Jurnal
Teknologi

Full Paper

ASSESSMENT OF THE SPATIAL VARIATION AND SOURCE APPORTIONMENT OF AIR POLLUTION BASED ON CHEMOMETRIC TECHNIQUES: A CASE STUDY IN THE PENINSULAR MALAYSIA

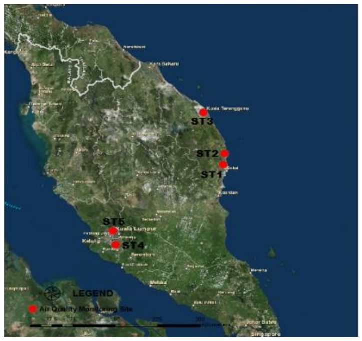
Hamza Ahmad Isiyaka*, Hafizan Juahir, Mohd Ekhwan Toriman, Azman Azid, Barzani M. Gasim, Mohd Khairul Amri Kamarudin

East Coast Environmental Research Institute (ESERI), Universiti Sultan ZainalAbidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia

Article history
Received
17 March 2015
Received in revised form
9 April 2015
Accepted
1 October 2015

*Corresponding author
isiyakas@gmail.com

Graphical abstract



Abstract

This study aims to investigate the spatial variation in the source of air pollution, identify the percentage contribution of each pollutant and apportion the mass contribution of each source category using chemometric techniques. Hierarchical agglomerative cluster analysis (HACA) successfully grouped the five air monitoring sites into three groups (cluster 1, 2 and 3). Principal component analysis (PCA) was used to spot out the sources of air pollution which are attributed to anthropogenic activities. Multiple linear regression (MLR) was used to develop an equation model that explains the contribution of pollutants in each cluster. However, it was observed that particulate matter (PM_{10}) and Ozone (O_3) are the most significant pollutants influencing the value of air pollutant index (API). Meanwhile, the source apportionment indicates that cluster 1 is influenced by gas and non-gas pollutants to a degree of 84%, weather condition 15% and 1% by gas and secondary pollutants. Cluster 2 is affected by gas and secondary pollutants to a tune of 87% and 13% by weather condition while cluster 3 is apportioned with 98% secondary gas and non-gas pollutants and 2% weather condition. This study reveals the usefulness of chemometric technique in modeling and reducing the cost and time of monitoring redundant stations and parameters.

Keywords: Chemometric technique, source apportionment, principal component analysis, multiple linear regressions, air pollution index

Abstrak

Kajian ini bertujuan untuk mengkaji variasi ruang dalam sumber pencemaran udara, mengenal pastisumbangan peratusan bagi setiap pencemar, dan mengagihkan sumbangan jisim bagi setiap kategori sumber dengan menggunakan teknik kemometrik. Analisis hierarki agglomerative kelompok (AHAK) telah Berjaya membahagikan lima tapak pemantauan udara kepada tiga kumpulan (kelompok 1, 2 dan 3). Analisis komponen utama (AKU) telah digunakan untuk mengesan sumber pencemaran udara yang disebabkan oleh aktiviti antropogenik. Regresi linear pelbagai (RLP) telah digunakan untuk membangunkan model persamaan yang menerangkan sumbangan pencemar dalam setiap kelompok. Walau bagaimanapun, ia telah menunjukkan bahawa zarah terampai (PM_{10}) dan ozon (O_3) adalah pencemar yang paling ketara mempengaruhi nilai Indeks Pencemaran Udara (IPU). Sementara itu, pembahagian sumber menunjukkan bahawa kelompok 1 dipengaruhi oleh pencemar gas danbukan gas ke tahap 84%, keadaancuaca 15%, dan 1% oleh gas dan bahan cemar kedua. Kelompok 2 dipengaruhioleh gas dan bahan cemar kedua iaitu sebanyak 87%, dan 13% disebabkan oleh keadaan cuaca, manakala kelompok 3 telah dibahagikan kepada 98% pencemar dari gas dan bukan gas kedua di samping keadaan cuaca sebanyak 2%. Kajian ini mendedahkan betapa pentingnya kegunaan teknik kemometrik dalam pemodelan bagi mengurangkan kos dan masa dalam memantau stesen pemantaun yang bertindih disamping parameter-parameter lain.

Kata kunci: Teknik-teknik kemometrik, sumber pembahagian, analisis komponen utama, Regresi linear pelbagai, indeks pencemaran udara

1.0 INTRODUCTION

Environmental protection and air quality management constitutes an important issue in public policy in this current dispensation. The 21st century has seen a rapidly change in climate [1] and an increase in the level of pollutants emitted into the air as countries become industrialized [2]. Air pollutants found in urban areas represent a mixture of primary particles emitted from various sources and secondary particles from aerosols formed by chemical reactions [3], since the general characteristics of the troposphere has the ability to emit, transform, disperse as well as deposit pollutants [4].

