature i and j, N is equal to the total number of features, and S_o is the aggregate of all spatial weight. The values of both indices that show higher than 0 refer to positive spatial correlation while the values less than 0 indicates a negative correlation and 0 values mean insignificant (Habibi *et al.*, 2017).

III. RESULT AND DISCUSSION

1. *Interpolation calculation output*

The cross-validation approach was mainly to verify best interpolation model before being implied for further data analysis. We have chosen the RMSE as the primary criterion for best selection of interpolation results. Table 1 depicted the statistic errors for each tested model. Results showed that Kriging was a better model compared to IDW generally for all five air pollutants. The RMSE sorting for PM10, SO2, and O3 was a similar trend: SK<OK=UK<IDW. In contradiction, RMSE sorting for NO2 and CO was on different direction: OK=UK<IDW<SK. By combining the overall result, SK and OK/UK was the best model for three pollutants (namely PM10, SO2, and O3) and the rest of two pollutants (NO2 and CO) respectively. The best model selection was based on RMSE parameters that meet the principle of RMSE should close to 1 (Xiao *et al.*, 2016). A series of studies of testing and modeling air quality across Europe and the US also found that kriging method is generally preferred over IDW for PM10 data (Wong *et al.*, 2004; Horalek *et al.*, 2007). The best interpolation method preference is highly depended to the nature and availability of data from existing monitoring network (Wong *et al.*,

2004) apart from consideration of another selection criterion such as spatial coverage quality, continuity, and robustness (Horalek *et al.*, 2007). Kriging is known as a geostatistical method unlike IDW is a deterministic method. The latter interpolation method uses assumptions that closer monitoring station has more similar nature of data value than monitoring station that is farther apart. Unlike for the geostatistical method (kriging), it assumes all values are the result of the random process with dependence (Johnston *et al.*, 2001). Due to both interpolation methods have their own strong and weakness as well as considering the RMSE value, we finally choose kriging method as the best interpolation method that suitable with our air quality data.

2. Spatial pattern analysis

Analysis of spatial pattern in this research was performed to identify spatial clustering or dispersion of the five types of air pollutants across Peninsular Malaysia. We compute the pattern analysis using two different tools/methods namely Getis-Ord general G and global Moran's I. The selected spatial pattern tools generate a value known as z-score which describe the degree of spatial dispersion or concentration for the air pollutants variables. Table 2 showed the Z-score value for global Moran's I and Getis-Ord general G. All p-values for Getis-Ord general G were significant ($p < 0.05$) for 16 years PM10, SO2, NO2, and CO except for O3. Therefore, the null hypothesis is rejected with 95% confident level, resulted in significant clustering for the observed of PM10, SO2, NO2, and CO. Statistical tests for the Getis-Ord general G also showed that 16 years O3 and NO2 has minimum and maximum Z-score value of 0.33 and 5.38 respectively.

On the other hand, the Moran's I result depicted significant p-values for 16 years average of SO2, NO2, and CO only. Similarly, with Getis, it is possible to reject the null hypothesis which leads to further investigation needed to find the cause of significant spatially related to air pollution in particular for SO2, NO2 and CO phenomena. Highest Z-score value of 16 years analyzed through Moran's I occurred for NO2, followed by CO, SO2, and PM10 whereas Z-score for O3 was in a negative value. Having that pattern analysis of 16 years air pollutants, the Z-score values of Getis were higher than Moran's I. Our finding was in contrast with previous research that reported Z-score for Moran's I were larger as compared to Getis (Habibi *et al.*, 2017). However, the previous study used IDW as their interpolation method unlike in the current study we employed kriging as the interpolation method due to smaller RMSE result in cross-validation test.

Looking at the overall 16 years of spatial pattern analysis, a low negative Z-score either in Getis or Moran's I method indicates a significant data outlier spatially. In addition, the results showed a minimum of outliers for SO2, NO2, and CO within the 16 years dataset. Moreover, the high Z-score values for SO2,

NO2, and CO suggested that such pollutants are highly and spatially clustered in Peninsular Malaysia. Theoretically, sources of outdoor air pollutants are associated with number of vehicles, level of urbanization, industrialization as well as trans boundaries pollution. A major development in Peninsular Malaysia is rapidly driven by industrial sector that interrelated with an increment of vehicle numbers and urban population. Types of air pollutants that emitted from such human activities vary with notably NO2 concentration is coming from motor vehicles emission (Awang *et al.*, 2000). Hence, our results are showing that emission of NO2 is highly clustered at a specific location where vehicles are the primary emission sources. The subsequent phase of this study will assess quantitatively the spatial association between different air pollutants clustering with their potential emission sources.

Moreover, Table 2 shows positive correlation value for the contaminants (95% confidence level) which demonstrate that locations with the relatively high level of pollutants are close together, the similar pattern with contaminants of low level (Moore & Carpenter, 1999). Table 2 also showed seasonal Z-score values for PM10, SO2, NO2, CO, and O3. Only SO2 has a significant p-value for both Getis and Moran's I value ($p<0.05$) during wet, dry and inter-season. It can be seen that in Figure 2, a similar trend for seasonal Z-score between Getis and Moran's I value, however, Getis Z-score was still more significant than Moran's I.

