Sains Malaysiana 47(3)(2018): 471–479
http://dx.doi.org/10.17576/jsm-2018-4703-06

Applied Chemometric Approach in Identification Sources of Air Quality Pattern in Selangor, Malaysia

(Aplikasi Pendekatan Kimometriks dalam Mengenal Pasti Corak Sumber Kualiti Udara di Selangor, Malaysia)

ANG KEAN HUA*

ABSTRACT

In recent years, Malaysia has experienced quite a few number of chronic air pollution problems and it has become a major contributor to the deterioration of human health and ecosystems. This study aimed to assess the air quality data and identify the pattern of air pollution sources using chemometric analysis through hierarchical cluster analysis (HCA), discriminant analysis (DA), principal component analysis (PCA) and multiple linear regression analysis (MLR). The air quality data from January 2016 until December 2016 was obtained from the Department of Environment Malaysia. Air quality data from eight sampling stations in Selangor include the selected variables of nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO) and particulate matter (PM_{10}). The HCA resulted in three clusters, namely low pollution source (LPS), moderate pollution source (MPS) and slightly high pollution source (SHPS). Meanwhile, DA resulted in two and four variables for the forward stepwise mode and the backward stepwise mode, respectively. Through PCA, it was identified that the main pollutants of LPS, MPS and SHPS came from industrial and vehicle emissions, agricultural systems, residential factors and natural emission sources. Among the three models yielded from the MLR analysis, it was found that SHPS is the most suitable model to be used for the prediction of Air Pollution Index. This study concluded that a clearer review and practical design of air quality monitoring network would be beneficial for better management of air pollution. The study also suggested that chemometric techniques have the ability to show significant information on spatial variability for large and complex air quality data.

Keywords: Discriminant analysis; hierarchical cluster analysis; multiple linear regression analysis; principal component analysis

ABSTRAK

Sejak beberapa tahun kebelakangan ini, Malaysia telah mengalami beberapa masalah pencemaran udara yang kronik dan ia telah menjadi salah satu penyebab utama dalam kemerosotan kesihatan manusia dan ekosistem. Matlamat kajian ini adalah untuk menilai data kualiti udara dan mengenal pasti corak sumber pencemaran udara menggunakan teknik kimometriks melalui analisis pengkelasan hierarki (HCA), analisis diskriminan (DA), analisis komponen berprinsip (PCA) dan analisis regrasi linear pelbagai (MLR). Data kualiti udara bermula dari Januari hingga Disember 2016 telah diperoleh daripada Jabatan Alam Sekitar Malaysia. Data kualiti udara dari lapan stesen persampelan di Negeri Selangor melibatkan pemboleh ubah terpilih nitrogen dioksida (NO_2), ozon (O_3), sulfur dioksida (SO_2), karbon monoksida (CO) dan zarahan terampai (PM_{10}). Proses HCA telah menghasilkan tiga kluster iaitu sumber pencemar rendah (LPS), sumber pencemar sederhana (MPS) dan sumber pencemar sedikit tinggi (SHPS). Sementara itu, proses DA telah menghasilkan dua pemboleh ubah bagi mod ke hadapan berperingkat dan empat pemboleh ubah bagi mod ikut langkah kebelakang. Melalui proses PCA, ia telah dikenal pasti bahawa bahan pencemar utama bagi LPS, MPS dan SHPS berasal daripada hasil pelepasan industri dan pengangkutan, sistem pertanian, faktor kediaman dan sumber pelepasan semula jadi. Antara ketiga-tiga model yang dihasilkan melalui analisis MLR, ia didapati bahawa SHPS adalah model yang paling sesuai untuk digunakan bagi kerja-kerja ramalan Indeks Pencemaran Udara. Kajian ini menyimpulkan bahawa ulasan serta reka bentuk praktikal ke atas rangkaian pengawasan kualiti udara akan memberi manfaat dalam usaha mengurus pencemaran udara dengan lebih baik. Kajian ini juga mencadangkan bahawa teknik kimometriks mempunyai keupayaan untuk mendedahkan maklumat yang penting tentang pemboleh ubah reruang bagi data kualiti udara yang besar dan rumit.

Kata kunci: Analisis diskriminan; analisis komponen berprinsip; analisis pengkelasan hierarki; analisis regrasi linear pelbagai

INTRODUCTION

Air pollution is an important factor that could influence the quality of life and it requires serious and immediate attention in both developed and developing countries. Air pollution can be explained as a circumstance where air pollutants concentration in the atmosphere exceeded the normal ambient levels (Seinfeld & Pandis 1998). The types of pollutants can be irregular depending on the time

period and it can also be in either scattered or concentrated form in the atmosphere. Air pollution problems often occurred in highly populated focus area such as urban and manufacturing industrial areas (Azid et al. 2013). This condition is not unfamiliar to a developing country such as Malaysia. Malaysia's experiences with industrial pollution and urban environmental degradation can be credited to the rapid economic growth which started when it aimed to achieve the status of developed country by the year 2020. Nowadays, air pollution has become a major issue of debate because of its negative influence on humans, buildings, crops, and the ecosystems (Moustris et al. 2010). Continuous exposure to air pollution would threaten the wellbeing of public health and this condition requires close government monitoring especially at certain areas as to prevent the deterioration of air quality.

