MALAYSIAN JOURNAL OF ANALYTICAL SCIENCES

Published by The Malaysian Analytical Sciences Society

ISSN 1394 - 2506

SPATIAL AIR QUALITY MODELLING USING CHEMOMETRICS TECHNIQUES: A CASE STUDY IN PENINSULAR MALAYSIA

(Pemodelan Ruang Kualiti Udara Menggunakan Teknik-Teknik Kemometrik: Satu Kajian Kes di Semenanjung Malaysia)

Azman Azid[1]*, Hafizan Juahir[1], Mohammad Azizi Amran[1], Zarizal Suhaili[2], Mohamad Romizan Osman[3], Asyaari Muhamad[1,4], Ismail Zainal Abidin[1], Nur Hishaam Sulaiman[1], Ahmad Shakir Mohd Saudi[1]

[1]East Coast Environmental Research Institute,
Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia
[2]Faculty of Bioresources and Food Industry,
Universiti Sultan Zainal Abidin, Tembila Campus, 22200 Besut, Terengganu, Malaysia
[3]Kulliyyah of Science,
International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia
[4]Institute of the Malay World and Civilisation (ATMA),
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

**Corresponding author: azmanazid@unisza.edu.my*

Received: 14 April 2015; Accepted: 9 July 2015

Abstract

This study shows the effectiveness of hierarchical agglomerative cluster analysis (HACA), discriminant analysis (DA), principal component analysis (PCA), and multiple linear regressions (MLR) for assessment of air quality data and recognition of air pollution sources. 12 months data (January-December 2007) consisting of 14 stations in Peninsular Malaysia with 14 parameters were applied. Three significant clusters - low pollution source (LPS), moderate pollution source (MPS), and slightly high pollution source (SHPS) were generated via HACA. Forward stepwise of DA managed to discriminate eight variables, whereas backward stepwise of DA managed to discriminate nine variables out of fourteen variables. The PCA and FA results show the main contributor of air pollution in Peninsular Malaysia is the combustion of fossil fuel from industrial activities, transportation and agriculture systems. Four MLR models show that PM_{10} account as the most and the highest pollution contributor to Malaysian air quality. From the study, it can be stipulated that the application of chemometrics 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 novelty design of air quality monitoring network for better management of air pollution can be achieved *via* these methods.

Keywords: air quality, chemometrics, pattern recognition, Peninsular Malaysia

Abstrak

Kajian ini menunjukkan keberkesanan kaedah hirarki algorithma analisa kelompok (HAAK), analisis pembezalayan (AP), analisis komponen utama (AKU), dan kepelbagaian regresi linear (KRL) untuk penilaian data kualiti udara dan pengenalpastian punca pencemaran udara. Data 12 bulan (Januari-Disember 2007) terdiri daripada 14 stesen di Semenanjung Malaysia dengan 14 parameter telah digunakan. Tiga kelompok besar - sumber pencemaran rendah (SPR), sumber pencemaran sederhana (SPS), dan sumber pencemaran sedikit tinggi (SPST) diwujudkan melalui HAAK. Melalui AP, kaedah langkah demi langkah ke hadapan berjaya membezalayan lapan pembolehubah, manakala kaedah langkah demi langkah kebelakang berjaya membezalayan sembilan pembolehubah daripada 14 belas pembolehubah. Keputusan AKU menunjukkan bahawa penyumbang utama pencemaran udara di Semenanjung Malaysia adalah disebabkan oleh pembakaran bahan api fosil melalui aktiviti perindustrian, pengangkutan dan sistem pertanian. Empat model KRL menunjukkan bahawa PM_{10} bertindak sebagai penyumbang utama

kepada pencemaran udara Malaysia. Dari kajian ini, ia dapat membuktikan bahawa penggunaan teknik kemometrik boleh memberikan maklumat yang bermakna terhadap kebolehubahan ruang bagi data yang besar dan kompleks. Kajian yang lebih jelas mengenai kualiti udara dan rangkaian pemantauan reka bentuk kualiti udara yang baru dalam pengurusan pencemaran udara yang lebih baik dapat dicapai melalui kaedah-kaedah tersebut.

Kata Kunci: kualiti udara, kemometrik, pengenalan corak, Semenanjung Malaysia

Introduction

Air pollution control is needed to prevent the situation from deteriorating in the long term period [1,2]. Therefore, air quality monitoring network is a part of the preliminary strategy for the air pollution deterrence plan in Malaysia. A properly-designed of 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. Consequently, the application of chemometrics can be utilized to complement the monitoring strategy [3].

Chemometrics in the environmental field is verified to be a functional tool to identify the sources of pollution such as in [2,4,5]. Chemometrics techniques include the interrelationship of faunal structure, physical-chemical characterization, and toxicity data that received from *in-situ* measurement and laboratory analysis. The analysis is considered to be the most suitable tool for the reduction and interpretation of meaningful data [5,6,7]. Unbiased methods such as hierarchical agglomerative cluster analysis (HACA), discriminant analysis (DA), principal component analysis (PCA), and multiple linear regressions (MLR) were applied in air quality analysis.

The application of diverse chemometrics statistical techniques for interpretation of the complex databases, permits a better understanding of air quality in the study region. Chemometrics 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 [2,8]. The objectives of this study are to recognize the pollution sources and identify the most significant pollutant.

