Aerosol and Air Quality Research, 15: 1545–1558, 2015

ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2014.04.0073

Identification Source of Variation on Regional Impact of Air Quality Pattern Using Chemometric

Azman Azid[1], Hafizan Juahir[1*], Ezureen Ezani[1], Mohd Ekhwan Toriman[1], Azizah Endut[1], Mohd Nordin Abdul Rahman[2], Kamaruzzaman Yunus[3], Mohd Khairul Amri Kamarudin[1], Che Noraini Che Hasnam[1], Ahmad Shakir Mohd Saudi[1], Roslan Umar[1]

[1] *East Coast Environmental Research Institute (ESERI), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia*
[2] *Center of Research & Innovation Management (CRIM), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia*
[3] *Kulliyyah of Science, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia*

ABSTRACT

This study intends to show the effectiveness of hierarchical agglomerative cluster analysis (HACA), discriminant analysis (DA), principal component analysis (PCA), factor analysis (FA) and multiple linear regressions (MLR) for assessing the air quality data and air pollution sources pattern recognition. The data sets of air quality for 12 months (January–December) in 2007, consisting of 14 stations around Peninsular Malaysia with 14 parameters (168 datasets) were applied. Three significant clusters - low pollution source (LPS) region, moderate pollution source (MPS) region, and slightly high pollution source (SHPS) region were generated via HACA. Forward stepwise of DA managed to discriminate 8 variables, whereas backward stepwise of DA managed to discriminate 9 out of 14 variables. The method of PCA and FA has identified 8 pollutants in LPS and SHPS respectively, as well as 11 pollutants in MPS region, where most of the pollutants are expected derived from industrial activities, transportation and agriculture systems. Four MLR models show that PM_{10} categorize as the primary pollutant in Malaysia. From the study, it can be stipulated that the application of chemometric techniques can disclose meaningful information on the spatial variability of a large and complex air quality data. A clearer review about the air quality and a novel design of air quality monitoring network for better management of air pollution can be achieved.

Keywords: Air quality; Chemometric; Pattern recognition; HACA; DA; PCA; FA; MLR.

INTRODUCTION

Nowadays, the ambient air quality is a matter of serious concern to both developed and developing countries. The status of air quality is described according to biological, chemical and physical properties. Based on the properties, the quality of air can be expressed via a numerical index such as an air pollutant index (API), by combining measurements of selected air quality variables. The index is important in evaluating the air quality of different sources (Azid *et al.*, 2014a). Once the compliance or lack of compliance determined, the data can be used to advise or caution the public in lieu of health effects (Kamal *et al.*, 2006; Azid *et al.*, 2014a). Poor air quality has both acute and chronic effects, especially to human health (Moustris *et al.*, 2010). In Malaysia, the government applied the API as an indicator of air quality since 1989 (Dominick *et al.*, 2012; Mutalib *et al.*, 2013; Azid *et al.*, 2014a). The API value calculated based on the sub-index of five types of air pollutants - ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and particulate matters under 10 μm (PM_{10}) (Mutalib *et al.*, 2013). The higher level of API value, the greater the level of air pollution and the greater the health concern. For example, an API value of 50 represents good air quality, while an API value over 300 indicates hazardous to air quality (DOE, 2013). The status of air quality in Malaysia is monitored by the establishment of Recommended Malaysian Air Quality Guideline (RMAQG) issued by the Malaysian Department of Environment (DOE) since 1989 (Dominick *et al.*, 2012; Mutalib *et al.*, 2013; Azid *et al.*, 2014a).

Malaysia strives to achieve industrialized country status by 2020, which is highly correlated with rapid economic growth. This may contribute tremendously to industrial

* Corresponding author.
Tel.: +609-666 3410
E-mail address: hafizanjuahir@unisza.edu.my

pollution and degradation of urban environments. With the rapid economic growth, air pollution is the main subject that has been adversely affecting human health, agricultural crops, animals and ecosystems (Moustris *et al.*, 2010; Azid *et al.*, 2015). Hence, it can unavoidably cause damages to buildings, monuments and statues. Simultaneously, not only it reduces visibility, it even interferes aviation. Most of the air pollution sources derived from land transportation (mobile source), industrial emissions (stationary source), and open burning sources (Afroz *et al.*, 2003; Azmi *et al.*, 2010; Abdullah *et al.*, 2012; Azid *et al.*, 2013, 2014a, b). The studies and monitoring data on ambient air quality have traced that, several air pollutants in various regions are increasing with time and are not constantly at satisfactory levels according to the national ambient air quality standards.

Air pollution control is needed to prevent the situation from worsening in the long run (Moustris *et al.*, 2010; Azid *et al.*, 2015). Consequently, air quality monitoring network is a part of the preliminary strategy for the air pollution deterrence plan in Malaysia. The common approach to the network design involves placing monitoring stations at selected representative spots, chosen on the basis of required data, and recognized dispersion or emission patterns of pollutants in the study area. This study is considered as the scientific approach, which will provide a cost-effective air quality monitoring program. Stations must be carefully selected. Moreover, chemometric analysis may need to be utilized to complement such monitoring strategy (Lu *et al.*, 2011). A properly-designed air monitoring network is a main component of any air quality control program. The operation and maintenance of air quality monitoring stations and tools for measuring the parameters of air quality are costly, so it is more favourable to use as few stations and parameters as possible to achieve the objectives of monitoring.

Instead of traditional statistical methods, the chemometric techniques (also known as multivariate techniques) believed as a better tool for analysing air quality. Chemometric in the environmental field is verified to be a functional tool to identify the sources of pollution (Simeonov *et al.*, 2002; Mutalib *et al.*, 2013; Azid *et al.*, 2015). Chemometric analysis includes the interrelationship of faunal structure, physical-chemical characterization, and toxicity data that received from laboratory analysis. The analysis is considered to be the most suitable tool for the reduction and interpretation of meaningful data (Kannel *et al.*, 2007; Satheeshkumar and Khan, 2011; Mutalib *et al.*, 2013). Unbiased methods such as hierarchical agglomerative cluster analysis (HACA), discriminant analysis (DA), principal component analysis (PCA), factor analysis (FA) and multiple linear regressions (MLR) were used in air quality analysis. The application of diverse chemometric statistical techniques for interpretation of the complex databases, permits a better understanding of air quality in the study region. Chemometric methods also offer the recognition of the potential sources that are accountable for variations in air quality and manipulate the air quality. Therefore, the methods have been proven as priceless tools for developing suitable plans for efficient management of the air monitoring network (Singh *et al.*, 2005; Azid *et al.*, 2015).

The objectives of this study are to illustrate a clearer view of air quality in Peninsular Malaysia by recognizing the pollution source, identifying which air quality variables are the most significant in this study, and make a predictive performance of air quality along the study area.

MATERIALS AND METHODS

Study Sites

Peninsular Malaysia is a part of Malaysia, where the economic growth is more rapid and dominant, and has the same air pollution problems as other developed and developing countries in the world. Its area is approximately 131,600 square kilometres. It shares a land boundary with Thailand in the north and the island of Singapore in the south. To the west of it lays the Strait of Malacca, and across of the strait is an island of Sumatra. With an estimated population of 29 million in Malaysia, Peninsular Malaysia accounts for the majority (approximately 80%) of Malaysia's population and economy. There are no main natural disasters occurred in Peninsular Malaysia (such as typhoon, volcanic eruption and earthquake), which makes the air quality in Peninsular Malaysia is under controls. However, the emerging economic growth these days have worsened the existing air quality. This study is vital to show the latest status of air quality in Peninsular Malaysia.

The air quality data in this study acquired from 14 stations across the Peninsular Malaysia (Fig. 1). These stations were chosen due to located in urban, suburban, and industrial area. The chosen stations are demonstrated in Table 1.

Data Collection

The air quality data were gathered from the DOE, from January to December 2007. The variables such as ambient temperature (°C), methane (CH_4, ppm), carbon monoxide (CO, ppm), relative humidity (%), non-methane hydrocarbons (NmHC, ppm), nitrogen monoxide (NO, ppm), nitrogen dioxide (NO_2, ppm), nitrogen oxides (NO_x, ppm), ozone (O_3, ppm), particulate matter (PM_{10}, µg/cu.m), sulfur dioxide (SO_2, ppm), total hydrocarbons (THC, ppm), ultra-violet B (J/m^2hr) and wind speed (km/hr) were selected to study the influence of API values and the sources of pollution. For the statistical analysis in this study, the hourly data were used to form a monthly average which comprises 168 datasets (12 data per stations × 14 stations) with a total of 2,352 observations (12 data per stations × 14 variables × 14 stations) were employed.