In Malaysia, the main sources of air pollutants came from mobile, stationary and trans-boundary sources (Azid et al. 2014, 2013; Khan et al. 2015; Makmom et al. 2012; Mutalib et al. 2013; Sulong et al. 2017). Generally, mobile source pollution referred to any air pollution emitted by motor vehicles (Azid et al. 2014). Meanwhile, stationary source pollution originated from fossil fuel burning power plants, food processing plants, heavy industrial sources and open burning (Dominick et al. 2012). On the other hand, trans-boundary pollution comprises of forest burning or volcanic eruptions in neighboring countries which causes air pollution to the home ground (Makmom et al. 2012). The air quality status in Malaysia is measured using Malaysia Ambient Air Quality Standard (MAAQS) which was established by the Department of Environment (DOE), Malaysia. Air Pollutant Index (API) is calculated based on sub-index of SO_2, NO_2, CO, O_3, and PM_{10}; where only the highest value of sub-index in individual pollutants are taken as the API value (DOE 2012). Normally, PM_{10} and O_3 were detected as major air pollutants in urban cities and sub urban areas and they both have quite an influenced on human health in the country (Mahiyuddin et al. 2013). This condition of air pollution was particularly influenced by the high traffic volume and industrial activities (Azmi et al. 2010).

Chemometric techniques also known as multivariate techniques analysis is an excellent tool that is often used in the environmental field to identify the sources of pollution (Azid et al. 2014; Mutalib et al. 2013). Chemometric analysis include the interrelationship of faunal structure, physic-chemical and biological characteristic, as well as toxicity data that could be obtain from the laboratory analysis (Azid et al. 2015). This made the tool suitable for the reduction and interpretation of meaningful data (Azid et al. 2015; Mutalib et al. 2013) through hierarchical cluster analysis (HCA), discriminant analysis (DA), principal component analysis (PCA) and multiple linear regressions (MLR). Chemometric techniques is not only beneficial for the recognition of potential sources variations in air quality and manipulation of air quality, but it is also beneficial for the interpretation of complex databases for better understanding of the condition of air quality in a specific region (Azid et al. 2015; Hua et al. 2016; Mutalib et al. 2013). Therefore, these methods are appropriate for the development of efficient management of air quality monitoring network (Azid et al. 2015).

This study was carried out to identify the spatial and temporal variations of air quality parameters using chemometric techniques and to determine the origin of pollution sources in Selangor, Malaysia. Specifically, the objective of this study were to identify the level of air quality in Selangor by recognizing the pollution source, to determine the most significant air quality variable and to produce a model for the prediction of air quality performance in Selangor.

Materials and Methods

Study Area

Selangor is located in the western part of Malaysia, lying within the latitude of 2°35'23.53"N to 3°47'55.09"N and longitude of 100°56'25.09"E to 101°57'58.50"E (Table 1 & Figure 1). Selangor has an approximate square area of around 8104 km^2 and it has an estimated number of populations at about five million people. While the west side of Selangor is facing the Straits of Malacca, Selangor shares its terrestrial boundary with Perak along the north, Pahang in the east and Negeri Sembilan along the south. Malaysia can be considered as a country which is free from natural disaster such as typhoon, volcanic eruption and earthquake; and this has helped in keeping the air quality under control. Nevertheless, the rapid economic growth experienced by the state and the country has become an aggravating factor towards the air quality level and hence,

TABLE 1. Geographical Coordinate of 8 monitoring stations details in Selangor

Station ID	Location	Latitude	Longitude
Station 1	Klang, Selangor	3° 0'53.72"N	101°24'47.02"E
Station 2	Petaling Jaya, Selangor	3° 7'59.37"N	101°36'28.53"E
Station 3	Shah Alam, Selangor	3° 6'17.03"N	101°33'21.66"E
Station 4	Kuala Selangor, Selangor	3°19'16.13"N	101°15'22.61"E
Station 5	Putrajaya, Wilayah Persekutuan	2°54'52.49"N	101°41'23.69"E
Station 6	Cheras, Kuala Lumpur	3° 6'22.62"N	101°43'5.00"E
Station 7	Batu Muda, Kuala Lumpur	3°12'45.08"N	101°40'56.47"E
Station 8	Banting, Selangor	2°48'59.98"N	101°37'23.21"E

this study would assist in determining the latest status of air quality in Selangor.

DATA COLLECTION

Air quality data was retrieved from DOE, Malaysia, starting from January 2016 to December 2016. The variable of pollutants selected for the study such as nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO), and particulate matter (PM_{10}) were used to evaluate the API status and determine the pollution sources. Generally, the air quality assessment was done at the eight sampling stations in Selangor (Figure 1), where majority of the stations are located at urban, suburban and industrial areas. The statistical data used in this study consisted of 96 dataset (12 data per stations × 8 stations) and a total number of 480 observations (12 data per stations × 8 stations × 5 variables). All data were obtained from a monthly average that was established from the hourly monitoring sites.

CHEMOMETRIC ANALYSIS

HIERARCHICAL CLUSTER ANALYSIS (HCA)

HCA uses an unsupervised pattern method to identify the sources, by splitting a large group of data into smaller ones based on their similarities and produced the analysis in the form of 'cluster' (Azid et al. 2015). HCA involved with several procedures (Aris et al. 2013; Azid et al. 2015; Juahir et al. 2011; Mahiyuddin et al. 2013), namely; Ward's method using variance analysis to determine the distance between two clusters by minimizing the sum of squares (SS) from each step; Euclidean distance to determine the similarity between two samples and a distance to characterized the differences between analytical values from the samples, which can be defined by (1):

$$d(x,y) = \sum_{m-1}^{p}(x_m - y_m) \quad (1)$$

where $d(x,y)$ is the Euclidean distance between two samples represented in x_m and y_m; and p is the dimensional space of the variables (Azid et al. 2015); and dendogram that shows the high similarity in small distances between clusters and dissimilarity of possible distances between clusters.