Materials and Methods

Study Sites

14 stations (Figure 1 and Table 1) around Peninsular Malaysia were selected as a monitoring site in this study. There are no main natural disasters occurred in Peninsular Malaysia such as typhoon, volcanic eruption and earthquake. These stations were chosen due to the type of region which are urban, suburban, and industrial area.

Data Collection

The data for 14 air quality parameters from 14 stations were gathered from the Department of Environment (DOE), from January to December 2007. Ambient temperature (ºC), 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), sulphur dioxide (SO_2, ppm), total hydrocarbons (THC, ppm), ultra-violet B (J/m2hr) and wind speed (km/hr) were selected to study the influence of API values and the sources of pollution. The hourly data were used to form a monthly average, which comprises 168 datasets (12 data per stations x 14 stations) with a total of 2,352 observations (12 data per stations x 14 variables x 14 stations).

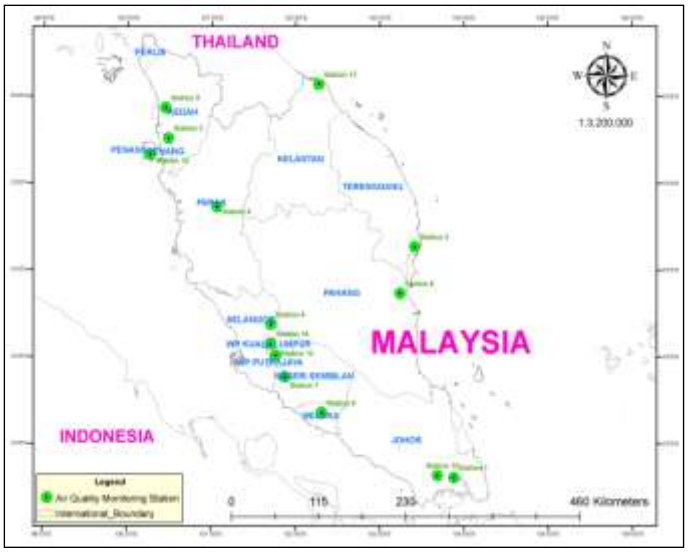


Figure 1. Fourteen selected air quality monitoring stations in the Peninsular Malaysia

Table 1. The details of 14 monitoring stations around Peninsular Malaysia

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	N 03° 06.376	E 101°43.072

Hierarchical Agglomerative Cluster Analysis (HACA)

In this study, HACA was used for clustering the spatial air monitoring station based on the Air Pollutant Index (API) data. HACA is a statistical method that can classify the object in a data set [9], and known as the art of finding groups in data processing [10]. In this method, the samples to be clustered are defined in *n*-dimensional hyperspace and distances are computed accordingly. After a criterion of distance is defined in several algorithms, then it can be used to detect groups of the samples. Samples that are close to each other are expected to be similar in one group [11].

In this study, HACA is employed on the normal distribution dataset through the Ward's method by means of Euclidean distances, as a measure of the relationship [12,13]. The outcome of this method depicted by a *treelike* structure known as a dendrogram. The dendrogram demonstrated a visual summary of the clustering process, presenting a picture of the groups, and their proximity, with a reduction in dimensionality of the original data [14]. The linkage distance by a Euclidean distance accounted as D_{link}/D_{max}, which signifies the measure between the linkages 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 [14]. Euclidean distance can be defined by equation 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.

Discriminant Analysis (DA)

DA is used to classify the object of unknown origin to one of several naturally occurring groups [15]. In this study, DA has been coupled with HACA to establish the significant different variables as well as for reducing the errors in groups such as in [6]. In each cluster, it creates a discriminant functions (DFs) [16], which can be determined by equation 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, the air quality parameters were treated as independent variables, whereas the three significant groups were treated as dependent variables. Three modes of DA were applied, which are standard mode, forward stepwise mode and backward stepwise mode. DFs was created by a standard mode in order to evaluate the 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. 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, the interrelated variables were analysed and interpreted by PCA. Theoretically, PCA is a method of creating new variables (known as principal components, PCs), which are linear composites of the original variables. The values of PCs created by PCA is known as principal component scores (PCS). The maximum number of new variables is equivalent to the number of original variables [12]. PCA can be utilized to identify the emission source [17]. 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 [2, 8,18]. The PCs can be calculated as equation 3:

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

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.

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 PCA. Eigenvalues more than 1.0 were considered as significant and subsequently varimax factors (VFs), which are the new groups of variables are generated [19]. 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 [2,20,21]. In this study, only factor loadings with absolute values greater than 0.75 were selected for the interpretation [12,21]. 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 equation 4:

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

where, *z* is the measured value of a variables, *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, PCA were applied to the classified datasets (14 variables) independently, based on the LPS, MPS and SHPS region that were classed by HACA.

Multiple Linear Regressions (MLR)

MLR is widely used for investigating the relationship among various independent and dependent variables by fitting a linear equation to observed data [22,23,24] and gives the percentage of the contribution of each parameter to the atmospheric pollution [25]. In this study, MLR was used to justify the relationship between the air quality parameters and the API data. The model of the original air quality parameters-API was compared to 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 was calculated using the equation 5:

$$Y_i = \beta_o + \beta_1 x_1 + \beta_2 x_{1i2} ... + \beta_{p-1} \quad (5)$$

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.