Chemometric Analysis

Hierarchical Agglomerative Cluster Analysis (HACA)

HACA is an unsupervised pattern identification method to split a large group into smaller ones (Almeida *et al.*, 2007). In this study, HACA is used to identify unseen 'clusters' which are illustrated by numerical, symbolical or structural data, such that the members of a significant cluster share the similarities among them and the clusters are confidently well parted (Bock, 1996). HACA is employed on the normal distribution dataset via the Ward's method by means of Euclidean distances, as a measure of the relationship (Juahir *et al.*, 2011). The outcome of this

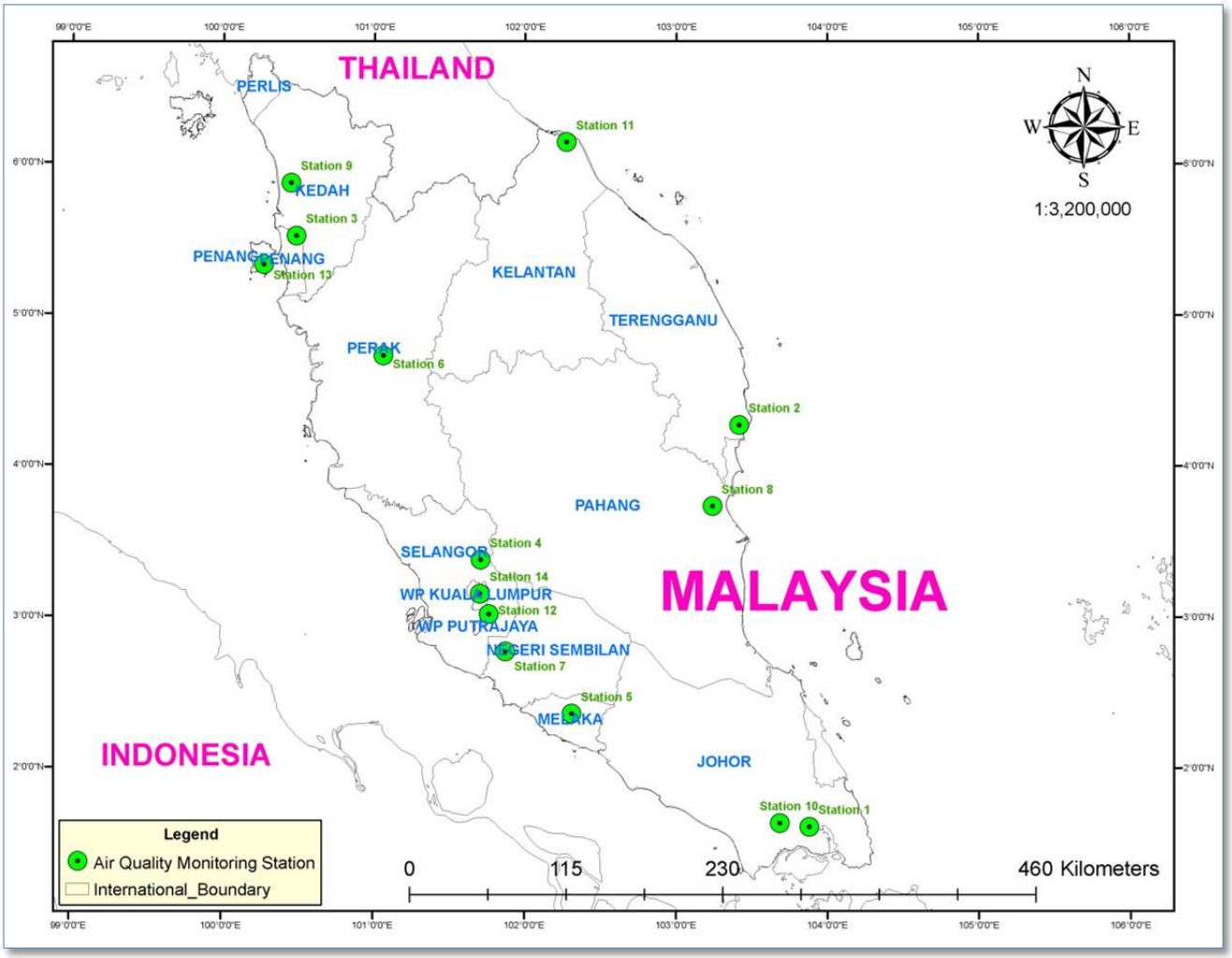

Fig. 1. 14 selected air quality monitoring stations across Pe2ninsular Malaysia.

Table 1. The details of 14 monitoring stations.

Station No.	Site State	Location	Latitude	Longitude
Station 1	Johor	SM Pasir Gudang 2, Pasir Gudang	N01° 28.225	E103° 53.637
Station 2	Terengganu	SRK Bukit Kuang, Teluk Kalung, Kemaman	N04° 16.260	E103° 25.826
Station 3	Pulau Pinang	Sek. Keb. Cenderawasih, Tmn. Inderawasih, Perai	N05° 23.470	E100° 23.213
Station 4	Selangor	Jab. Bekalan Air Daerah Gombak	N03° 15.702	E101° 39.103
Station 5	Melaka	Sek. Men. Keb. Bukit Rambai, Melaka	N02° 15.510	E102° 10.364
Station 6	Perak	SM Jalan Tasek, Ipoh	N04° 37.781	E101° 06.964
Station 7	Negeri Sembilan	Taman Semarak (Phase II), Nilai	N02° 49.246	E101° 48.877
Station 8	Pahang	SK Indera Mahkota, Kuantan	N03° 49.138	E103° 17.817
Station 9	Kedah	SK Bakar Arang, Sungai Petani	N05° 37.886	E100° 28.189
Station 10	Johor	SM Vok. Perdagangan, Johor Baru	N01° 29.815	E103° 43.617
Station 11	Kelantan	Maktab Sultan Ismail, Kota Bharu	N06° 09.520	E102° 15.059
Station 12	Selangor	Country Heights, Kajang	N02° 59.645	E101° 44.417
Station 13	Pulau Pinang	USM, Minden	N05° 21.528	E100° 17.864
Station 14	Kuala Lumpur	S. M. Keb. Seri Permaisuri, Cheras, Kuala Lumpur	N03° 06.376	E101° 43.072

method depicted by a *treelike* method, which known as a dendrogram. The Euclidean distance (known as linkage distance) accounted as D_{link}/D_{max}. It signifies the measure between the linkage distances divided by the maximal distance. The measure will be multiplied by 100 as a way to standardize the linkage distance signified by the y-axis (Shrestha and Kazama, 2007). Euclidean distance can be defined by Eq. (1):

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

where, $d(x,y)$ is the Euclidean distance between two items represented by x_m and y_m; p is the dimensional space of the variables.

Analysis of variance (ANOVA) is used to analyse the distances between clusters in Ward's method, which is established to minimize the total of squares of any two achievable clusters at every step (Ward, 1963). Then, DA is performed to confirm the groups clustered by HACA.

Discriminant Analysis (DA)

The fundamental point of DA is to classify an object of unknown origin to one of several naturally occurring groups (Manjunath *et al.*, 2012). In this study, DA was coupled with the HACA for the goal of establishing the significantly different variables and reducing the errors of these groups (Kannel *et al.*, 2007). For every cluster, it creates a discriminant function (DF) (Johnson and Wichern 1992). Then, the DFs can be determined by Eq. (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 inherent to each group, n is the number of parameters used to classify a set of data into a given group, and w_j is the weight coefficient assigned by DF analysis (DFA) to a given parameter (Pj).

In this study, DA was applied on three modes, which are standard mode, forward stepwise mode and backward stepwise mode. A standard mode was performed to create DFs for evaluating spatial variations in the air quality raw data. In the forward stepwise mode, variables were gradually eliminated starting with the most significant variable until no significant changes were found. Nevertheless, in the backward stepwise mode, variables were eliminated gradually, starting with the least significant variable until no significant changes were found.

Principal Component Analysis (PCA)

In this study, PCA was used in order to analyse and interpret set of interrelated variables. PCA is a method of creating new variables, which are linear composites of the original variables. The new variables known as principal components (PCs), while the values of PCs known as principal component scores (PCS). The maximum number of new variables is equivalent to the number of original variables (Juahir *et al.*, 2011). The PCA was utilized to identify the emission source (Hopke, 1985). In this study, the HACA was coupled with PCA in order to create the most powerful model recognition of emission sources. It presents the details on the most significant variables due to spatial and temporal variations, by putting them from the less significant variables with minimum loss of the original information (Singh *et al.*, 2004; 2005; Azid *et al.*, 2015). The principal components (PC) can be assessed as Eq. (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.

Factor Analysis (FA)

The FA were performed to infer relationships between variables (Lioy *et al.*, 1989). Varimax method is used in the FA techniques. The varimax rotation guarantees that every variable associated with only one principal component as encompassing a near-zero relationship with the other components (Soares *et al.*, 2008). Sometimes, the PCs produced by PCA are not interpreted well. Consequently, the varimax rotation has been applied to rotate the PCs for the interpretation purposes. Eigenvalues obtained from varimax rotation are the precursor of the FA. Eigenvalues over than 1 were deemed as significant and subsequently varimax factors (VFs), which are the new groups of variables are generated (Yu and Chang, 2000). The VFs values which are greater than 0.75 (> 0.75) is considered as "strong", the values range from 0.50–0.75 (0.50 ≥ factor loading ≥ 0.75) is considered as "moderate", and the values range from 0.30–0.49 (0.30 ≥ factor loading ≥ 0.49) is considered as "weak" factor loadings (Liu *et al.*, 2003; Azid *et al.*, 2014a; Azid *et al.*, 2015). In practice, only factor loadings with absolute values greater than 0.75 are selected for the principal component interpretation (Juahir *et al.*, 2011, Azid *et al.*, 2014a). Emission source recognition of different air pollutants was completed based on different activities in the three significant clustered regions. The fundamental model of FA is stated as Eq. (4):

$$z_{ij} = a_{f1}f_{1i} + a_{f2}f_{2i} + \ldots + a_{fm}f_{mi} + e_{fi} \quad (4)$$

where, z is the measured value of a variable, a is the factor loading, f is the factor score, e is the residual term accounting for errors or other sources of variation, i is the sample number, j is the variable number, and m is the total number of factors.