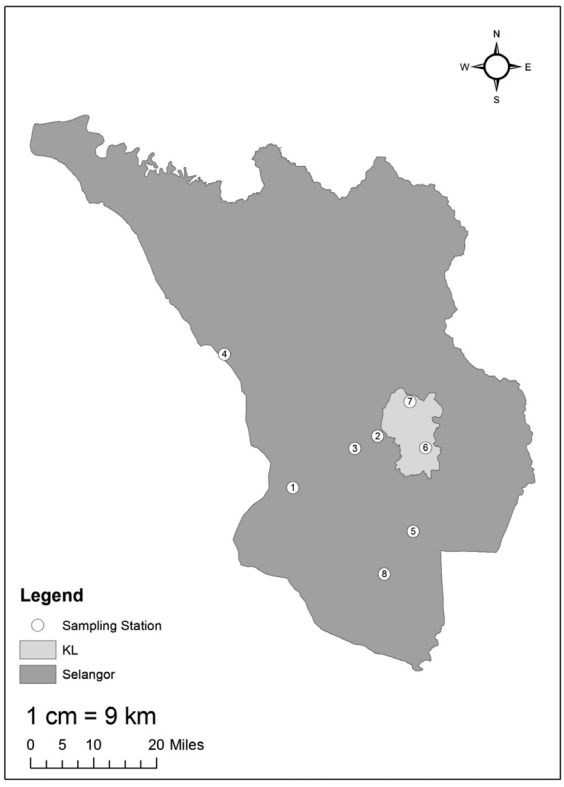

FIGURE 1. Study Area in Selangor, Malaysia

DISCRIMINANT ANALYSIS (DA)

DA is usually applied to evaluate an object of unknown origin to one of several naturally occurring groups (Manjunanth et al. 2012). This study applies DA together with HCA to establish significantly different variables and reduce the errors of these groups (Aris et al. 2013; Azid et al. 2015; Juahir et al. 2011). Every cluster from HCA will create discriminant function (DF) in DA, where the DF can be defined in (2):

$$f(G_i) = k_i + \sum_{j=1}^{n} W_{ij}P_{ij} \quad (2)$$

where i is the number of groups (G); k_i is the constant to each group; n is the number of parameters used to classify a set of data into a given group; and w_{ij} is the weight coefficient assigned by DF analysis to a given parameter (P_{ij}). This study applied DA in three modes, namely standard mode, forward stepwise mode and backward stepwise mode (Aris et al. 2013; Juahir et al. 2011). Standard mode would provide DFs for the evaluation of spatial variations in the air quality raw data. Meanwhile, forward stepwise mode would eliminate the variables from the most-significant to no-significant changes and backward stepwise mode would eliminate variables from the less-significant to no-significant changes.

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is used to interpret interrelated variables to create new variables, which is known as principal components (PCs) and the value are known as principal component scores (PCS). In other words, the newly maximum number is equivalent to the original number (Juahir et al. 2011). In this study, PCA is used together with HCA to recognize the emission sources by presenting the details of most-significant variables in the spatial and temporal variation and putting them with the less-significant variables, with minimum loss of the original information (Azid et al. 2015; Juahir et al. 2011). The PCA can be defined in (3):

$$Z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + \ldots + a_{im}x_{mj} \quad (3)$$

where *z* is the component score; *a* is the component loading; *x* is the measured value of the variable; *i* is the component number; *j* is the sample number; and *m* is the total number of variables. The general procedures applied in PCA are: the hypothesis obtained from the original data will be reduced to dominant factors that influence the observed data variance; and the whole data set is extracted through eigenvalues and eigenvectors from the square matric produced by multiplying the data matrix (Aris et al. 2013; Azid et al. 2015; Juahir et al. 2011). To be considered as significant, varimax factors (VFs) of new group variables were defined based on the eigenvalues that were greater than 1 (Aris et al. 2013; Juahir et al. 2011). VFs coefficient that are greater than 0.75 would be considered as 'strong', 0.75 to 0.50 as 'moderate' and 0.50 to 0.30 as 'weak' (Juahir et al. 2011). In this study, PCA was applied to classified datasets (five variables) independently based on the different spatial regions obtained from the HCA techniques.

MULTIPLE LINEAR REGRESSIONS (MLR)

MLR is widely used in atmospheric modeling (Azid et al. 2015; Dominick et al. 2012). This method is an appropriate way to investigate the relationship between independent and dependent variables through the formation of linear equation of observed data (Ul-Saufi et al. 2011) and it would also provide the percentage of atmospheric pollution on each parameter (Aertsen et al. 2010). This study applied MLR to justify the relationship between air quality parameter (the most-significant among the five parameters) with the API data. The MLR model is defined by (4):

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \varepsilon \quad (4)$$

where *Y* is the response variable; *p-1* is the explanatory variable for $x_1, x_2, \ldots, x_{p-1}$ with p is the parameter (regression coefficient) of β_0, β_1, β_2, …, β_{p-1}; and ε is the error associated with the regression.

Determination of the best fitting linear regression equation was done using the coefficient of determination (R^2), adjusted coefficient of determination (Adjusted R^2) and root mean square error (RMSE). The value of R^2 provides the information of how well the model performs using the external data (Dominick et al. 2012). Adjusted R^2 considered all possible number of variables (Mutalib et al. 2013) and RMSE measure the residual error and the mean difference between observed and modeled value of API (Azid et al. 2015). Generally, higher R^2 value (which is near to 1) will be considered as the best linear model (Azid et al. 2015; Dominick et al. 2012; Mutalib et al. 2013). Consequently, chemometric techniques analysis through HCA, DA, PCA and MLR was performed using SPSS version 23.