The coefficient of determination (R^2) and root mean square error (RMSE) are the components that need to be considered in model performance. The value of R^2 provides information about how well the model performs on external data [26]. RMSE is 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 [5,21,26,27].

Results and Discussion

Spatial Classification of Air Quality by HACA

This part measures the historical values of API step by step to categorize the air quality station based on their homogeneity level by means of HACA. Figure 2 and Figure 3 shows the three significant regions illustrated by HACA and the potential pollution sources within the study regions. Three clusters that generated from the clustering method are known as: 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. Cluster 2 (station 5) corresponds on the MPS region. Cluster 3 (station 3, station 6, station 7, station 9, station 10, station 12 and station 14) corresponds on the SHPS region.

This result suggests that, for a shorter period of air quality assessment, the number of monitoring station can be reduced to only one station per each cluster of region. Three monitoring stations which are representing three significantly clustered regions are adequate to construct the whole monitoring network.

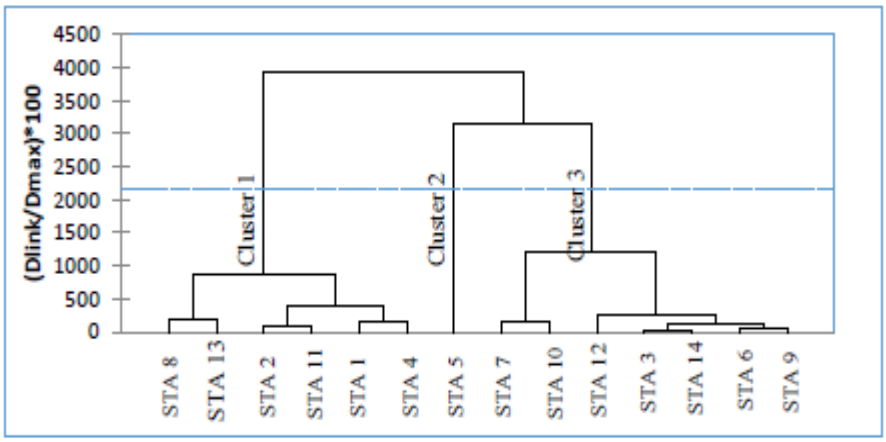


Figure 2. Dendrogram showing different clusters of sampling stations located across Peninsular Malaysia based on API

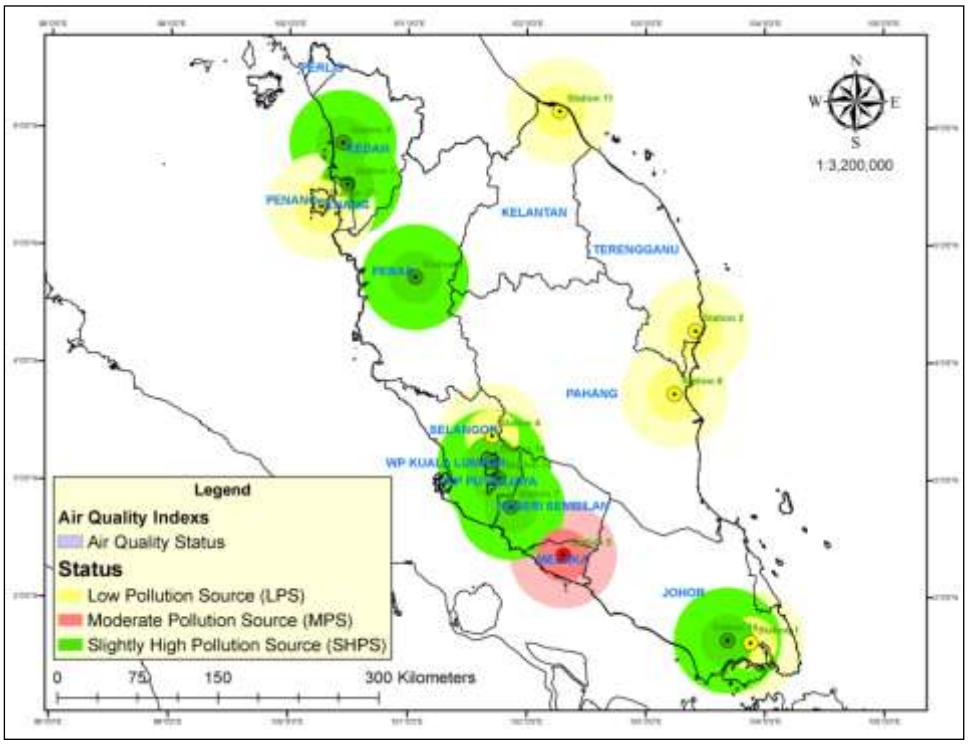


Figure 3. Classification of regions due to air quality by HACA in Peninsular Malaysia

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. The finding from this analysis shows that the accuracy of spatial variation by means of standard mode, forward stepwise mode, and backward stepwise mode were 95.83% (14-variables), 94.05% (8-variables), and 94.05% (9-variables), respectively such in 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, while in the backward stepwise included NO_x as the additional variable for having a high spatial variation. Figure 4 shows the box and whisker plots of three significant regions. Nine selected air quality variables that showed high spatial variations in backward stepwise mode of DA were then applied for further discussion.