In this study, the PCA and FA were applied to the classified datasets (14 variables) independently, according to regions (LPS, MPS and SHPS) that were classed by HACA method.

Multiple Linear Regressions (MLR)

Multiple linear regression (MLR) is widely used in atmospheric modelling (Dominick *et al.*, 2012). This technique has been used for investigating the relationship among various independent and dependent variables by fitting a linear equation to observed data (Pai *et al.*, 2009; Ul-Saufi *et al.*, 2011) and gives the percentage of the contribution of each parameter to the atmospheric pollution (Aertsen *et al.*, 2010). In this study, it was used to justify the relationship between the air quality parameters and total API data. The model of the original air quality parameters-API was compared to the nine air quality parameters (the most significant parameters)-API, in order to get a better model within clusters. The model generalizes of the simple

linear regression, in which each value of the independent variable is associated with a value of the dependent variable. The model is obtained using the Eq. (5):

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + \varepsilon \quad (5)$$

where, Y is the response variable, and there are $p-1$ explanatory variable x_1, x_2, …, x_{p-1}, with p parameters (regression coefficients) β_0, β_1, β_2, …, β_{p-1} and ε is an error associated with the regression.

To get the best fitting linear regression equation, the coefficient of determination (R^2), adjusted coefficient of determination (Adjusted R^2) and root mean square error (RMSE) are the values that need to be considered (Aertsen *et al.*, 2010). Unfortunately, the value of R^2 only provides information about how well it performs on external data (Dominick *et al.*, 2012). However, the highest values of R^2 (which was near to 1) will be declared as the best linear model (Norusis 1990; Mutalib *et al.*, 2013; Azid *et al.*, 2013, 2014a). The adjusted R^2 value is calculated by considering all the possible number of variables, since R^2 tends to over-estimate the success of the model when applied to the real world. Meanwhile, RMSE was used to measure the residual error and it will be taken into account for estimation of the mean difference between observed and modelled value of the API. The smallest RMSE and the closest R^2 value to 1, the better model shall be performed (Dominick *et al.*, 2012; Mutalib *et al.*, 2013; Azid *et al.*, 2013, 2014a).

In this study, HACA, DA, PCA, FA and MLR were performed using XLSTAT 2014 add-in software.

RESULTS AND DISCUSSION

Spatial Classification Based on Air Quality Parameters

This part observes the historical values of air quality parameters step by step to categorize the air quality station based on their homogeneity level by means of HACA. Figs. 2 and Fig. 3 show the three significant regions illustrated by HACA and the potential pollution sources within the study regions. Three clusters were generated from the clustering method in an incredibly convincing way, as the stations in these clusters share the homogeneity characteristics.

The study areas are diversified into three significant groups of regions, which are the low pollution source (LPS), moderate pollution source (MPS), and slightly high pollution source (SHPS) region. Cluster 1 (station 1, station 2, station 4, station 8, station 11 and station 13) corresponds to the LPS region due to no severe air pollution occurred during the year, with the average value of the API is 38 during the year. Cluster 2 (station 5) corresponds on the MPS region due to the average API values is 52 during the year. Cluster 3 (station 3, station 6, station 7, station 9, station 10, station 12 and station 14) corresponds on the SHPS region with the API values is 78 during the year.

This result implies for a shorter period of air quality assessment, each cluster of region only requires one station to correspond to a practically precise spatial assessment of the air quality. By applying HACA technique, the number of monitoring station can be reduced. Three monitoring stations which are representing three significantly clustered regions are adequate to construct the whole monitoring network. It has clearly shown that the HACA technique is practiced in suggesting dependable categorization of air quality for the entire region and can be employed to design an improvement monitoring network in future.

Discrimination of Spatial Variation

The air quality data post clustering of the monitoring stations into three significant clusters obtained by HACA was then undergoing with DA. Through DA, the spatial variation among the diverse regions can be studied. The air quality parameters were treated as independent variables, whereas the three significant groups (LPS, MPS and SHPS)

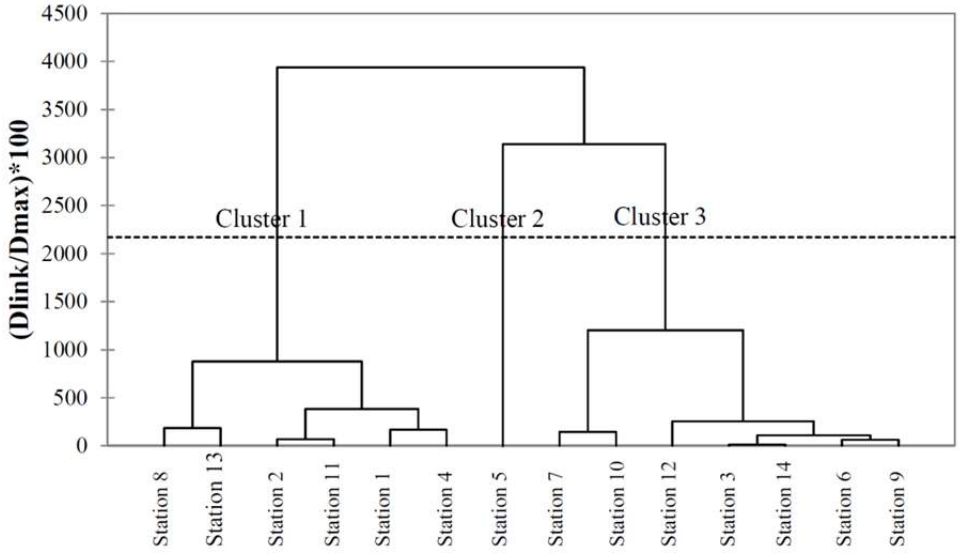

Fig. 2. Dendrogram showing different clusters of sampling stations located across Peninsular Malaysia based on air quality parameters.

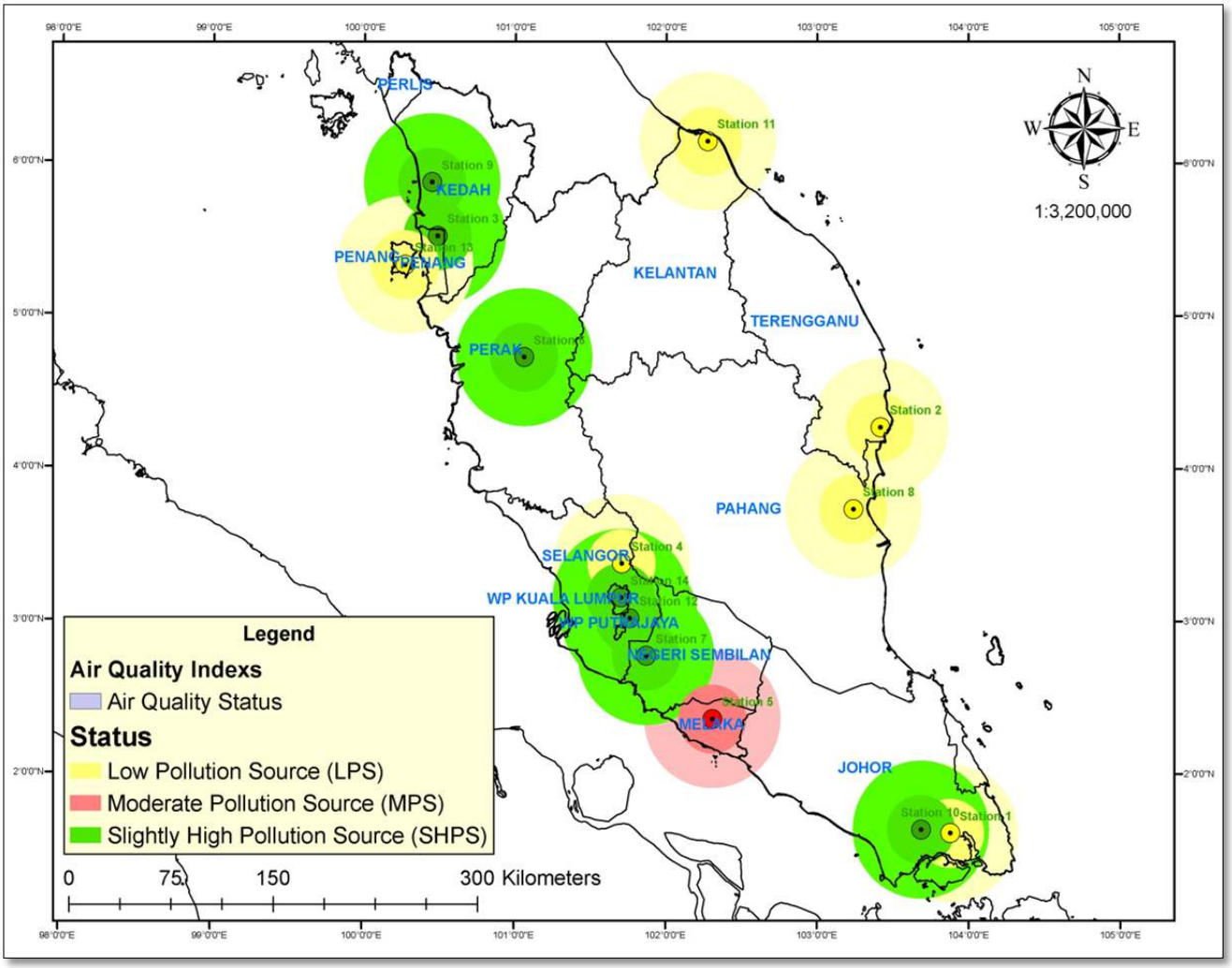

Fig. 3. Classification of regions due to air quality by HACA for the Peninsular Malaysia.

were treated as dependent variables. DA was implemented for three modes which were standard, forward stepwise, and backward stepwise.