RESULTS AND DISCUSSION

AN OVERVIEW ON THE RECORD OF AIR QUALITY DATA

The compilation of data from eight sampling stations in Selangor can be summarized in Table 2. The result showed that the average concentration of O_3, CO, NO_2, SO_2 and PM_{10} were detected far below the value suggested by MAAQS for the average concentration (0.1 ppm for O_3, 30 ppm for CO, 0.18 ppm for NO_2, 0.15 ppm for SO_2, 120 μgm^{-3} for PM_{10}). Nevertheless, the average concentration of PM_{10} detected in most stations exceeded the value of 50 μgm^{-3} as per recommended by the European Commission for PM_{10}. Therefore, it can be said that this condition is a common occurrence in urban and suburban area, especially in areas with large number of motor vehicles, industrial areas and areas with high level of street dust, which increases the amount of suspended particulate in the atmosphere at all the stations in Selangor.

SPATIAL CLASSIFICATION BASED ON AIR QUALITY PARAMETERS

Analysis of HCA indicates the result of three clusters that were formed using the data collected from the eight

TABLE 2. Overall data on air quality at different stations in Selangor in 2016

Parameters	AT		Sampling station								MAAQS (2018)
			S1	S2	S3	S4	S5	S6	S7	S8	
O_3 (ppm)	1h	M	0.021	0.018	0.010	0.023	0.023	0.023	0.021	0.025	0.1
		SD	0.005	0.004	0.011	0.004	0.008	0.004	0.004	0.004	
CO (ppm)	1h	M	0.287	0.196	0.535	0.736	0.730	0.845	0.845	0.712	30
		SD	0.367	0.458	0.572	0.251	0.300	0.550	0.273	0.261	
NO_2 (ppm)	1h	M	0.017	0.028	0.016	0.018	0.006	0.020	0.022	0.013	0.18
		SD	0.013	0.004	0.015	0.009	0.009	0.002	0.004	0.002	
SO_2 (ppm)	1h	M	0.002	0.003	0.001	0.004	0.001	0.002	0.002	0.003	0.15
		SD	0.002	0.002	0.002	0.001	0.002	0.001	0.001	0.002	
PM_{10} (ugm^{-3})	24h	M	76.64	60.04	65.95	52.25	61.93	58.02	61.01	71.12	120
		SD	35.60	31.93	36.96	28.35	33.37	27.44	34.04	34.43	

(AT=Average Time; M=Mean; SD=Standard Deviation; S=Station; MAAQS=Malaysian Ambient Air Quality Standard; O_3=Ground Level Ozone; CO=Carbon Monoxide; NO_2=Nitrogen Dioxide; SO_2=Sulfur Dioxide; PM_{10}=Particulate Matter with the Size of less than 10 Micron)

sampling stations (Figure 2). Cluster 1 consisted of results of data from Stations 1, 3, and 8; while Cluster 2 consisted of results of data from station 2, 5, 6, and 7; and Cluster 3 consisted of results of data from station 4. To summarize, Cluster 3 have an average API value of 42 which is classified as low pollution source (LPS). Meanwhile Cluster 1 have an API average of 55 (moderate pollution source (MPS)); and Cluster 2 recorded a slightly high pollution source (SHPS) with an average API value of 78. As a result, HCA technique has shown the ability to reduce the number of monitoring stations and basically suggested the category of air quality based on the regions and this is a beneficial finding for the process of improving the monitoring network system in future.

DISCRIMINANT ANALYSIS BASED ON SPATIAL VARIATION

Further analysis using DA method were carried out based on the clustering obtained from the HCA, which is the LPS, MPS and SHPS. DA techniques comprises of three modes, namely standard, forward stepwise and backward stepwise. The analysis of DA indicated that the accuracy of spatial variation for the three modes are at 95.38% with 5 variables for standard mode, 89.05% with 2 variables for forward stepwise mode and 93.23% with 4 variables for backward stepwise mode (Table 3). In this study, null hypothesis (H_0) stated that at least one of the mean vectors is different from the others, while alternative hypothesis (H_a) stated that the mean vectors of the three classes are equal. Simultaneously, the *p*-value is lower than the significant level of alpha (0.05) and hence the null hypothesis (H_0) will be rejected and the alternative hypothesis (H_a) would be accepted instead. Since the Pillai's Trace test for standard mode, forward mode and backward mode provided the result of $p<0.0001$ with 1.321, $p<0.0001$ with 1.185 and $p<0.0001$ with 1.274, respectively, which are above 0.01%; is acceptable to discard the H_0 which is lower than 0.05% and have the same mean vectors of in the three classes. The four selected variables of air quality which showed high spatial variations (with most-significant *p*-value of less than 0.05) for the backward stepwise mode were applied into the box and whisker plots for further discussion (Figure 3).

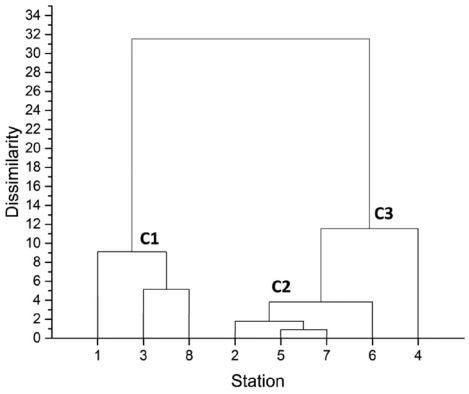

FIGURE 2. HCA using Ward linkage method to generate dendogram

IDENTIFICATION SOURCE OF VARIATION

PCA were applied to the air quality data to determine the pattern of air quality variables and further identify the factor based on the discovery regions (LPS, MPS and SHPS). As shown in Table 4, the result indicated that two VFs were obtained in the three regions with the eigenvalue of higher than 1. Meanwhile, the total variance for LPS, MPS and SHPS regions are at 66.27%, 66.39% and 62.26%, respectively.