Table 2 . Classification matrix for spatial variations across the 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

Source Identification of Air Pollutants

PCA was applied for identifying the source of air pollutants in this study. Four VFs were obtained in LPS and MPS region, and five PCs in the SHPS region based on the eigenvalues more than 1.0. The total variance for LPS, MPS, and SHPS region were correspond to 77.20%, 88.71%, and 79.95%, respectively. The finding of VFs, loadings of variables, and variance are illustrated in Table 3.

Low Pollution Source (LPS) Region

In the LPS region, VF1 contributes 41.73% of the total variance and has strong positive loadings on CO, NO_2, non-methane hydrocarbons, NO and NO_x. In VF1, the presence of CO, NO_2, NO and NO_x are related to the fossil fuel combustion from agricultural systems [28], while the presence of non-methane hydrocarbons is related to the fossil fuel combustion from transportation [29]. Additional carbon can be sequestered as the effect of nitrogen deposition caused by agricultural practices [30]. 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 [31]. 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 [32].

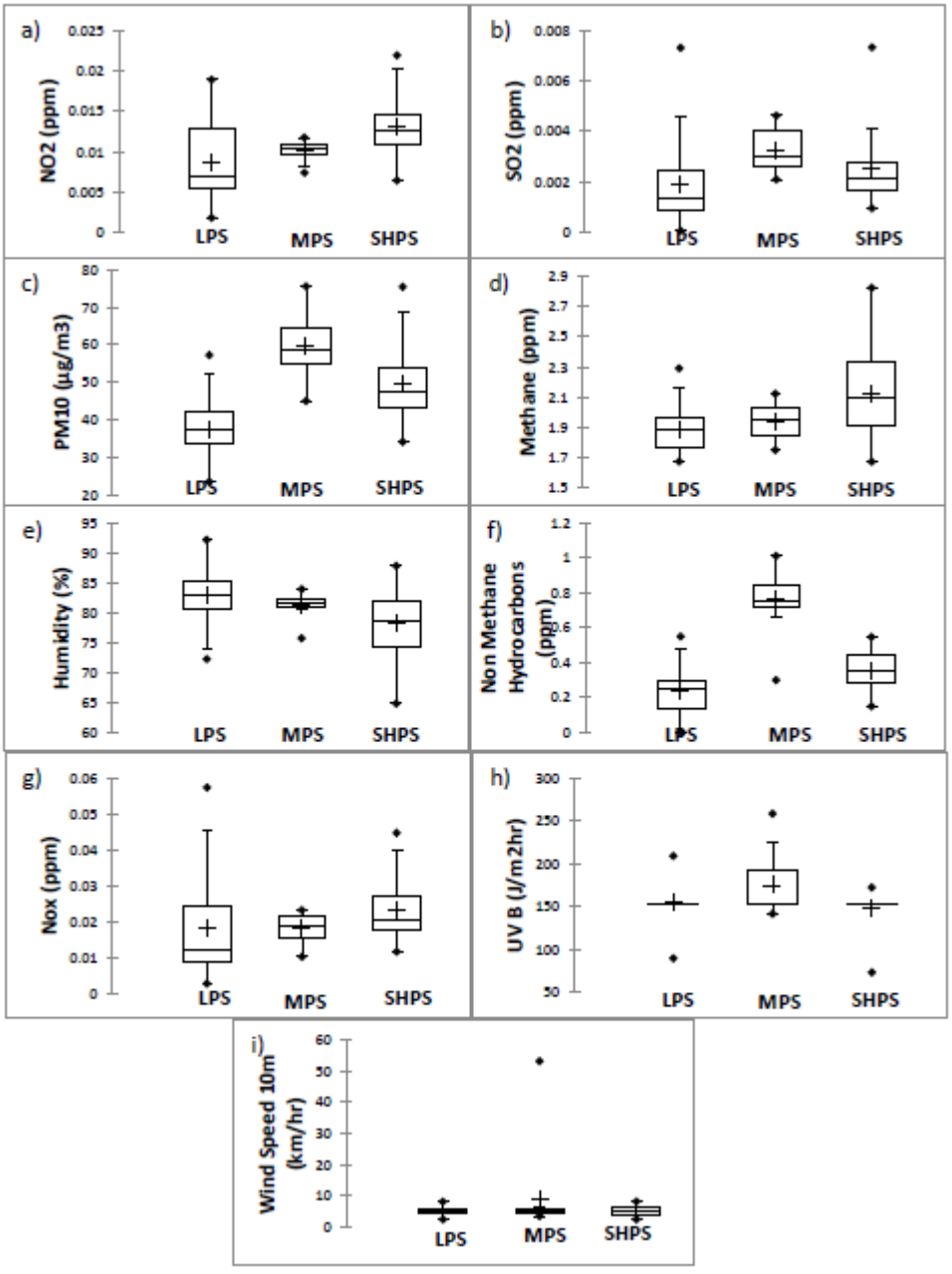


Figure 4. Box and whisker plots of (a) NO_2, (b) SO_2, (c) $PM_{10,}$ (d) Methane, (e) Humidity, (f) Non Methane Hydrocarbons, (g) NO_x, (h) Ultraviolet B, and (i) Wind Speed generated by backward stepwise of DA

Source Identification of Air Pollutants

PCA was applied for identifying the source of air pollutants in this study. Four VFs were obtained in LPS and MPS region, and five PCs in the SHPS region based on the eigenvalues more than 1.0. The total variance for LPS, MPS, and SHPS region were correspond to 77.20%, 88.71%, and 79.95%, respectively. The finding of VFs, loadings of variables, and variance are illustrated in Table 3.