The accuracy of spatial variation by means of standard mode, forward stepwise mode, and backward stepwise mode were 95.83% (fourteen variables), 94.05% (eight variables), and 94.05% (nine variables), respectively (Table 2). The discriminant variables resulting from the forward stepwise mode are NO_2, SO_2, PM_{10}, CH_4, humidity, NmHC, ultra-violet B, and wind speed. It shows that these eight discriminant variables have high variation due to their spatial distribution. In contrast, backward stepwise mode included NO_x as the additional variable for having a high spatial variation. The Pillai's Trace test for standard, backward and forward mode gave a trace value of 1.414 and $p < 0.0001$, 1.401 and $p < 0.0001$, and 1.391 and $p < 0.0001$ respectively. The null hypothesis, H_0 states that the mean vectors of the three classes are equal. The alternative hypothesis, H_a state that at least one of the means vector is different from others. As the computed p-value is lower than the significance level alpha (0.05), one should reject the null hypothesis H_0, and accept the alternative hypothesis H_a. The risk to reject the H_0 while it is true is lower than 0.01%. Fig. 4 shows the box and whisker plots of three significant regions of these air quality variables over the 12-month (January–December) in 2007. Nine selected air quality variables that showed high spatial variations (the most significant which their p-value were less than 0.05) in DA with backward stepwise mode were then applied for further discussion.

Identification Source of Variation

PCA was applied on the air quality data set to evaluate the compositional patterns among the studied air quality variables and to recognize the factors that manipulate each of the discovered regions (LPS, MPS and SHPS). Four PCs were obtained in LPS and MPS region, respectively; and five PCs in the SHPS region with the eigenvalue higher than 1. The total variance for LPS, MPS, and SHPS region were cumulated to 77.20%, 88.71%, and 79.95%, respectively. Four VFs were obtained for LPS and MPS regions, and five VFs were obtained for the SHPS region by means of FA carried out on the PCs. The finding of VFs, loadings of variables, and variance are illustrated in Table 3.

Table 2. Classification matrix for DA of spatial variations across Peninsular Malaysia.

Sampling Regions	% Correct	Regions assigned by the DA		
		SHPS	MPS	LPS
Standard DA mode (14 variables)				
SHPS	97.22	70	0	2
MPS	91.67	1	11	0
LPS	95.24	4	0	80
Total	95.83	75	11	82
Stepwise forward DA mode (8 variables)				
SHPS	91.67	66	0	6
MPS	91.67	1	11	0
LPS	96.43	3	0	81
Total	94.05	70	11	87
Stepwise backward DA mode (9 variables)				
SHPS	93.06	67	0	5
MPS	91.67	1	11	0
LPS	95.24	4	0	80
Total	94.05	72	11	85

(i) Low Pollution Source (LPS) Region

In the LPS region, VF1 contributes about 41.73% of the total variance and has strong positive loadings on CO, NO_2, non-methane hydrocarbons, NO and NO_x. VF1 is associated with fossil fuel combustion. The presence of CO, NO_2, NO and NO_x are related to the fossil fuel combustion of agricultural systems (Mukhopadhyay and Forssell, 2005), while the presence of non-methane hydrocarbons are related to the fossil fuel combustion of transportation (Kopmann 2007). Additional carbon can be sequestered as the effect of nitrogen deposition caused by agricultural practices (De Vries *et al.*, 2006). This assumption is realistic, as the air quality in this region is good and most activities are restricted to agriculture and transportation. VF2 contributes 16.14% of the total variance which has strong positive loadings on methane and wind speed. Strong negative loading is also shown by O_3. VF2 is associated with biogenic emissions. The emission of CH_4 is commonly occurring at the peat swamp area. Most of the LPS regions are located nearby the coastal area. The CH_4 and O_3 are closely correlated and near-simultaneous, though opposite in sign. The processes that had led to the accumulation of CH_4 appeared to have led to the depletion of O_3, to be precise, accumulation and depletion under a shallow night-time inversion (Simmonds *et al.*, 2005). VF3 and VF4 contribute 11.07% and 8.25% of the total variance, respectively; have a strong positive loading on ultra-violet B and ambient temperature and strong negative loading on humidity, which are considered as meteorological factors. When ultra-violet B intensity is increased, automatically the ambient temperature is increased. However, the humidity will decrease due to the evaporation process. Despite of emission sources, ambient air quality can be strongly influenced by meteorological factors through the complex relations between diverse processes - emissions, transport, chemical transformation and wet and dry deposition (Demuzere *et al.*, 2009).

(ii) Moderate Pollution Source (MPS) Region

In the MPS region, VF1 contributes 50.03% of the total variance and has strong positive loadings on CO, NO_2, SO_2, NO and NO_x; and strong negative loading on PM_{10}. VF1 could be related to the composition of chemicals for a range of anthropogenic activities that comprise point source pollution, particularly from industrial, residential, and vegetation areas in MPS region. Most of the pollutants in the MPS region are originated from burning of biomass and fossil fuels, particularly from industrial, residential and vegetation areas, motor vehicles, and natural emission sources (Mutalib *et al.*, 2013; Azid *et al.*, 2014b). VF2 contributes 18.01% of the total variance and proves strong positive loadings on non-methane hydrocarbons and total hydrocarbons, which are pointed to mobile source of pollution (Kopmann, 2007). Access route for land transportation has been developed rapidly in the MPS region recently which makes the number of transportation on the road increased drastically. VF3 contributes 12.04% of the total variance and proves strong positive loadings on ultra-violet B and wind speed. VF3 is commonly related to meteorological factor. The life cycle of pollutants is influenced by chemical and meteorological factors, such as wind speed, temperature, precipitation, and solar radiation (Giorgi and Meleux, 2007). VF4 contributes 8.64% of the total variance, and has a strong positive loading on O_3, which is related to small-scale fossil fuel combustion.

(iii) Slightly High Pollution Source (SHPS) Region

In the SHPS region, VF1, VF2, VF4 and VF5 contribute 28.26%, 20.21%, 10.29% and 9.40% of the total variance, respectively. They have strong positive loadings on CO, NO_2, NO, NO_x, CH_4, total hydrocarbons, O_3 and SO_2. These factors contain chemical compositions that are involved with fossil fuel combustion in various means. The combustion of these fuels in industries and vehicles has been a main source of air pollution (Romieu and Hernandez, 1999; Mutalib *et al.*, 2013). VF3 contributes 11.80% of the total variance and has a strong positive loading on humidity, and strong negative loading on ambient temperature and wind speed. VF3 is associated with meteorological factor. Air