Low pollution source (LPS) region LPS region shows that VF1 contributed about 36.32% of the total variance to produce strong negative loadings of CO and moderate

TABLE 3. Classification matrix of DA for spatial variation in Selangor

Sampling Regions	% Correct	Regions assigned by the DA		
		Cluster 1	Cluster 2	Cluster 3
Standard Stepwise				
Cluster 1	94.29%	70	0	3
Cluster 2	91.67%	1	13	0
Cluster 3	93.24%	6	0	78
Total	95.38%	77	13	81
Forward Stepwise				
Cluster 1	91.57%	65	0	2
Cluster 2	90.67%	3	11	0
Cluster 3	96.43%	2	0	82
Total	89.05%	70	11	84
Backward Stepwise				
Cluster 1	93.06%	69	0	5
Cluster 2	91.37%	2	12	0
Cluster 3	95.42%	4	0	80
Total	93.23%	75	12	85

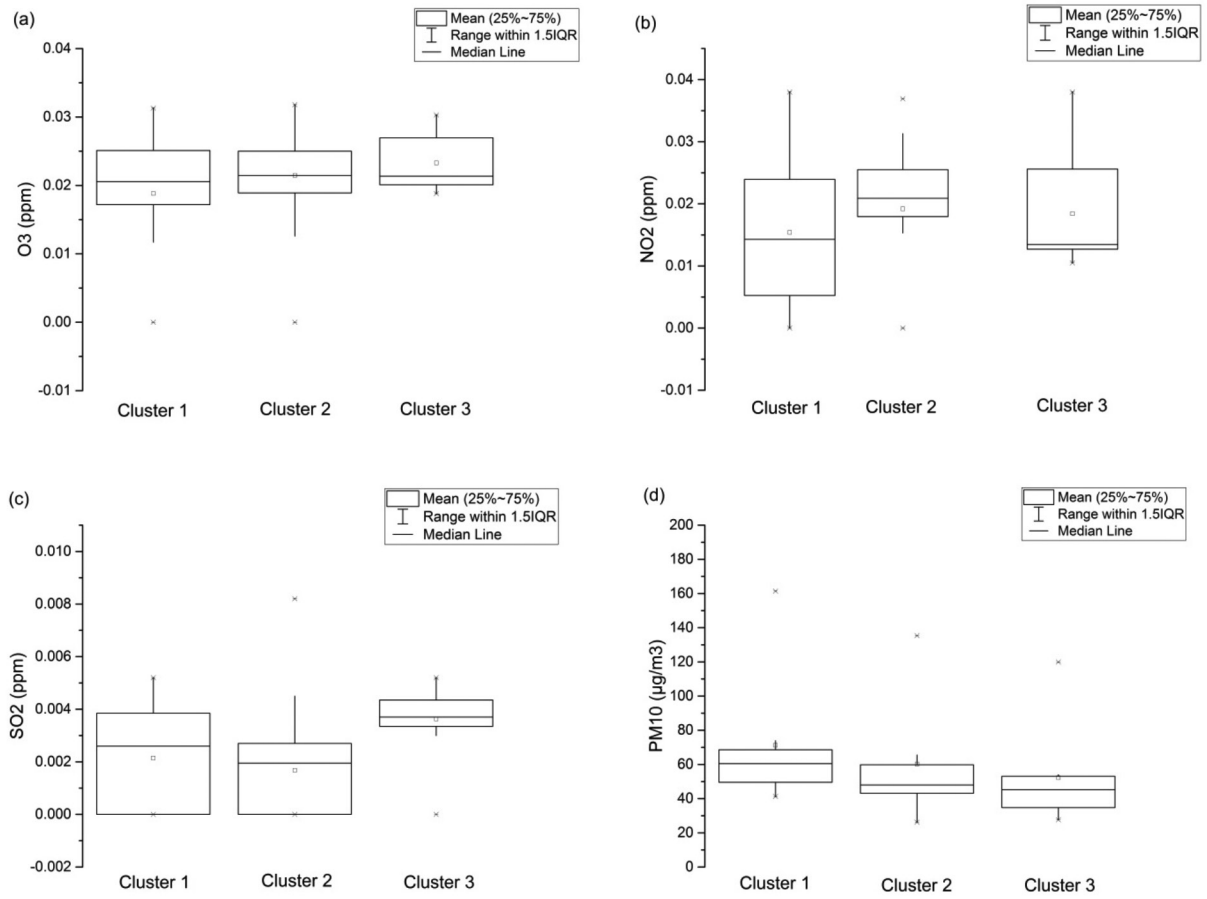

FIGURE 3. Box and whisker plot for (a) O_3, (b) NO_2, (c) SO_2 and (d) PM_{10}, that generated from backward stepwise mode in DA of air quality in Selangor

TABLE 4. Varimax rotation PCs for air quality data based on three clusters in the Selangor

Variables	LPS		MPS		SHPS	
	VF1	VF2	VF1	VF2	VF1	VF2
O_3	.038	**.811**	**.599**	.437	**.578**	-.015
CO	-**.909**	.034	**.789**	.291	**.813**	.016
NO_2	.005	-**.895**	**.711**	.221	-.260	**.882**
SO_2	**.745**	-.089	0.96	**.921**	**.664**	.198
PM_{10}	**.659**	.174	.049	-.300	**.622**	**.637**
Eigenvalue	1.816	1.497	2.057	1.263	1.890	1.224
Variability (%)	36.328	29.947	41.145	25.252	37.796	24.473
Cumulative (%)	36.328	66.274	41.145	66.397	37.796	62.269

*The bold value are factor loadings above 0.5 that were taken after Varimax rotation was performed

positive significant loadings of SO_2 and PM_{10}. Meanwhile, VF2 indicated a total variance of 29.94% resulting in strong positive loadings of O_3 and strong negative loadings of NO_2. In other words, the existence of CO and NO_2 can be linked to the process of biomass burning and grazing and the residual of agricultural products from agricultural activities, as well as long range transportation air pollutants and domestic fuel sources (Haiduc & Beldean-Gale 2011; Rajab et al. 2011). Meanwhile O_3, SO_2, and PM_{10} can be link to motor vehicles, industrial activities and construction sites (Sadanaga et al. 2012; Wei et al. 2012). According to the Ministry of Transport (MOT), Malaysia, the total amount of newly registered motor vehicles in Malaysia increased by 7.24% from 1,160,082 in 2010 to 1,638,498 in 2015; which increase the possibility of motor vehicles in becoming a major factor that contribute to the deterioration of atmospheric conditions.