Low Pollution Source (LPS) Region

In the LPS region, VF1 contributes 41.73% of the total variance and has strong positive loadings on CO, NO_2, non-methane hydrocarbons, NO and NO_x. In VF1, the presence of CO, NO_2, NO and NO_x are related to the fossil fuel combustion from agricultural systems [28], while the presence of non-methane hydrocarbons is related to the fossil fuel combustion from transportation [29]. Additional carbon can be sequestered as the effect of nitrogen deposition caused by agricultural practices [30]. 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 [31]. 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 [32].

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, CH_4, 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 [5,24,33]. 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 [29]. 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 [24,34]. 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.

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 [5,24,35]. 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 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 [24,36]. It is tremendously vital to consider the consequence of meteorological states on air pollution, because they directly influence the emission effect of the atmosphere.

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				0.933				0.801				
NO_2	0.939				0.773				0.747				
SO_2					0.906								0.766
PM_{10}					-0.748								
O_3		-0.749						0.827				0.888	
Ambient Temp				0.852							-0.752		
CH_4		0.746			0.913					0.956			
Humidity			-0.756								0.768		
Non Methane Hydrocarbons	0.887					0.941							
NO	0.873				0.857				0.918				
NO_x	0.921				0.881				0.945				
Total Hydrocarbons						0.880				0.911			
UV B			0.769				0.799						
Wind Speed		0.749					0.961				-0.884		
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

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

In this study, the source apportionment of air pollutant parameters (known as independent variable) was used to identify the potential of API (known as dependent variable) values. Four models were developed. To develop the models, 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)).

The finding of the study shows that the values of R^2 and RMSE for the original air quality parameters-API were 0.873 and 3.108, respectively from the goodness of fit statistics. The values of R^2 and RMSE for LPS were 0.865 and 2.187, respectively. The values of R^2 and RMSE for MPS were 0.999 and 1.430, respectively. Meanwhile, the values of R^2 and RMSE for SHPS were 0.868 and 2.195, respectively. The proposed equation with R^2 and RMSE can be seen in equation 6 - 9:

Original air quality parameters (14 variables)

Total API = - 0.15(CO) - 501.09(NO_2) - 210.13(SO_2) +0.70(PM_{10}) + 58.94(O_3) + 0.19(Temp) – 0.24(CH_4) - 0.14(Humidity) - 7.94(NMHC) - 576.91(NO) + 597.93(NO_x) + 1.26(THC) + 6.43e^{-03}(UVB) - 0.88(Wind Speed 10m) + 18.30 [R^2=0.873 and RMSE=3.108] (6)

LPS (9 variables)

Total API = - 524.57(NO_2) + 59.48(SO_2) + 0.82(PM_{10}) + 2.71(CH_4) - 6.99e^{-02}(Humidity) + 6.70(NMHC) + 87.74(NO_x) + 1.47e^{-}02(UVB) - 0.36(Wind Speed 10m) + 8.72 [R^2=0.865, RMSE=2.187] (7)

MPS (9 variables)

Total API = - 8216.13(NO_2) + 1803.87(SO_2) + 1.61(PM_{10}) - 49.44(CH_4) - 2.18(Humidity) – 13.64(NMHC) + 5478.53(NO_x) - 3.95e^{-02}(UVB) - 1.28(Wind Speed 10m) + 229.68 [R^2=0.999, RMSE=1.430] (8)

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 10m) + 16.37 [R^2=0.868, RMSE=2.195] (9)

Based on the equations 6 - 9, the 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 10m 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 10m 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 10m 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 10m, 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 [5,26,27].