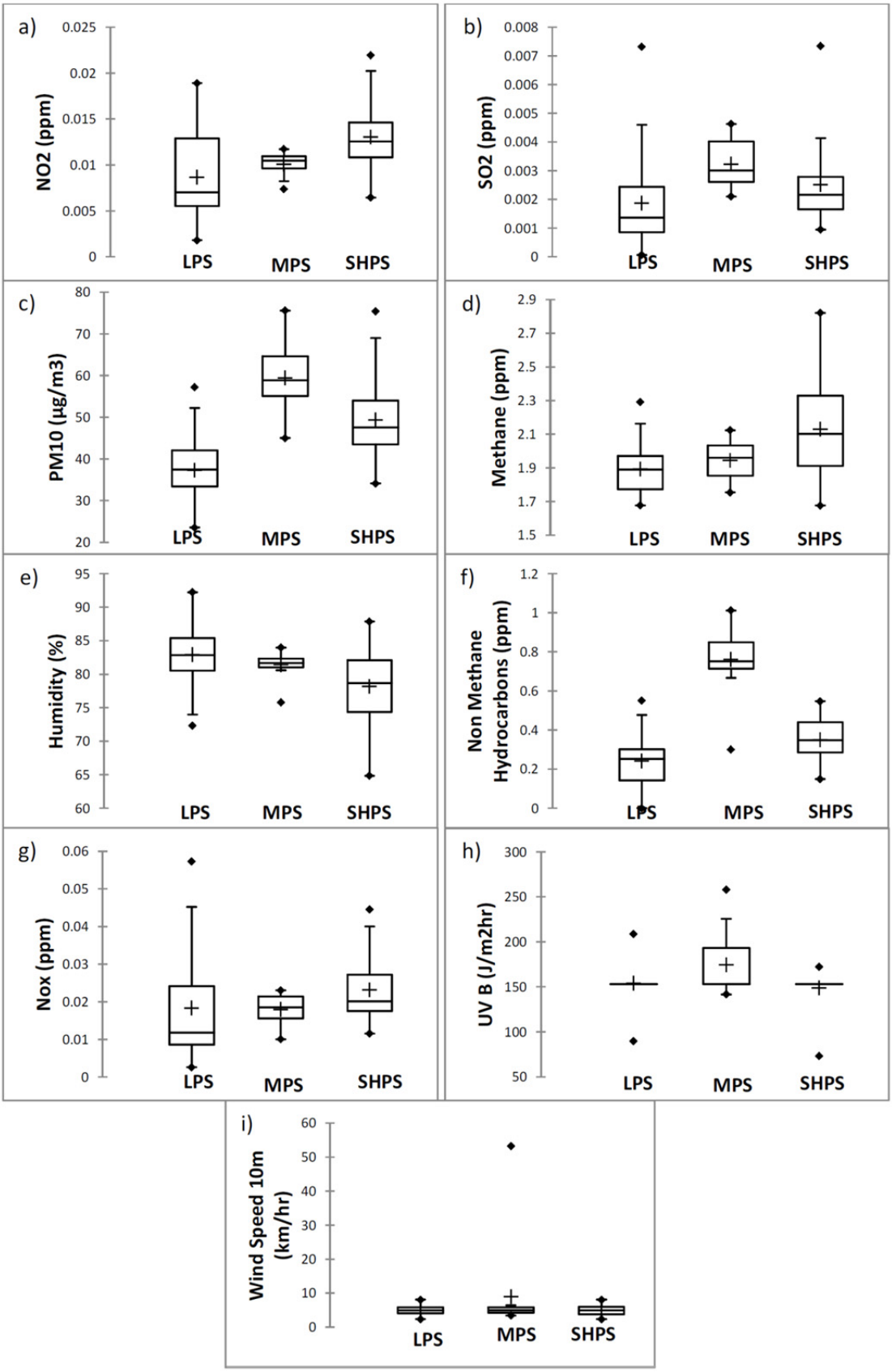

Fig. 4. Box and whisker plots of (a) NO_2, (b) SO_2, (c) $PM_{10,}$ (d) Methane, (e) Humidity, (f) Non Methane Hydrocarbons, (g) NO_x, (h) Ultraviolet B, and (i) Wind Speed generated by backward stepwise DA related to air quality across Peninsular Malaysia.

Table 3. Loadings of environmental variables on the varimax-rotated PCs for water quality data collected from LPS, MPS and SHPS of the Peninsular Malaysia.

Variables	LPS				MPS				SHPS				
	VF1	VF2	VF3	VF4	VF1	VF2	VF3	VF4	VF1	VF2	VF3	VF4	VF5
CO	**0.896**	0.012	0.184	–0.135	**0.933**	–0.020	0.191	0.247	**0.801**	0.245	–0.030	0.330	–0.272
NO_2	**0.939**	0.046	–0.067	0.003	**0.773**	0.048	–0.063	0.524	**0.747**	–0.218	0.239	0.417	0.213
SO_2	0.697	–0.022	–0.491	0.256	**0.906**	–0.240	0.023	0.116	0.108	0.436	–0.077	0.03	**0.766**
PM_{10}	0.646	0.432	–0.237	0.296	**–0.748**	–0.046	0.076	0.434	–0.215	0.514	0.027	0.179	0.508
O_3	0.343	**–0.749**	0.093	0.097	0.213	–0.277	–0.013	**0.827**	0.024	–0.024	0.139	**0.888**	–0.024
Ambient Temp	–0.300	–0.178	0.189	**0.852**	0.000	0.000	0.000	0.000	0.090	–0.180	**–0.752**	0.139	0.206
CH_4	0.263	**0.746**	–0.154	–0.228	**0.913**	0.312	0.012	–0.009	–0.154	**0.956**	0.017	0.023	–0.047
Humidity	–0.251	0.344	**–0.756**	–0.167	0.419	0.553	0.068	0.457	0.426	0.049	**0.768**	0.216	–0.069
Non Methane Hydrocarbons	**0.887**	–0.04	0.143	–0.046	–0.120	**0.941**	–0.066	–0.124	0.218	–0.061	0.373	–0.333	0.589
NO	**0.873**	–0.33	0.071	–0.281	**0.857**	0.253	0.387	0.029	**0.918**	–0.157	0.071	–0.248	0.070
NO_x	**0.921**	–0.236	0.04	–0.222	**0.881**	0.268	0.300	0.185	**0.945**	–0.192	0.141	–0.019	0.117
Total Hydrocarbons	0.591	0.403	0.163	–0.031	0.382	**0.88**	–0.056	–0.103	–0.066	**0.911**	0.159	–0.105	0.185
UV B	0.073	0.078	**0.769**	0.189	0.499	–0.048	**0.799**	0.131	0.168	–0.395	–0.159	0.432	0.458
Wind Speed	–0.403	**0.749**	0.047	–0.047	0.020	–0.075	**0.961**	–0.077	–0.118	–0.063	**–0.884**	–0.097	–0.185
Eigenvalues	5.92	2.49	1.37	1.02	6.5	2.34	1.56	1.12	3.96	2.83	1.65	1.44	1.32
Variability (%)	41.73	16.14	11.07	8.25	50.03	18.01	12.04	8.64	28.26	20.21	11.8	10.29	9.4
Cumulative (%)	41.73	57.88	68.95	77.2	50.03	68.04	80.08	88.71	28.26	48.47	60.27	70.56	79.95

pollutant chemical reactions rely on ambient air states and are normally manipulated by short-wave radiation, air temperature, wind speed, wind direction and relative humidity (Elminir, 2005). It is tremendously vital to consider the consequence of meteorological states on air pollution, because they directly influence the emission effect of the atmosphere.

Multiple Linear Regression Result (MLR) of Air Pollutant Index (API)

The purpose of MLR modelling in this study is to describe the behaviour of the variables, which is based on a linear least-square fitting process, and a trace element or property is required to be determined for each source (Henry *et al.*, 1984). In this study, the source of apportionment of air pollutant parameters (independent variable) was used to identify the potential of the total API (dependent variable) in the study area. Four models were developed. To develop the models, the API values were used as dependent variable, while the independent variables were the air quality parameters (using original air quality parameters (14 variables), air quality parameters from LPS (9 variables), air quality parameters from MPS (9 variables), and air quality parameters from SHPS (9 variables)).

Due to the importance of R^2, adjusted R^2, and RMSE values for better coefficient results, the finding of the study shows that the values of R^2, adjusted R^2 and RMSE for the original air quality parameters-API were 0.873, 0.861, and 3.108 respectively from the goodness of fit statistics. The values of R^2, adjusted R^2 and RMSE for LPS were 0.865, 0.846, and 2.187 respectively. Followed by the values of R^2, adjusted R^2 and RMSE for MPS were 0.999, 0.993, and 1.430 respectively. Meanwhile, the values of R^2, adjusted R^2 and RMSE for SHPS were 0.868, 0.852, and 2.195 respectively. The proposed equation with R^2, adjusted R^2, and RMSE can be seen in Eqs. (6*i*)–(6*iv*):

a) Original air quality parameters (14 variables)
$$\text{Total API} = -0.15(CO) - 501.09(NO_2) - 210.13(SO_2) + 0.70(PM_{10}) + 58.94(O_3) + 0.19(\text{Temp}) - 0.24(CH_4) - 0.14(\text{Humidity}) - 7.94(\text{NMHC}) - 576.91(\text{NO}) + 597.93(NO_x) + 1.26(\text{THC}) + 6.43e^{-03}(\text{UVB}) - 0.88(\text{Wind Speed 10m}) + 18.30 \quad [R^2 = 0.873, \text{adjusted } R^2 = 0.861, \text{and RMSE}=3.108] \quad (6i)$$

b) LPS (9 variables)
$$\text{Total API} = -524.57(NO_2) + 59.48(SO_2) + 0.82(PM_{10}) + 2.71(CH_4) - 6.99e^{-02}(\text{Humidity}) + 6.70(\text{NMHC}) + 87.74(NO_x) + 1.47e^{-}02(\text{UVB}) - 0.36(\text{Wind Speed 10 m}) + 8.72 \quad [R^2 = 0.865, \text{adjusted } R^2 = 0.846, \text{and RMSE} = 2.187] \quad (6ii)$$

c) MPS (9 variables)
$$\text{Total API} = -8216.13(NO_2) + 1803.87(SO_2) + 1.61(PM_{10}) - 49.44(CH_4) - 2.18(\text{Humidity}) - 13.64(\text{NMHC}) + 5478.53(NO_x) - 3.95e^{-02}(\text{UVB}) - 1.28(\text{Wind Speed 10 m}) + 229.68 \quad [R^2 = 0.999, \text{adjusted } R^2 = 0.993, \text{and RMSE} = 1.430] \quad (6iii)$$

d) SHPS (9 variables)
Total API = 332.99(NO_2) – 241.01(SO_2) + 0.60(PM_{10}) + 1.50(CH_4) – 0.10(Humidity) + 2.00(NMHC) – 54.50(NO_x) + 2.13e^{-02}(UVB) – 0.15(Wind Speed 10 m) + 16.37 [R^2 = 0.868, adjusted R^2 = 0.852, and RMSE = 2.195] (6*iv*)

Based on the Eqs. (6*i*)–(6*iv*), Cluster of MPS shows the highest coefficient of determination, R^2 (0.999) contributed by the nine air pollutant parameters. The daily average concentrations of NO_2, CH_4, Humidity, NMHC, UVB, and wind speed 10 m have a negative influence on the total API value, in contrast to the average concentration of SO_2, NO_x, and PM_{10}. The second highest is from original air quality parameters (14 air pollutant parameter) model with the R^2 value of 0.873. The concentrations of CO, NO_2, SO_2, CH_4, Humidity, NMHC, NO, and wind speed 10 m show a negative influence compared to O_3, PM_{10}, NO_x, Ambient Temperature, THC, and UVB. The third highest is in Cluster SHPS with R^2 value is 0.868, in which SO_2, Humidity, NO_x, and wind speed 10 m show a negative influence on the total API while NO_2, PM_{10}, CH_4, NMHC, and UVB positively influenced to the total API. Meanwhile, Cluster LPS is the lowest of R^2 value (0.865) in this study. Apart from NO_2, Humidity, and Wind speed 10 m, it is positively influenced by the SO_2, PM_{10}, CH_4, NMHC, NO_x, and UVB. From the finding, Cluster MPS has been selected as the best model due to the smallest RMSE and the closest R^2 value of 1 when compared among tested parameters. This is because the better model shall be performed if the value of RMSE is smaller than the other and closest of R^2 value to 1 (Dominick *et al.*, 2012; Mutalib *et al.*, 2013).