Moderate pollution source (MPS) region MPS region indicated that the VF1 contributed around 41.14% of

total variance and has strong positive loadings on CO and moderate positive loadings of NO_2 and O_3. Simultaneously, VF2 resulted in 25.25% of total variance producing strong positive loading of SO_2. The result showed that factors containing chemical compositions that were involved with fossil fuel combustion especially from industrial activities and vehicles have become the main source of air pollution (Mutalib et al. 2013). Generally, the release of O_3 into the atmosphere can be related to photochemical oxidation and the main component of smog (Banan et al. 2013). Moreover, urban and suburban activities which causes the release of mono-nitrogen oxide (NO_x) (Sadanaga et al. 2012), as well as industrial activities which released SO_2 (Wei et al. 2013) could assist in the increase of O_3 concentration in the atmosphere.

Slightly high pollution source (SHPS) region SHPS region indicate that the VF1 with total of variance of 36.32% produced a strong positive loading of CO and moderate positive loadings of O_3, SO_2 and PM_{10}. Meanwhile, VF2 with a total of variance of 29.94% produced strong positive loading of NO_2 and moderate positive loading of PM_{10}. Major pollution occurred in SHPS region could be related to the composition of chemicals from anthropogenic activities which consist of point source pollution. In other words, the pollutants originated from the burning of biomass and fossil fuels, particularly from industrial, residential and vegetation areas; as well as from motor vehicles and natural emission sources (Azid et al. 2015; Dominick et al. 2012; Mutalib et al. 2013). It should be noted that the concentration of PM10 was detected to be higher than other pollutants due to high traffic congestion of motor vehicles, industrial activities in construction site, soil dust and open burning activities (Azid et al. 2015, 2014, 2013; Mutalib et al. 2013).

MULTIPLE LINEAR REGRESSION (MLR) OF AIR POLLUTANT INDEX (API)

The MLR modeling was done to identify the behavior of variables, which can be done using the linear least-square fitting process and to determine all trace element sources (Henry et al. 1984). For that reason, this study used the source of apportionment of air pollutant parameter to identify the potential API. Three models were developed using the API value as a dependent variable, whereas the independent variables will be based on the air quality parameters taken from LPS (4 variables), MPS (4 variables) and SHPS (4 variables).

R^2, Adjusted R^2 and RSME value are essential for better coefficient result, which is why they were used in the LPS, MPS and SHPS. The value for R^2, Adjusted R^2 and RSME for LPS are 0.837, 0.816 and 3.324, respectively; 0.878, 0.822, and 2.829 for MPS; and 0.894, 0.834, and 1.808 SHPS. The proposed equation of R^2, Adjusted R^2 and RMSE is shown in (5i) to (5iii):

LPS (4 variables)

$$\text{Total API} = 3.612 + 32.792\,(NO_2) - 875.640(CO) + 0.783(PM_{10}) + 210.13(SO_2)$$

($R^2 = 0.837$; Adjusted $R^2 = 0.816$; RMSE = 3.324) (5i)

MPS (4 variables)

$$\text{Total API} = 3.977 + 5.423(CO) - 178.579(NO_2) + 223.63\,(SO_2) + 54.29\,(O_3)$$

($R^2 = 0.878$; Adjusted $R^2 = 0.822$; RMSE = 2.829) (5ii)

SHPS (4 variables)

$$\text{Total API} = 4.596 - 58.943(O_3) + 2.065(CO) + 501.098(NO_2) + 0.821(PM_{10})$$

($R^2 = 0.894$; Adjusted $R^2 = 0.834$; RMSE = 1.808) (5iii)

Measured results of (5i) to (5iii) shows that the highest coefficient of determination (R^2) came from SHPS with 0.894 for NO_2, CO and PM_{10}, as well as a negative for O_3; followed by MPS with $R^2 = 0.878$ for CO, SO_2, and O_3, but negative for NO_2; and the lowest came from LPS with $R^2 = 0.837$ for NO_2, PM_{10}, and SO_2, with negative for CO. From the finding, Cluster SHPS has been determined as the best model due to the closest R^2 value to 1 and the smallest RMSE when compared to other parameters (Azid et al. 2015; Dominick et al. 2012; Mutalib et al. 2013).

Figure 4 shows the perceived residual analysis and prediction of total API using LPS, MPS and SHPS. The results indicated that the deficiency of the model contained some differences in the range of -0.5 to 1.5 for LPS, -1.7 to 1.3 for MPS and -2.3 to 1.5 for SHPS. Meanwhile, the standard predicted values for LPS, MPs and SHPS ranged between -0.8 and 1.6, -2.5 and 2.2, and -1.9 and 1.8, respectively. The main objective of the scatter plot diagram was to prove that the SHPS model is suitable to be used for total API prediction. This is because the model provided a results with greater difference between the predicted API and the calculated API.