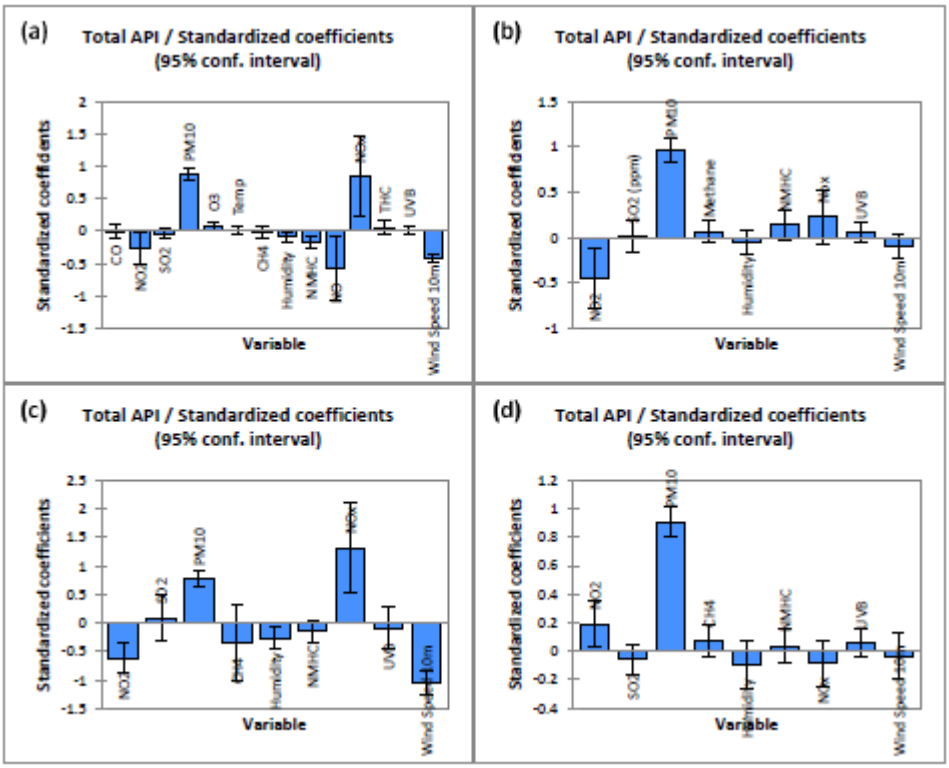
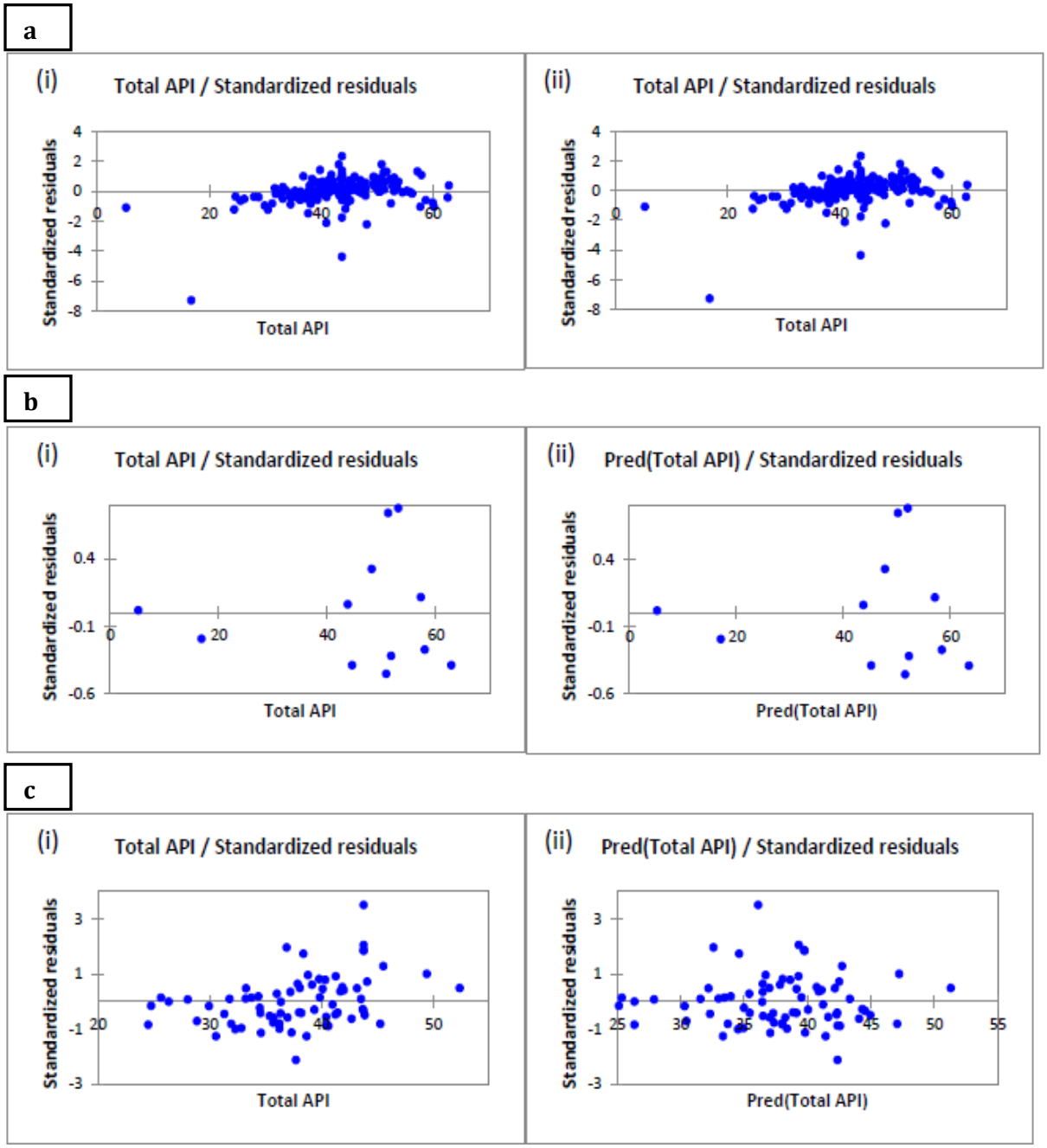


Figure 5. Bar chart of the standardized coefficient for the independent variables of (a) Original air quality parameters, (b) LPS, (c) MPS, and (d) SHPS