Table 1. Statistical errors of the compared interpolation models

Statistical error		IDW	OK	SK	UK
PM_{10}	Mean	1.29	0.56	0.45	0.56
	RMSE	9.37	7.84	7.80	7.84
SO_2	Mean	0.00	-0.00	0.00	-0.00
	RMSE	0.00	0.00	0.00	0.00
NO_2	Mean	0.00	0.00	0.00	0.00
	RMSE	0.00	0.00	0.01	0.00
CO	Mean	0.04	-0.00	0.06	-0.00
	RMSE	0.20	0.19	0.48	0.19
O_3	Mean	-0.00	-0.00	0.00	-0.00
	RMSE	0.00	0.00	0.00	0.00

IDW=Inverse Distance Weighting, OK=Ordinary Kriging, SK=Simple Kriging, UK=Universal Kriging

Table 2. Comparison between global Moran's I and Getis-Ord General G for the five pollutants during the different period

		Wet	Dry	Inter	16 years average
PM_{10}	Getis Z-score	-0.17	-0.67	-0.74	2.51*
	Moran's I Index	0.213	0.218	0.238	0.119
	Moran's Z-score	1.56	1.59	1.71	0.95
SO_2	Getis Z-score	4.67*	3.15*	3.48*	3.91*
	Moran's I Index	0.505	0.339	0.336	0.413
	Moran's Z-score	3.47*	2.42*	2.36*	2.86*
NO_2	Getis Z-score	-0.8	-0.82	-0.86	5.38*
	Moran's I Index	-0.159	-0.118	-0.134	0.768
	Moran's Z-score	-0.86	-0.58	-0.69	5.23*
CO	Getis Z-score	-0.56	-0.51	-0.51	4.02*
	Moran's I Index	-0.134	-0.14	-0.109	0.561
	Moran's Z-score	-0.71	-0.75	-0.54	3.95*
O_3	Getis Z-score	-0.43	-0.17	-0.88	0.33
	Moran's I Index	0	0.071	0.103	-0.089
	Moran's I Z-score	0.18	0.63	0.85	-0.39