Atmospheric air pollutants have the ability to penetrate into the gas exchange region of the lung when inhaled, which causes lung cancer, asthma, cardiovascular and respiratory diseases [5, 6] as well as impair visibility. It also affect the amount of oxygen fed to the fetus through the mother [7], thereby retarding the anthropometric developmentby reducing its length and head circumference, cause global warming and affect the general ecosystem. According to [8], an estimate of 2.4 million people dies each year from causes directly attributed to air pollution.

However, atmospheric air pollutants are usually transported from point source to non-point source depending on the meteorological characteristics of the area (temperature and wind speed) [9]. An estimate of about 82% of the total pollutants in Malaysia are emitted from mobile sources especially motor vehicles; the stationary source usually come from static points such as industrial activities, power stations and construction sites [10, 11]; while the transboundary pollutants are emitted from neighboring Indonesia with the worst episode in 1997 due to adverse bush burning [12, 13].

An increase in the economic level and standard of living in Malaysia has led to an increase in the ownership of vehicles, making the number of newly registered motor vehicles to increase by 4.42% from 934,367 vehicles in 2004 to 1,160,082 vehicles in the year 2010 [14, 15]. A report by Department of Statistics Malaysia in 2011 shows that Malaysia population increased from 23.3 million in 2000 to 28.3 million in 2010. Malaysia experiences tropical climate with a uniform temperature, high humidity, strong wind as well as copious rainfall throughout the year. These climatic characteristics are usually influenced by the northeast and southwest monsoon wind [16, 17]. High temperature helps in photochemical oxidation that transforms primary pollutants to secondary pollutants (O_3 and PM_{10}) [18, 19]. Strong wind helps to transport O_3 precursors and PM_{10} from point source to non-point source [18, 20, 21].

Most urban cities within Peninsular Malaysia faces serious air pollution episodes [22] and about 5.4 metric tons of CO_2 is produced in Malaysia as at 2002 which exceeds the benchmark of 3.9 metric ton per capita provided by the world resource institute [23]. Although, based on air pollutant index (API) classification, the air quality status in Malaysia fall within good and moderate but still show a slight fluctuation in trend from 2008 to 2011 [24]. In 2008, the status of good air quality fall around 59%, 55.6% in 2009, 63% in 2010 and 55% in 2011 [24].

API is a non-dimensional number calculated according to the urban daily average concentration of pollutants [13, 24]. API is used to indicate and classify the ambient air quality in Malaysia base on the possible health implications to the public [25]. It is calculated based on the sub-index using five air pollution parameters such as; PM_{10}, SO_2, NO_2, CO_2 and O_3. The highest sub-index value of the individual pollutants is used as the API value for a specific time period [10, 13, 25, 26]. Malaysia has a standard threshold for measuring the level of its air quality, which is based on the Recommended Malaysia Ambient Air Quality Guideline (RMAQG) issued by the Department of Environment DOE since 1989 as good, moderate, unhealthy, very unhealthy and hazardous. This conformsto international standard provided by United State Environmental Protection Agency [25]. Meteorological parameters (ambient temperature and wind speed) were used in this study because of their ability to influence the source and percentage contribution of each pollutant to the API [27].

Atmospheric air quality monitoring involves observation of large complex data sets from stations which require the integration of modern and robust statistical techniques for simplification, avoid misinterpretation and to show spatial variation [28, 29, 30, 31]. However, chemometric technique (HACA, PCA, FA and MLR) was used based on the statistical principles that involve a simultaneous observation and analysis of more than one variable by simplifying the process within a convenient size [26]. Chemometric techniques reduce the dimensionality of the data as well as extract meaningful information from it [32, 33, 34, 35, 36].

Using the XLSTAT 2014 add-in software, the spatial variation of air quality was achieved using HACA. PCA after varimax rotation was used to identify the major possible source of pollution with a strong factor loading (>0.75).Principal component scores (PCS) together with MLR was used to obtain source apportionment in each cluster developed by HACA [35, 36, 37, 38, 39]. The percentage contribution of each parameter to the level of pollution was also predicted using MLR [40, 41]. Many researchers have adopted the use of XLSTAT software (11, 13, 23, 29, 31, 35, 36, 39) in modeling air pollution because of its flexibility, multidimensionality and ability to synthesis complex data sets. The objectives of this study are: to identify the spatial variation in air monitoring sites; to identify the major possible source of the pollutants base on their spatial variability; to predict the percentage contribution of each pollutants and apportion the mass contribution of each source category.