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



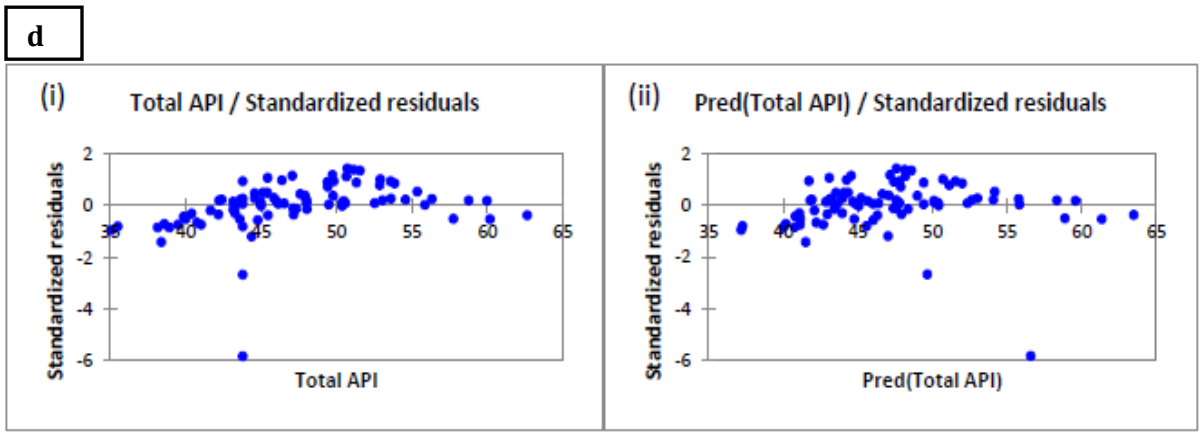


Figure 6. Scatter plot diagram of standardized residuals of (i) actual API, and (ii) predicted API for: (a) original air quality parameter model, (b) LPS model, (c) MPS model, and (d) SHPS model.

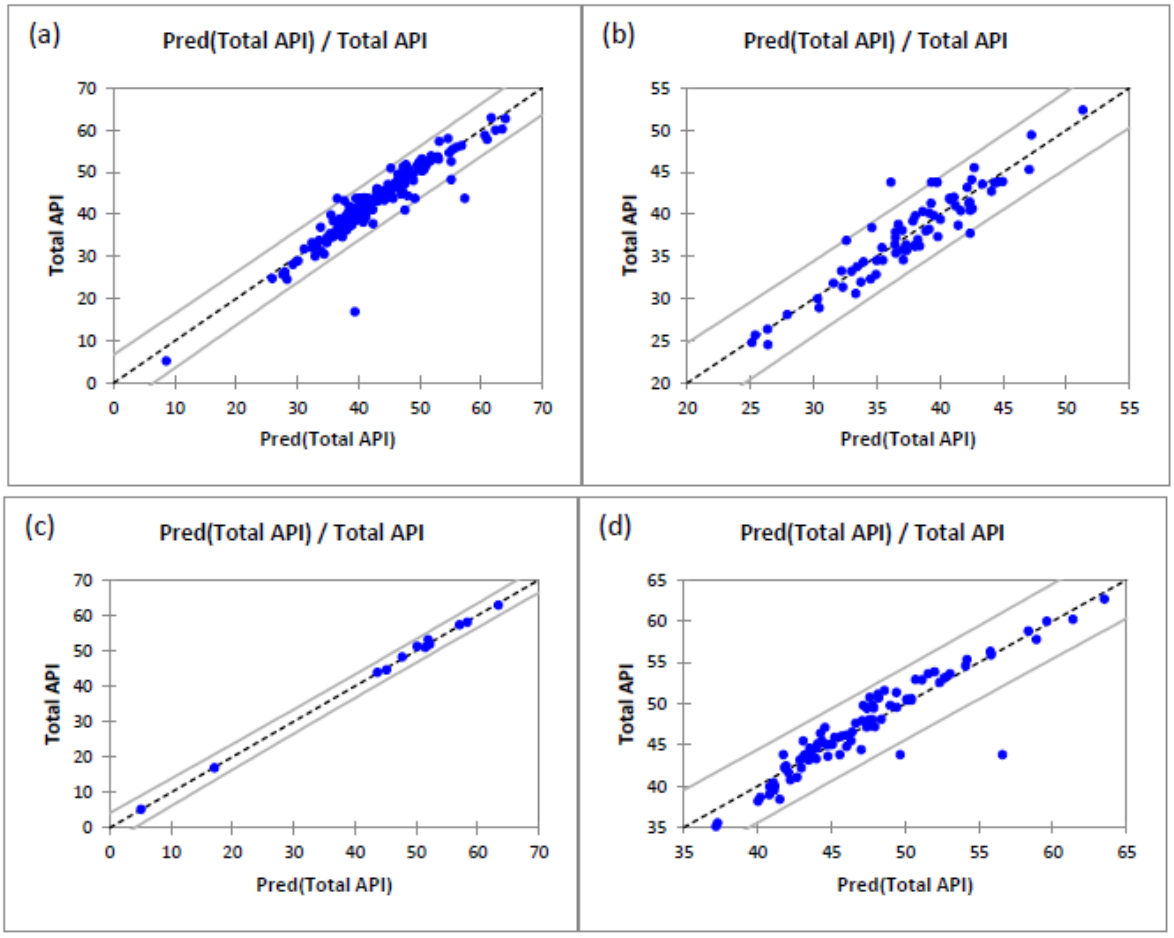


Figure 7. Scatter plot diagram of the API (predicted) versus the actual API of (a) original air quality parameter model, (b) LPS model, (c) MPS model, and (d) SHPS model.