*p-value < 0.05

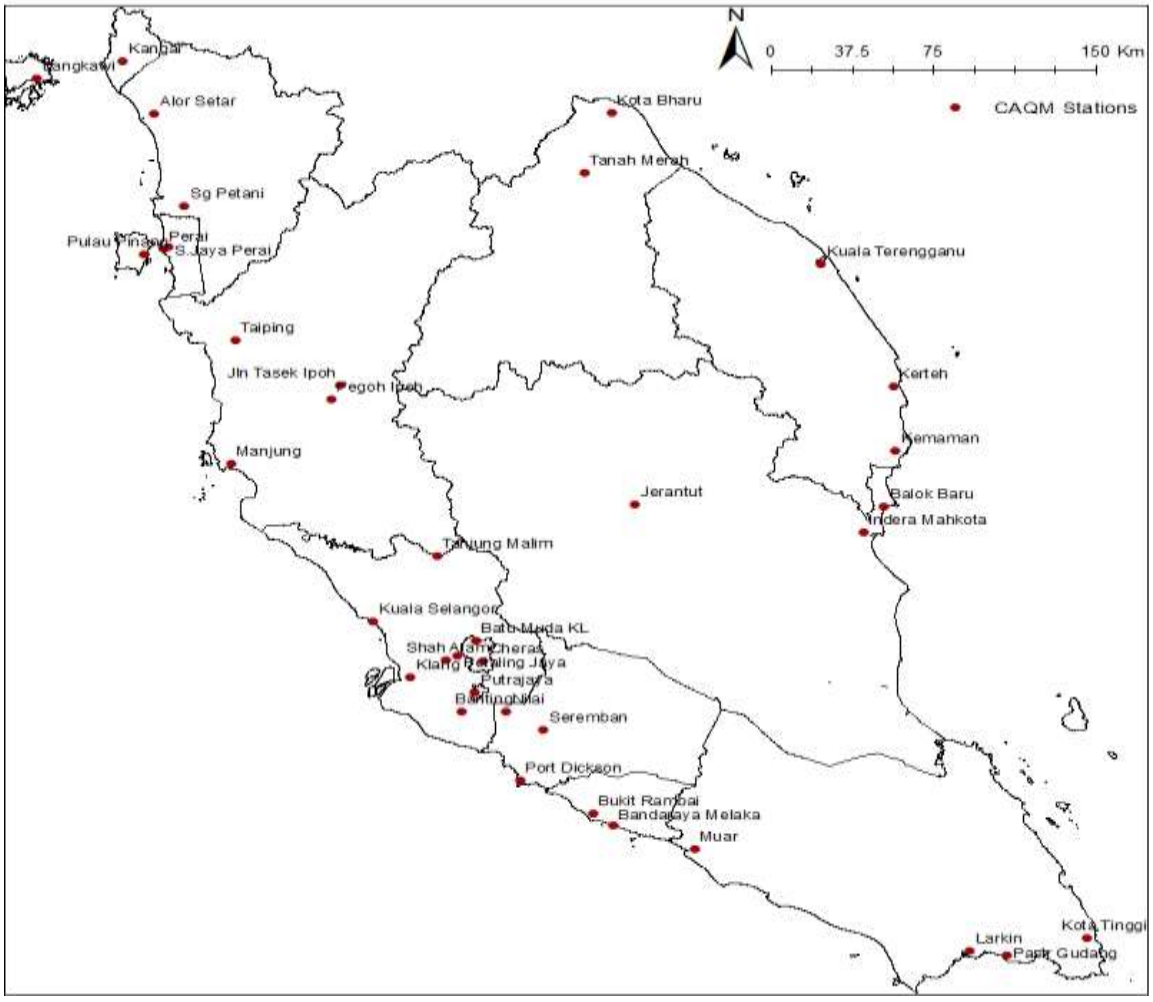

Figure 1. Continuous Air Quality Monitoring (CAQM) stations

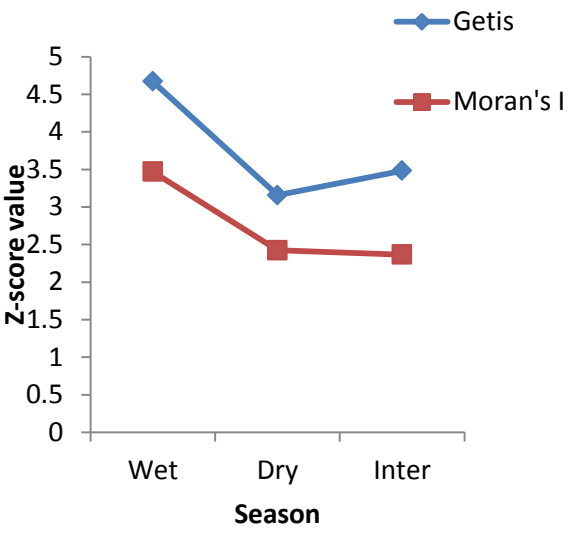

Figure 2. Seasonal Z-score values for SO2

IV. CONCLUSION

A GIS-based air pollution study in Peninsular Malaysia during the period of 2000-2015 is reported in this paper. In the end, this study seeks the best interpolation method for primary data before being used in the further geo-statistical analysis. Kriging is chosen than IDW after cross-validation test was performed and the value of root mean square error was used as the selection basis. This study also suggests that Getis-Ord general G as a better pattern analysis after considering the z-score value that consistently significant. This research finding will be used as initial consideration for subsequent air pollution phenomena profiling.

V. ACKNOWLEDGEMENTS

The authors appreciate the Ministry of Higher Education Malaysia for funding this research under grants R/FRGS/A08.00/01228A/001/2016/000373. We also would like to thank Department of Environment Malaysia for providing air quality and meteorological data in this research.

[1] Awang, M., Jaafar, A. B., Abdullah, A. M., Ismail, M., Hassan, M. N., Abdullah, R., Johan, S. & Noor, H. (2000). Air quality in Malaysia:impacts, management issues and future challenges', Respirology. 5, 183-196.

[2] Habibi, R., Alesheikh, A., Mohammadinia, A. & Sharif, M. (2017). An assessment of spatial pattern characterization of air pollution: A case study of CO and PM2.5 in Tehran, Iran. ISPRS International Journal of Geo-information. 6 (9), 270.

[3] Horalek, J., Denby, B., Smet, P. D, de Leeuw, F., Swart, R. & van Noije, T. (2007). Spatial mapping of air quality for European scale assessment, European Topic Centre on Air and Climate Change.ETC/ACC Technical Paper 2006/6.

[4] Johnston, K., Hoef, J. M. V., Krivoruchko, K. & Lucas, N. (2001). Using ArcGIS geostatistical analyst, ESRI.

[5] Moore, D. & Carpenter, T. (1999). Spatial analytical methods and geographic information systems:use in health research and epidemiology. Epidemiologic Reviews, 30(2), 143-161.

[6] Mosley, S. (2014). Environmental history of air pollution and protection. World environmental history, EOLSS.

[7] Scott, L. M. & Janikas, M. V. (2010). Spatial statistics in ArcGIS', eds Getis, A. & Fischer, M. in Handbook of applied spatial analysis: Software tools, methods and applications, Springer, 27-41.

[8] Wong, D. W., Yuan, L. & Perlin, S. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. Journal of Exposure Analysis and Environmental Epidemiology. 14(5), 404-415.

[9] Xiao, Y., Gu, X., Yin, S., Shao, J., Cui, Y., Zhang, Q. & Niu, Y. (2016). Geostatistical interpolation model selection based on ArcGIS and spatio-temporal variability analysis of groundwater level in piedmont plains, northwest China. SpringerPlus, 5(1), 425.

[10] Yerramili, A., DodlaV. B. R. & Yerramili, S. (2011). Air pollution, modeling and GIS based decision support systems for air quality risk assessment. Ed Farhad Nejadkoorki, in Advanced air pollution, Intechopen, 295-324.