CONCLUSION

The study concluded that spatial variations of air quality data in Selangor has been successfully studied using chemometric approach, such as HCA, DA, PCA and MLR. HCA has successfully grouped the eight sampling stations with five air quality variables into three significant clusters, namely LPS, MPS and SHPS. The HCA has benefitted the monitoring network approach by reducing the number of monitoring stations. Meanwhile, clusters delivered from HCA into DA has confirmed the standard mode, forward stepwise mode and backward stepwise mode with the accuracy of 95.38%, 89.05% and 93.23%, respectively, which confirmed the selection of the four variables of O_3, NO_2, SO_2, and PM_{10} in backward stepwise mode. In PCA, two VFs were detected in LPS, MPS and SHPS regions, with the total of variance of 66.27%, 66.39% and 62.26%,

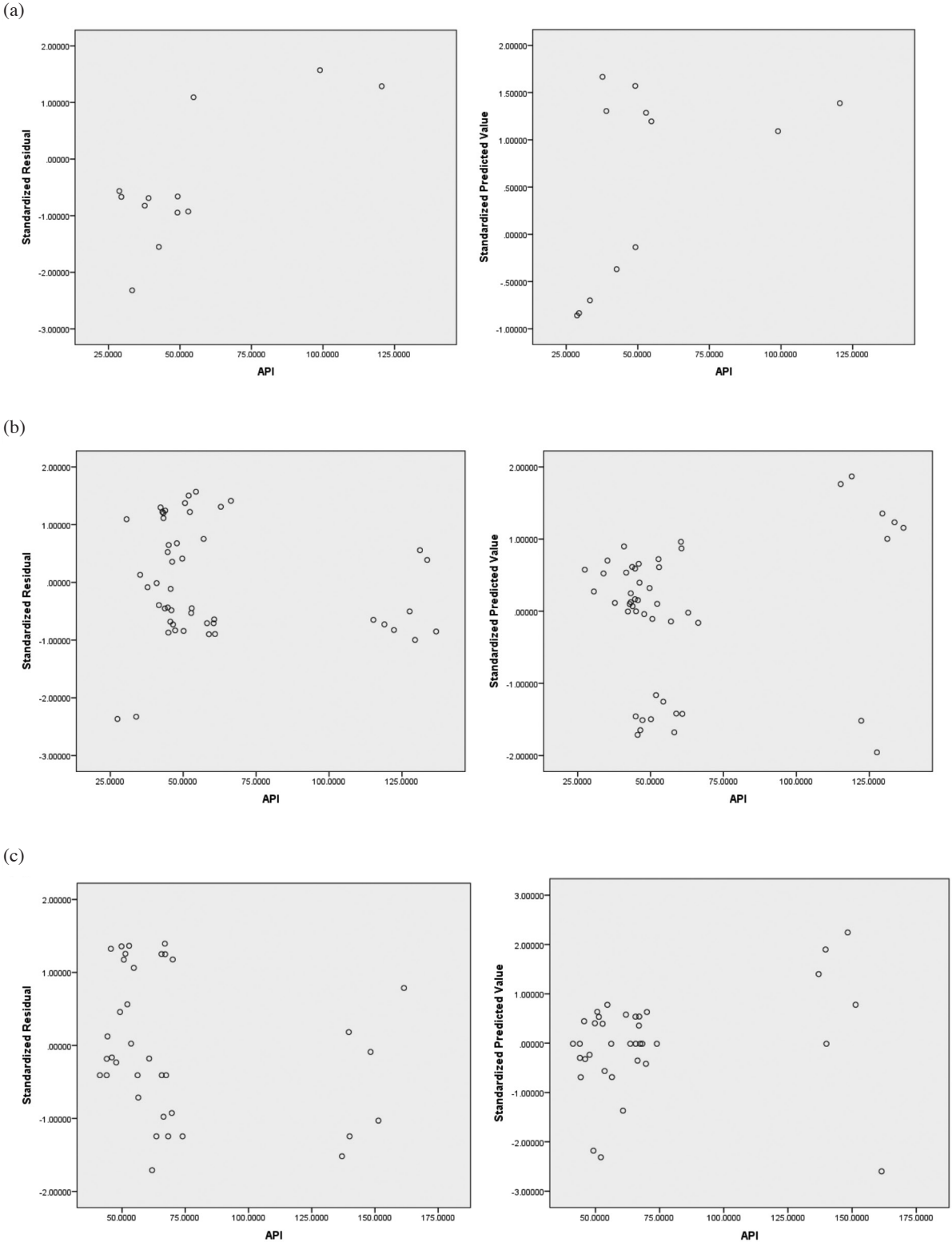

FIGURE 4. Scatter plot diagram of standardized residuals and standard predicted value for (a) LPS, (b) MPS and (c) SHPS

respectively. The sources of variations detected in this study are industrial emissions, transport emissions, agricultural systems, residential factors and natural emission sources. MLR analysis was carried out to determine the variability of proposed equation to predict the total values of the API. The strong result of R^2 value was due to high significant *p*-value of smaller than 0.05 when compared to the three developed models. The highest R^2 values are SHPS with

0.894, followed by MPS with 0.878 and LPS with 0.837. It was determined that the most suitable model to be used for total API prediction is the SHPS model due to the greater difference it provides between predicted API and calculated API. For effective air quality management, a new air monitoring network should be design for the purpose of practicality and cost saving.

REFERENCES

Aertsen, W., Kint, V., Van Orshoven, J., Özkan, K. & Muys, B. 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological modelling* 221(8): 1119-1130.

Aris, A.Z., Preveena, S.M., Isa, N.M., Lim, W.Y., Juahir, H., Yusoff, M.K. & Mustapha, A. 2013. Application of environmentric methods to surface water quality assessment of Langkawi Geopark (Malaysia). *Environmental Forensics* 14(3): 230-239.

Azid, A., Juahir, H., Ezani, E., Toriman, M.E., Endut, A., Rahman, M.N.A., Yunus, K., Kamarudin, M.K.A., Hasnam, C.N.C., Saudi, A.S.M. & Umar, R. 2015. Identification source of variation on regional impact of air quality pattern using chemometric. *Aerosol and Air Quality Research* 15(4): 1545-1558.