2.0 EXPERIMENTAL

2.1 Sampling Site

The study area comprises of two main strategic locations (Terengganu and Klang Valley) within the Peninsular Malaysia. Five continuous air quality monitoring stations were selected. The Kemaman station (ST1), Paka-Kertih (ST2), Kuala Terengganu (ST3), Putrajaya (ST4) and Kuala Lumpur (ST5) [42, 43]. The location of the monitoring site base on latitude and longitude are shown in Table 1 and Figure 1.

Table 1 Location and coordinates of monitoring station

Station ID	Location	coordinates	
		Latitude	Longitude
ST 1	Kemaman	N04° 16.260	E103° 25.826
ST2	Paka-Kertih	N04° 35.880	E103° 26.096
ST3	Kuala Terengganu	N05° 18.455	E103° 07.213
ST4	Putra Jaya	N02° 55.915	E101° 40.909
ST5	Kuala Lumpur	N03° 06.376	E 101°43.072

Source: Department of Environment (DOE) Malaysia

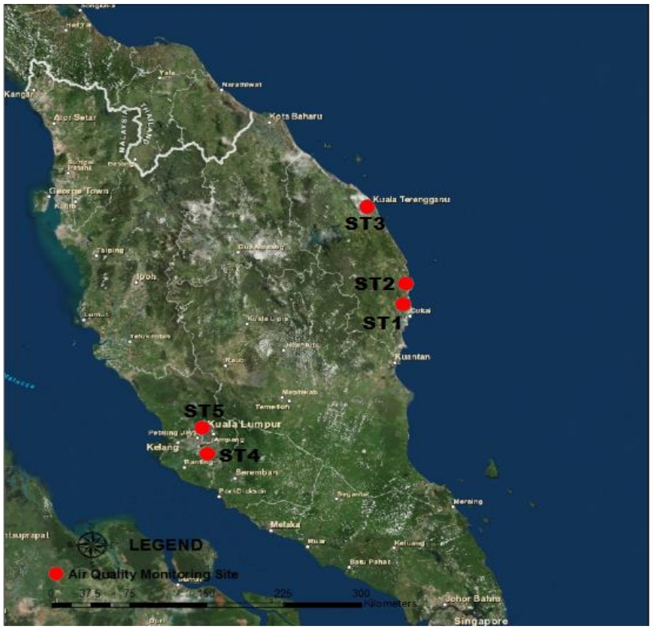


Figure 1 Map of the study area showing air quality monitoring stations

2.2 Data Collection and Processing

Five air pollutants and two meteorological parameters including carbon monoxide (CO), ozone (O_3), particulate matter (PM_{10}), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ambient temperature and wind speed were sourced from the Department of Environment Malaysia covering 2007- 2011 (5years). The average recording time varies from 1-24 hours which represents the period of time over which measurement are monitored and reported in order to assess the possible health impacts of specific air pollutants [25, 44]. The data were provided in the form of hourly readings by DOE Malaysia and were converted to daily average observations which gave a total of 64036 data sets (9148 observations × 7 parameters). Furthermore, to facilitate this analysis, the nearest neighbor method was applied using XLSTAT 2014 add-in software to estimate missing values [13, 45]. The missing data recorded is 4% of the original data sets. The nearest neighbor method can provide a simple scheme, where the endpoints of the gaps are used as estimates for all missing values [45]. This equation can be shown below:

$$y = y_1 \text{ if } x \leq x_1 + [(x_2 - x_1)/2] \text{ or}$$
$$y = y_1 \text{ if } x \geq x_1 + [(x_2 - x_1)/2] \quad (1)$$

Where y represents the interpolate, x is the time point of the interpolate, y_1 and x_1 are the coordinates of the starting point of the gap and y_2 and x_2 are the end points of the gaps.

2.3 Spatial Classification of Sampling Site Using HACA

HACA is used to spatially classify the monitoring sites into groups that are homogenous in their characteristics and differ from the observation in other groups [46, 47, 48, 49]. This spatial classification can be shown in a dendrogram that measures the degree of risk homogeneity through Ward's method and Euclidian distance measurement [50]. Euclidian distance is based on a single linkage which denotes the quotient between the linkage distance divided by the maximal distance [(Dlink/Dmax)], by multiplying the quotient by 100 in order to standardize the linkage distance represented by the y-axis [51, 52].