Fig. 5 depicts the bar chart of standardized coefficient values for all three clusters and original air quality parameters (14 independent variables) for total API's linear regression model. The outcome shows that PM_{10} account as the most and the highest pollution contributor to Malaysian atmospheric quality, which it is related to the chemical compositions from various anthropogenic activities such as industrial, residential, and agricultural areas.

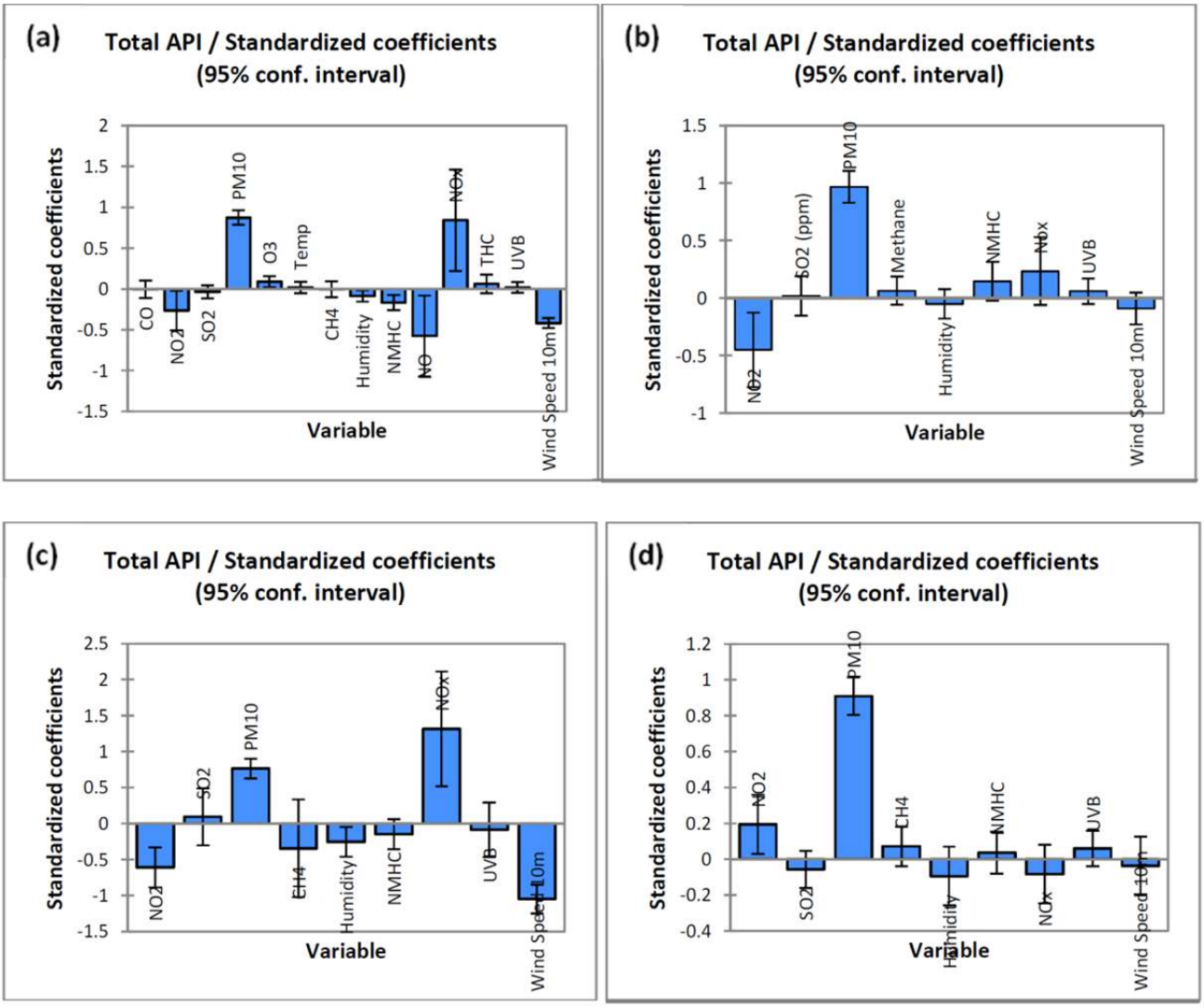

Fig. 5. Bar chart of the standardized coefficient for the independent variables: (a) Original air quality parameters, (b) LPS, (c) MPS, and (d) SHPS.

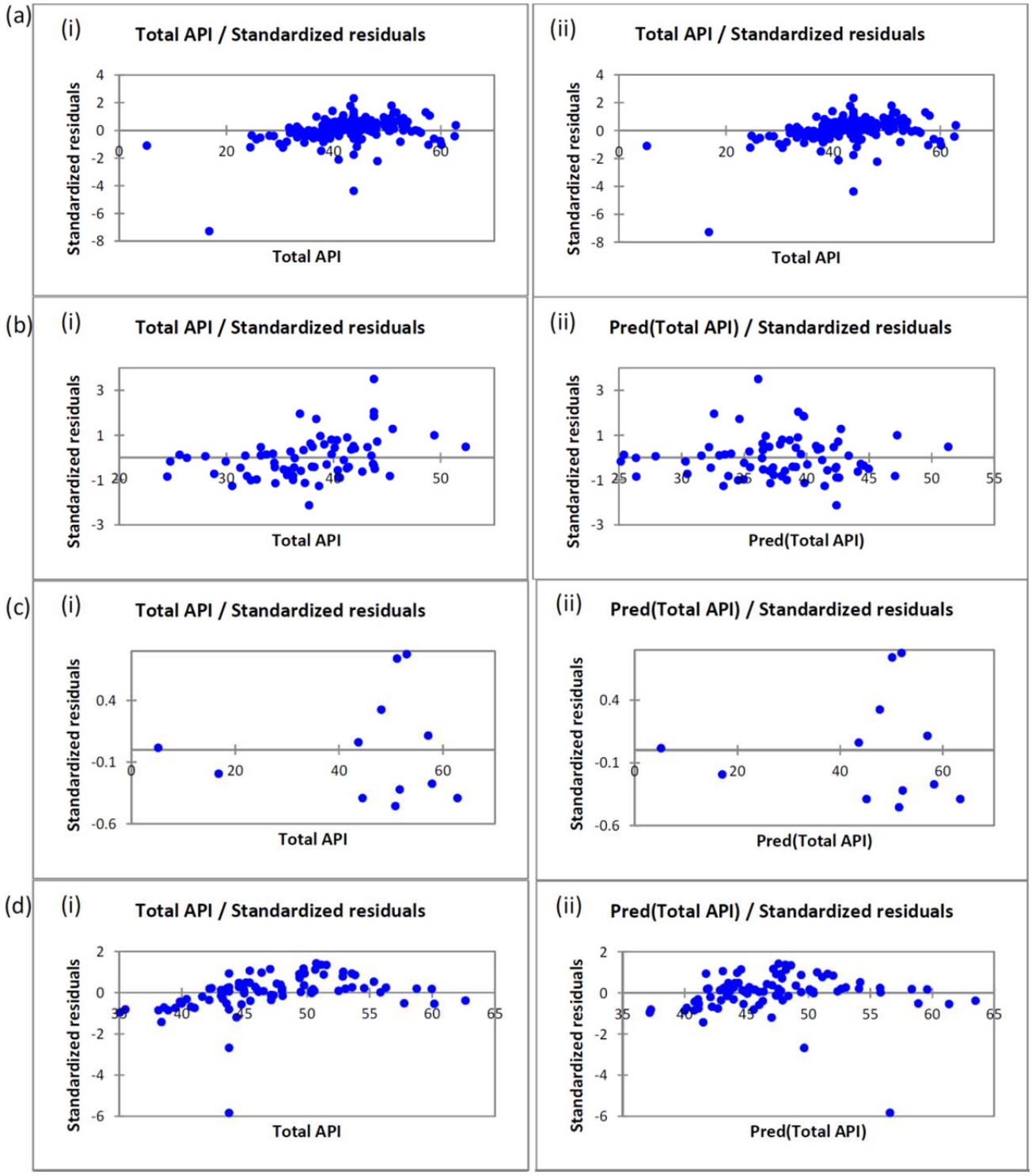

Fig. 6. Scatter plot diagram of standardized residuals of (i) actual total API, and (ii) predicted total API for: (a) original air quality parameter model, (b) LPS model, (c) MPS model, and (d) SHPS model.