Azid, A., Juahir, H., Toriman, M.E., Kamarudin, M.K.A., Saudi, A.S.M., Hasnam, C.N.C., Aziz, N.A.A., Azaman, F., Latif, M.T., Zainuddin, S.F.M. & Osman, M.R. 2014. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, & Soil Pollution* 225(8): 2063.

Azid, A., Juahir, H., Latif, M.T., Zain, S.M. & Osman, M.R. 2013. Feed-forward artificial neural network model for air pollutant index prediction in the southern region of Peninsular Malaysia. *Journal of Environmental Protection* 4(12A): 1-10.

Azmi, S.Z., Latif, M.T., Ismail, A.S., Juneng, L. & Jemain, A.A. 2010. Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Quality, Atmosphere & Health* 3(1): 53-64.

Banan, N., Latif, M.T., Juneng, L. & Ahamad, F. 2013. Characteristics of surface ozone concentrations at stations with different backgrounds in the Malaysian Peninsula. *Aerosol and Air Quality Research* 13(3): 1090-1106.

Department of Environment Malaysia (DOE). 2012. *Malaysia Environmental Quality Report*. Kuala Lumpur: Department of Environment, Ministry of Natural Resources and Environment.

Dominick, D., Juahir, H., Latif, M.T., Zain, S.M. & Aris, A.Z. 2012. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment* 60: 172-181.

Haiduc, I. & Beldean-Gale, M.S. 2011. Variation of greenhouse gases in urban areas-case study: CO_2, CO, and CH_4 in three Romanian cities. *Air Quality-Models and Applications*. Intech

Henry, R.C., Lewis, C.W., Hopke, P.K. & Williamson, H.J. 1984. Review of receptor model fundamentals. *Atmospheric Environment* (1967) 18(8): 1507-1515.

Hua, A.K., Kusin, F.M. & Praveena, S.M. 2016. Spatial variation assessment of river water quality using environmetric techniques. *Polish Journal of Environmental Studies* 25(6): 2411-2421.

Juahir, H., Zain, S.M., Yusoff, M.K., Hanidza, T.I.T., Armi, A.S.M., Toriman, M.E. & Mokhtar, M. 2011. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environmental Monitoring and Assessment* 173(1): 625-641.

Khan, M.F., Latif, M.T., Lim, C.H., Amil, N., Jaafar, S.A., Dominick, D., Nadzir, M.S.M., Sahani, M. & Tahir, N.M. 2015. Seasonal effect and source apportionment of polycyclic aromatic hydrocarbons in PM 2.5. *Atmospheric Environment* 106: 178-190.

Mahiyuddin, W.R.W., Sahani, M., Aripin, R., Latif, M.T., Thach, T.Q. & Wong, C.M. 2013. Short-term effects of daily air pollution on mortality. *Atmospheric Environment* 65: 69-79.

Makmom Abdullah, A., Armi Abu Samah, M. & Yee Jun, T. 2012. An overview of the air pollution trend in Klang Valley, Malaysia. *Open Environmental Sciences* 6(1): 13-19.

Manjunanth, B.G., Frick, M. & Reiss, R.D. 2012. Some notes on extremal discriminant analysis. *J. Multivar. Anal*. 103: 107-115.

Moustris, K.P., Ziomas, I.C. & Paliatsos, A.G. 2010. 3-Day-ahead forecasting of regional pollution index for the pollutants NO_2, CO, SO_2, and O_3 using artificial neural networks in Athens, Greece. *Water, Air, & Soil Pollution* 209(1-4): 29-43.

Mutalib, S.N.S.A., Juahir, H., Azid, A., Sharif, S.M., Latif, M.T., Aris, A.Z., Zain, S.M. & Dominick, D. 2013. Spatial and temporal air quality pattern recognition using environmetric techniques: A case study in Malaysia. *Environmental Science: Processes & Impacts* 15(9): 1717-1728.

Rajab, J.M., Tan, K.C., Lim, H.S. & MatJafri, M.Z. 2011. Investigation on the carbon monoxide pollution over Peninsular Malaysia caused by Indonesia fires from AIRS daily measurement. *Advanced Air Pollution*. InTech.

Sadanaga, Y., Sengen, M., Takenaka, N. & Bandow, H. 2012. Analyses of the ozone weekend effect in Tokyo, Japan: Regime of oxidant (O^{3+} NO_2) production. *Aerosol Air Qual. Res*. 12: 161-168.

Seinfeld, J.H. & Pandis, S.N. 1998. *Atmospheric Chemistry and Physics*. New York: John Wiley & Sons Inc.

Sulong, N.A., Latif, M.T., Khan, M.F., Amil, N., Ashfold, M.J., Wahab, M.I.A., Chan, K.M. & Sahani, M. 2017. Source apportionment and health risk assessment among specific age groups during haze and non-haze episodes in Kuala Lumpur, Malaysia. *Science of the Total Environment* 601: 556-570.

Ul-Saufie, A.Z., Yahya, A.S., Ramli, N.A. & Hamid, H.A. 2011. Comparison between multiple linear regression and feed forward back propagation neural network models for predicting PM10 concentration level based on gaseous and meteorological parameters. *International Journal of Applied* 1(4): 42-49.

Wei, X., Liu, Q., Lam, K.S. & Wang, T. 2012. Impact of precursor levels and global warming on peak ozone concentration in the Pearl River Delta Region of China. *Advances in Atmospheric Sciences* 29(3): 635-635.

Department of Environmental Sciences
Faculty of Environmental Studies
Universiti Putra Malaysia
43400 UPM Serdang, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: angkeanhua@yahoo.com

Received: 3 July 2017
Accepted: 17 October 2017