2.4 Principal Component Analysis for Spatial Source Identification (PCA)

PCA is a powerful statistical tool used to establish a small number of components that can explain the maximum variance possible in a data set [46, 47, 49, 53]. The equation is expressed as:

$$Z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + a_{i3}x_{3j} + \ldots\ldots\ldots\ldots + a_{im}x_{mj} \qquad (2)$$

Where z is the component score, *a* is the component loading, x is the measured value of variables, *i* is the component number, *j* is the sample number and *m* is the total number of variables.

Although, the principal components (PCs) generated by PCA are sometimes not readily available for interpretation, therefore, it is advisable to rotate it by varimax rotation with eigenvalues greater than one [29]. The varimax rotation is considered significant in order to obtain new groups of variables called varimax factors (VFs) (54, 55). This will help identify the different possible sources of pollution (10, 11). In addition, the VFs coefficient with a correlation from 0.75 are considered as strong significant factor loading, those that range from 0.50 - 0.74 are moderate, while 0.30 - 0.49 are classified as weak significant factor loading [56].

$$Z_{ij} = af_1 x_{1i} + a_{f2} x_{2i} + \ldots\ldots\ldots + a_{fm}f_{mi} + e_{fi} \qquad (3)$$

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 variation, *i* is the sample number, *j* is the variable number and *m* is the total number of factors.

2.5 Multiple Linear Regression Model (MLR)

MLR is a statistical technique that is used to predict the variability that exists between the dependent and independent variable [36, 40, 41, 57]. It is used to form explicit equations that are less complex [58]. Thus, the regress model can be represented as [19].

$$Y_i = \beta_o + \beta_1 x_{li} + \ldots\ldots\ldots\ldots\ldots + \beta_k x_{ki} + \varepsilon_i Eq \qquad (4)$$

where $i = 1\ldots\ldots.n$, β_0, β_1 and β_k are regression coefficient, x_1 and x_k are independent variables and $\varepsilon$is error associated with the regression.

Using this method, the percentage contribution of each parameter and the source apportioned were predicted base on the coefficient of determination "R^2", adjusted coefficient of determination "Adjusted R^2" and Root mean square error "RMSE" [40, 41].To achieve this, each parameter for percentage contribution was independently introduced to the linear model with API as the dependent variables [13]. PCs after varimax rotation for each factor scores were also introduced as independent variable and API as dependent variable for source category apportionment using the leave-one-out cross-validation method [37, 59]. The descriptive statistics of the entire data sets comprising of minimum value, maximum, mean and standard deviation were analyzed using the XL-STAT 2014 add-in software in Table 2.

3.0 RESULT AND DISCUSSION

3.1 Spatial Classification of Air Quality Monitoring Site

HACA was used to identify the spatial similarities and differences in the characteristics of air quality monitoring sites. Those with strong level of spatial homogeneity were grouped into one cluster. This process resulted in the formation of three clusters as presented in Figure 2. One station in each cluster can spatially represent the air quality monitoring process within the sampling sites. Cluster 1 is classified as a good healthy site (GHS) because it is associated with less pollutants base on the characteristics of activities in the air monitoring stations [43]. These areas comprise of Kemaman (ST1), Kertih (ST2) and Kuala Terengganu (ST3) located in the east coast Peninsular Malaysia. The activities in these areas range from residential, petrochemical industry, airport, seaports, city center with traffic, commercial and industrial practices [60, 61]. Cluster 2 is an area with moderate pollution called Putrajaya and can be classify as a moderate healthy site (MHS). Putrajaya (ST4) is the administrative nerve center of Malaysia and the third most populated state. It is associated with different commercial activities, residential areas and traffic [42, 60]. The third cluster is the most populated region in Malaysia, with vast industrial and commercial activities, construction sites, biomass burning, re-suspension and evaporation of soil and road dust and heavy traffic jam around the clock [62, 63]. This area is considered as high pollution site (HPS). The findings from HACA revealed that there are redundant monitoring stations in Terengganu which needs to be re-installed in a more strategic area with busy activities and likelihood of pollutant emission. This will indeed reduce the cost of equipment, staffing and time management. The equipment's can be moved to Kuala Lumpur which is considered as area of HPS. This is because one station in Terengganu can give a spatial representation of the monitoring processes.