Fig. 6 represents the residual analysis of the observed and predicted of the total API using the MLR modelling for original air quality parameters and 3 clusters. The findings have shown that the deficiency of the model for original air quality parameters, LPS, MPS, and SHPS, which the data sets indicate a great difference in the range of –8 to 4, –3 to 4, –0.6 to 0.8, and –6 to 2, respectively. The verification of the model was influenced by the outlier observation as illustrated in Fig. 7, which from the actual total API indicates that some of the observations were out from the 95% of the confidence interval range (lower and upper boundary) especially for the model of original air quality parameters, LPS, and SHPS, but contrast to MPS model. The main objective of plotting this graph is to prove that the MLR

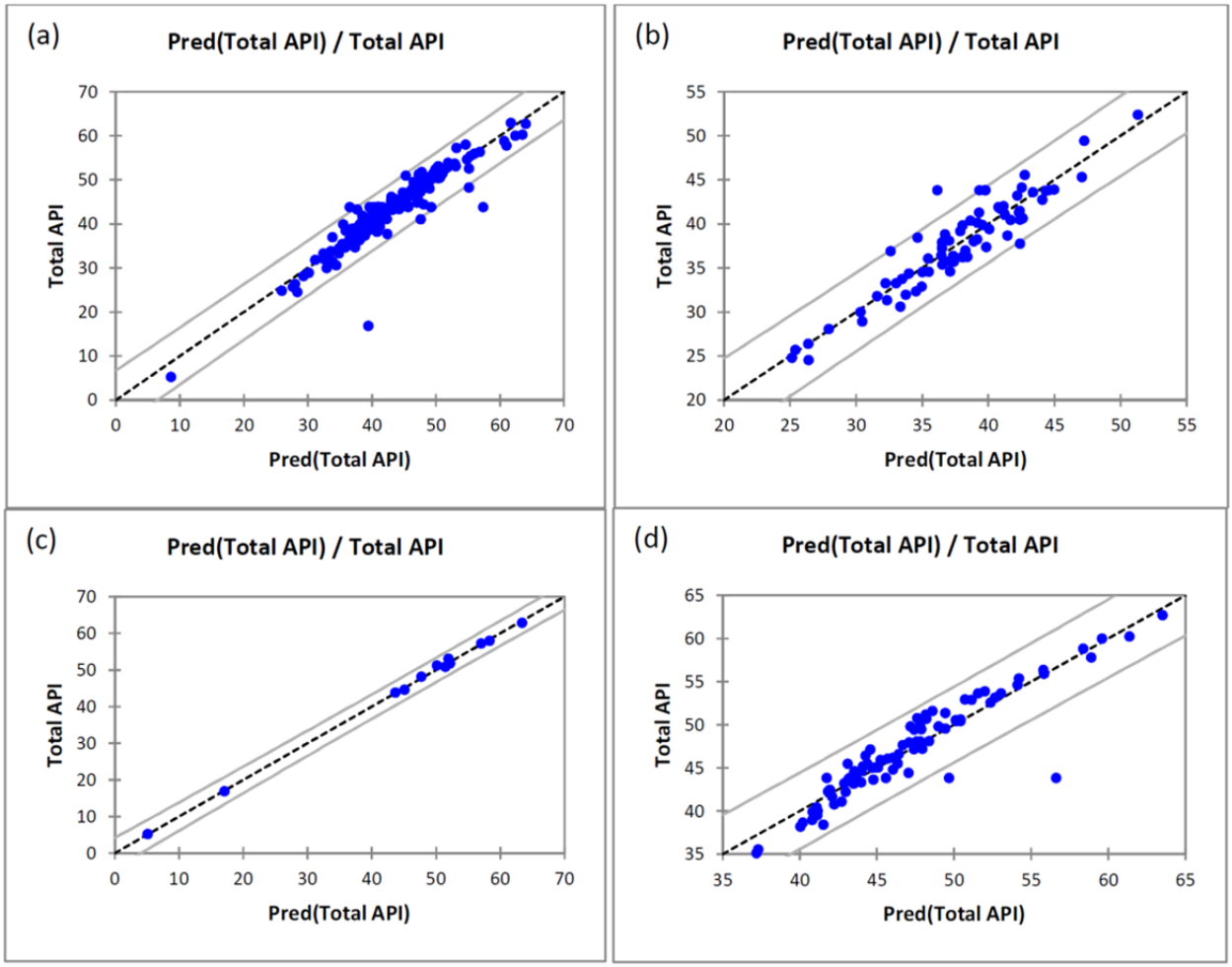

Fig. 7. Scatter plot diagram of total air pollution index (predicted) versus actual total air pollution index graph for: (a) original air quality parameter model, (b) LPS model, (c) MPS model, and (d) SHPS model.

model is suitable to be used for total API prediction, because it gives the great difference between predicted total API and calculated total API.

CONCLUSION

From this study, it can be concluded that the spatial variations of air quality data in Peninsular Malaysia were successfully studied by applying chemometric procedures, namely, HACA, DA, PCA, FA and MLR. By using HACA, 14 monitoring stations were well grouped into three significant diverse cluster regions, known as LPS, MPS, and SHPS. Based on the finding from HACA, a better monitoring network approach can be proposed which could lessen the quantity of monitoring stations. The grouped regions made by HACA were confirmed by DA with 94.05% accuracy of spatial variation for both forward and backward stepwise modes. Eight discriminant variables were selected for forward stepwise mode while nine discriminant variables were selected for backward stepwise mode. The nine variables obtained from backward stepwise mode can be used for the new design of an air quality monitoring network instead of taking fourteen air quality variables into account. To identify the source of air pollution, PCA and FA were done. Four VFs were found for the LPS and MPS regions, with total variance of 77.20% and 88.71%, respectively. In the SHPS region, with the total variance of 79.95%, only five VFs were obtained. In this study, the sources of variations are expected derived from industrial emissions, transportation emissions, agricultural systems, fuel combustions of peat swamp, and meteorological factors. For LPS and MPS regions, four variables were identified to be dependable for the major variations. For SHPS region, five variables were identified to be dependable for the major variations. Based on the PCA, air pollution sources are expected to come from fuel combustion of peat swamp, transportation emissions, large-scale agricultural systems and meteorological factors in the LPS region. The air pollution sources in MPS region are related to transportation emissions, small or medium industrial emissions, small-scale agricultural systems and meteorological factors. The major sources of variations in the SHPS region are expected derived from large-scale industrial emissions, transportation emissions, and meteorological factors. Multiple Linear Regressions (MLR) analysis was done to identify the variability of the proposed equation to predict values of the

total API. When comparing from four models developed, the R^2 values were found to be strong because they were high and significant at p-value (< 0.05). The MPS model shows the highest R^2 with the value of 0.999, followed by the original air quality parameter, SHPS, and LPS model with the value of 0.873, 0.868, and 0.865, respectively. In this study, the finding also shows that PM_{10} contributes the most of the total API in the atmosphere compared to the other pollutants and this pollutant can be categorized as the primary pollutant in Malaysia. For a better and effective air quality management, a new air quality monitoring network should be designed in term of practical and cost-effective.

ACKNOWLEDGEMENT

The authors are grateful to the Department of Environment (DOE) for the supply of air quality data required for the completion of this study.

REFERENCES

Abdullah, A.M., Samah, M.A.A. and Tham, Y.J. (2012). An Overview of the Air Pollution Trend in Klang Valley, Malaysia. *Open Environ. Sci.* 6: 13–19.

Aertsen, W., Kinta, V., Orshovena, J., Özkan, K. and Muysa, B. (2010). Comparison And ranking of Different Modelling Techniques for Prediction of Site Index in Mediterranean Mountain Forests. *Ecol. Modell.* 221: 1119–1130.

Afroz, R., Hassan, M.N. and Ibrahim, N.A. (2003). Review of Air Pollution and health impacts in Malaysia. *Environ. Res.* 92: 71–77, doi: 10.1016/S0013-9351(02)00059-2.

Almeida, J.A.S., Barbosa, L.M.S., Pais, A.A.C.C. and Formosinbo, S.J. (2007). Improving Hierarchical Cluster Analysis: A New Method with Outlier Detection and Automatic Clustering. *Chemom. Intell. Lab. Syst.* 87: 208–217.

Azid, A., Juahir, H., Latif, M.T., Zain, S.M. and Osman, M.R. (2013). Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia. *J. Environ. Prot.* 4: 1–10, doi: 10.4236/jep.2013.412A001.

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. and Yamin, M. (2014a). Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: A Case Study in Malaysia. *Water Air Soil Pollut.* 225: 2063, doi: 10.1007/s11270-014-2063-1.