Conclusion

From this study, it can be concluded that the spatial variations of air quality data in Peninsular Malaysia were successfully studied by applying chemometrics techniques. 14 monitoring stations were grouped into three significant diverse cluster regions, known as LPS, MPS, and SHPS by using HACA. 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 a new design of air quality monitoring network instead of taking 14 air quality variables into account. To identify the source of air pollution, PCA was done. Four VFs were found for 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 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. 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 API in 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) and East Coast Environmental Research Institute for the supply of air quality data required for the completion of this study.

References

1. Moustris, K.P., Ziomas, I.C. and Paliatsos, A.G. (2010). 3-day-ahead forecasting of regional pollution index for the pollutants NO2, CO, SO2, and O3 using artificial neural networks in Athens, Greece. *Water, Air & Soil Pollution* 209(1-4): 29-43.
2. 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(1): 83-88.
3. 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. *Building and Environment* 46: 577-583.
4. Simeonov, V., Einax, J.W., Stanimirova, I. and Kraft, J. (2002). Envirometric modeling and interpretation of river water monitoring data. *Analytical and Bioanalytical Chemistry* 374: 898–905.
5. 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 chemometrics techniques: a case study in Malaysia. *Environmental Sciences: Processes & Impact* 15(9): 1717-1728.
6. Kannel, P. R., Lee, S., Kanel, S. R. and Khan, S. P. (2007). Chemometrics application in classification and assessment of monitoring locations of an urban river system. *Analytical Chimica Acta* 582: 390-399.
7. Satheeshkumar, P. and Khan, A.B. (2011). Identification of mangrove water quality by multivariate statistical analysis methods in Pondicherry coast, India. *Environment Monitoring Assessment* 184(6): 3761-3774.

8. 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. *Analytica Chimica Acta* 35: 3581–3592.
9. Giri, D., Murthy, V.K., Adhikary, P.R. and Khanal, S.N. (2007). Cluster analysis applied to atmospheric PM_{10} concentration data for determination of sources and spatial patterns in ambient air-quality of Kathmandu Valley. *Current Science.* 93(5): 684-688.
10. Kaufman, L and Rousseeuw, P.J. (I990). *Finding Groups in Data.* Wiley Interscience, New York.
11. Ibarra-Berastegi, G., Sáenz, I., Ezcurra, A., Ganzedo, U., Argendoña, J.D., Errasti, I.,Farnandez – Ferrero, A. and Polanco – Martínez, J. (2009). Assessing spatial variability of SO_2 field as detected by an air quality network using self-organizing maps, cluster, and principal component analysis. *Atmospheric Environment.* 43: 3829–3836.
12. 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 chemometrics techniques. *Environment Monitoring Assessment* 173: 625-641.
13. Saithanu, K. & Mekpatyup, J. (2014). Air quality assessment in the urban areas with multivariate statistical analysis at the east of Thailand. *International Journal of Pure and Applied Mathematics.* 9(2): 169-177.
14. Shrestha, S. and Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software* 22: 464–475.
15. Manjunath, B.G., Frick, M. and Reiss, R.D. (2012). Some Notes on Extremal Discriminant Analysis. *Journal of Multivariate Analysis.* 103: 107–115.
16. Johnson, R.A. and Wichern, D.W. (1992). *Applied multivariate statistical analysis*. 3rd ed. Prentice-Hall Int.: New Jersey.
17. Hopke, P.K. (1985). *Receptor modelling in environmental chemistry*. New York: Wiley.
18. 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 Research* 38: 3980–3992.
19. Yu, T.Y. and Chang, L.F.W (2000). Selection of the scenarios of ozone pollution at southern Taiwan area utilizing principal component analysis. *Atmospheric Environment* 34: 4499-4509.
20. 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. *The Science of the Total Environment* 313, 77–89.
21. 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. & Yamin, M. (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: 2063.
22. 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. *International Journal of Applied Science and Engineering* 7(2): 127-132.
23. 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. *International Journal of Applied Science and Technology* 1(4): 42-49.
24. 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. and Umar, R. (2015). Identification source of variation on regional impact of air quality pattern using chemometrics. *Aerosol and Air Quality Research*:
25. 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. *Ecological Modelling* 221: 1119-1130.
26. 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. *Atmospheric Environment* 60: 172-181.
27. 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. *Journal Environmental Protection* 4: 1-10.

28. 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. *Ecological Economics* 55: 235 – 250.
29. Koppmann, R. (2007). *Volatile organic compounds in the atmosphere*. Singapore: Blackwell Publishing Ltd.
30. 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 Biology* 12: 1151–1173.
31. 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? *Atmospheric Environment* 39: 2513–2517.
32. 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. *Atmospheric Chemistry and Physics* 9: 2695-2714.
33. 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. (2014). 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, pp 307.
34. Giorgi, F. and Meleux, F. (2007). Modelling the regional effects of climate change on air quality. *C. R. Geoscience* 339: 721–733.
35. Romieu, I. and Hernandez, M. (1999). *Air pollution and health in developing countries: review of epidemiological evidence*. In: Mc Granahan, G., Murray, F. (Eds.), Health and Air Pollution in Rapidly Developing Countries. Stockholm Environment Institute, Sweden, pp 43 – 66.
36. Elminir, H. (2005). Dependence of urban air pollutants on meteorology. *Science of Total Environment* 350: 225–237.