Azid, A., Juahir, H., Aris, A.Z., Toriman, M.E., Latif, M.T., Zain, S.M., Yusof, K.M.K.K. and Saudi, A.S.M. (2014b). Spatial Analysis of the Air Pollutant Index in the Southern Region of Peninsular Malaysia Using Environmetric Techniques. In *From Sources to Solution*, Proceeding of the International Conference on Environmental Forensics 2013, Aris, A.Z., Ismail, T.H.T., Harun, R., Abdullah, A.M. and Ishak, M.Y. (Eds.), Springer, New York, p. 307, doi: 10.1007/978-981-4560-70-2_56.

Azid, A., Juahir, H., Toriman, M.E., Endut A., Kamarudin, M.K.A., Rahman, M.N.A., Hasnam, C.N.C., Saudi, A.S.M. and Yunus, K. (2015). Source Apportionment of Air Pollution: A Case Study In Malaysia. *Jurnal Teknologi* 72: 83–88. doi: 10.11113/jt.v72.2934

Azmi, S.Z., Latif, M.T., Ismail, A.S., Juneng, L. and Jemain, A.A. (2010). Trend and Status of Air quality at Three Different Monitoring Stations in the Klang Valley, Malaysia. *Air Qual. Atmos. Health* 3: 53–64, doi: 10.1007/s11869-009-0051-1.

Bock, H.H. (1996). Probabilistic Models in Cluster Analysis. *Comput. Stat. Data Anal.* 23: 5–28, doi: 10.1016/0167-9473(96)88919-5.

De Vries, W., Butterbach B.K., Denier V.D.G.H. and Oenema, O. (2006). The Impact of Atmospheric Nitrogen Deposition on the Exchange of Carbon Dioxide, Nitrous Oxide and Methane from European Forests. *Global Change Biol.* 12: 1151–1173.

Demuzere, M., Trigo, R.M., Vila-Guerau, D.A.J. and Van L.N.P.M. (2009). The Impact of Weather and Atmospheric Circulation on O_3 and PM_{10} Levels at a Rural Mid-latitude Site. *Atmos. Chem. Phys.* 9: 2695–2714.

Department of Environment (DOE) (2013). *Air Pollutant Index (API).* Available from: http://www.doe.gov.my/webportal/en/info-umum/english-air-pollutant-index-api/.

Dominick, D., Juahir, H., Latif, M.T., Zain, S.M. and Aris, A.Z. (2012). Spatial Assessment of Air Quality Patterns in Malaysia Using Multivariate Analysis. *Atmos. Environ.* 60: 172–181.

Elminir, H. (2005). Dependence of Urban Air Pollutants on Meteorology. *Sci. Total Environ.* 350: 225–237.

Giorgi, F. and Meleux, F. (2007). Modelling the Regional Effects of Climate Change on Air Quality. *C.R. Geosci.* 339: 721–733, doi: 10.1016/j.crte.2007.08.006.

Henry, R.C., Lewis, C.W., Hopke, P.K. and Williamson, H.J. (1984). Review of Receptor Model Fundamentals. *Atmos. Environ.* 18: 1507–1515.

Hopke, P.K. (1985). *Receptor Modelling in Environmental Chemistry*. Wiley, New York.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. 3rd ed. Prentice-Hall Int., New Jersey.

Juahir, H., Zain, S.M., Yusoff, M.K., Hanidza, T.I.T., Armi, A.S.M., Toriman, M.E. and Mokhtar, M. (2011). Spatial Water Quality Assessment of Langat River Basin (Malaysia) Using Chemometric Techniques. *Environ. Monit. Assess.* 173: 625–641, doi: 10.1007/s10661-010-1411-x.

Kamal, M.M., Jailani, R. and Shauri, R.L.A. (2006). Prediction of Ambient Air Quality Based on Neural Network Technique. *4th Student Conference on Research and Development*, p. 115–119. doi: 10.1109/SCORED.2006.4339321.

Kannel, P.R., Lee, S., Kanel, S.R. and Khan, S.P. (2007). Chemometric Application in Classification and Assessment of Monitoring Locations of an Urban River System. *Anal. Chim. Acta* 582: 390–399.

Koppmann, R. (2007). *Volatile Organic Compounds in the Atmosphere*. Blackwell Publishing Ltd., Singapore.

Lioy, P.J., Zelenka, M.P., Cheng, M.D. and Reiss, N.M.

(1989). The effect of Sampling Duration on the Ability to Resolve Source Types Using Factor Analysis. *Atmos. Environ.* 23: 239–254.

Liu, C.W., Lin, K.H. and Kuo,Y.M. (2003). Application of Factor Analysis in the assessment of Groundwater Quality in a Blackfoot Disease Area in Taiwan. *Sci. Total Environ.* 313: 77–89, doi: 10.1016/S0048-9697(02)00683-6.

Lu, W.Z., He, H.D. and Dong, L.Y. (2011). Performance Assessment of Air Quality Monitoring Networks Using Principal Component Analysis and Cluster Analysis. *Build. Environ.* 46: 577–583, doi: 10.1016/j.buildenv.2010.09.004.

Manjunath, B.G., Frick, M. and Reiss, R.D. (2012). Some Notes on Extremal Discriminant Analysis. *J. Multivar. Anal.* 103: 107–115, doi: 10.1016/j.jmva.2011.06.012.

Moustris, K.P., Ziomas, I.C. and 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 Pollut.* 209: 29–43.

Mukhopadhyay, K. and Forssell, O. (2005). An Empirical Investigation of Air Pollution from Fossil Fuel Combustion and Its Impact on Health in India during 1973–1974 to 1996–1997. *Ecol. Econ.* 55: 235–250, doi: 10.1016/j.ecolecon.2004.09.022.

Mutalib, S.N.S.A., Juahir, H., Azid, A., Sharif, S.M., Latif, M.T., Aris, A.Z., Zain, S.M. and Dominick, D. (2013). Spatial and Temporal Air Quality Pattern Recognition Using Environmetric Techniques: A Case Study in Malaysia. *Environ. Sci. Processes Impacts* 15: 1717–1728, doi: 10.1039/c3em00161j.

Norusis, M.J. (1990). *SPSS Base System User's Guide*. USA: SPSS, Chicago, IL.

Pai, T.Y., Sung, P.J., Lin, C.Y., Leu, H.G., Shieh, Y.R., Chang, S.C., Chen, S.W. and Jou, J.J. (2009). Predicting Hourly Ozone Concentration in Dali Area of Taichung Country Based on Multiple Linear Regression Method. *Int. J. Appl. Sci. Eng.* 7: 127–132.

Romieu, I. and Hernandez, M. (1999). *Air Pollution and Health in Developing Countries: Review of Epidemiological Evidence*. In: Mc Granahan, G. and Murray, F. (Eds.), Health and Air Pollution in Rapidly Developing Countries. Stockholm Environment Institute, Sweden, p. 43– 6.

Satheeshkumar, P. and Khan, A.B. (2011). Identification of Mangrove Water Quality by Multivariate Statistical Analysis Methods in Pondicherry Coast, India. *Environ. Monit. Assess.* 184: 3761–3774, doi: 10.1007/s10661-011-2222-4.

Shrestha, S. and Kazama, F. (2007). Assessment of Surface Water Quality Using Multivariate Statistical Techniques: A Case Study of the Fuji River Basin, Japan. *Environ. Modell. Softw.* 22: 464–475, doi: 10.1016/j.envsoft.2006.02.001.

Simeonov, V., Einax, J.W., Stanimirova, I. and Kraft, J. (2002). Envirometric Modeling and Interpretation of River Water Monitoring Data. *Anal. Bioanal. Chem.* 374: 898–905, doi: 10.1007/s00216-002-1559-5.

Simmonds, P.G., Manning, A.J., Derwent, R.G., Ciais, P., Ramonent, M., Kazan, V. and Ryall, D. (2005). A Burning Question. Can Recent Growth Rate Anomalies in the Greenhouse Gases by Attributed to Large-scale Biomass Burning Events? *Atmos. Environ.* 39: 2513–2517.

Singh, K.P., Malik, A., Mohan, D. and Sinha, S. (2004). Multivariate Statistical Techniques for the Evaluation of Spatial and Temporal Variations in Water Quality of Gomti River (India): A Case study. *Water Res.* 38: 3980–3992, doi: 10.1016/j.watres.2004.06.011.

Singh, K.P., Malik, A. and Sinha, S. (2005). Water Quality Assessment and Apportionment of Pollution Sources of Gomti River (India) Using Multivariate Statistical Techniques: A Case Study. *Anal. Chim. Acta* 35: 3581–3592.

Soares, P.K., Bruns, R.E. and Scarminio, I.S. (2008). Statistical Mixture Design - Varimax Factor Optimization for selective Compound Extraction from Plant Material. *Anal. Chim. Acta* 613: 48–55. doi: 10.1016/j.aca.2008.02.051.

Ul-Saufie, A.Z., Ahmad Shukri, Y., Nor Azam, R. and Hazrul, A.H. (2011). Comparison between Multiple Linear Regression and Feed Forward Back Propagation Neural Network Models for Predicting PM_{10} Concentration Level Based on Gaseous and Meteorological Parameters. *Int. J. Appl. Sci. Technol.* 1: 42–49.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* 58: 236–244.

Yu, T.Y. and Chang, L.F.W (2000). Selection of the Scenarios of Ozone Pollution at Southern Taiwan Area Utilizing Principal Component Analysis. *Atmos. Environ.* 34: 4499–4509.

Received for review, April 13, 2014
Revised, September 9, 2014
Accepted, February 7, 2015