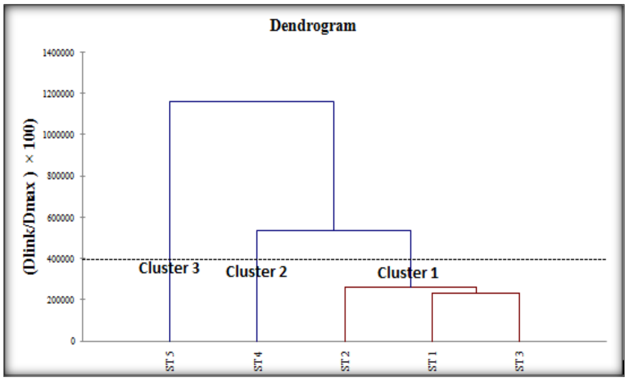


Figure 2 Dendrogram showing spatial classification of air monitoring sites

3.2 Spatial Identification of Major Pollutio Source

PCA was performed on the data set base on the unsupervisedspatialclassificationgeneratedby HACA in order to compare the compositional pattern between the examined parameters and to identify the major possible source of pollution in the study area. Three PCs were obtained for cluster 1 and two PCs for cluster 2 and 3 witheigenvaluesgreater than one (1), which account for more than 62%, 56% and 58% respectively of the total variance in the data sets (Figure 3). Three varifactors (VFs) were obtained for cluster 1, and two (VFs) for both cluster 2 and 3 through the FA performed on the PCs. Only the factor loading greater than 0.75 are considered for interpretation. Table 3 comprises of factor loadings, eigenvalues, total variance and cumulative percentages. The plot diagram for PCA loading after varimax rotation is shown in Figure 3.

Table 2 (a) Descriptive statistics of daily average air pollutants (b) meteorological parameters and API 2007- 2011

(a)

Variables	Statistics	Stations					Avg. time	RMAQG
		ST1	ST2	ST3	ST4	ST5		
CO, ppm	Min	0.000	0.000	0.000	0.000	0.000	1-h	30
	Max	3.180	0.810	2.740	2.830	4.750		
	Mean	0.452	0.049	0.923	1.045	1.624		
	Std. dev	0.250	0.117	0.426	0.423	0.722		
O_3, ppm	Min	0.000	0.000	0.000	0.000	0.000	1-h	100
	Max	3.180	0.050	0.140	0.120	0.170		
	Mean	0.452	0.003	0.027	0.047	0.055		
	Std. dev	0.250	0.008	0.012	0.022	0.031		
PM_{10}, µg/cu.m	Min	0.000	0.000	0.000	13.583	0.000	24-h	150
	Max	264.000	470.000	357.000	299.000	282.000		
	Mean	55.922	57.950	84.297	59.903	71.137		
	Std. dev	26.155	33.051	39.352	26.518	27.875		
SO_2, ppm	Min	0.000	0.000	0.000	0.000	0.000	1-h	130
	Max	0.060	0.010	0.010	0.060	0.060		
	Mean	0.002	0.001	0.000	0.004	0.003		
	Std. dev	0.005	0.002	0.001	0.006	0.006		
NO_2, ppm	Min	0.000	0.000	0.000	0.000	0.000	1-h	170
	Max	0.030	0.060	0.030	0.070	0.110		
	Mean	0.007	0.007	0.012	0.025	0.036		
	Std. dev	0.006	0.009	0.006	0.011	0.014		

(b)

Variables	Statistics	Stations				
		ST1	ST2	ST3	ST4	ST5
Ambient Temp °C	Min	22.208	21.463	21.463	22.371	23.913
	Max	30.763	31.267	31.267	29.679	29.679
	Mean	27.623	27.333	26.333	26.927	27.218
	Std. dev	1.597	1.701	1.848	1.252	1.124
Wind Speed 10m, KM/h	Min	0.313	1.000	1.508	1.508	1.508

	Max	27.213	210.200	30.700	13.983	7.546
	Mean	11.139	14.240	10.532	4.383	4.274
	Std. dev	3.571	14.143	3.530	0.986	0.891
API	Min	0.000	0.000	19.458	14.208	20.125
	Max	95.000	71.000	81.000	119.000	169.000
	Mean	45.126	37.926	50.053	53.513	64.298
	Std. dev	13.186	10.016	10.206	16.986	22.820

Table 3 Factor loadings after varimax rotation by clusters

Variables	Cluster 1			Cluster 2		Cluster 3	
	VF1	VF2	VF3	VF1	VF2	VF1	VF2
CO	0.776	0.27	-0.318	0.826	-0.033	0.789	-0.028
O_3	0.361	0.691	-0.111	0.756	0.057	0.763	0.013
PM_{10}	0.784	-0.089	-0.073	0.663	0.061	0.768	0.034
SO_2	-0.102	0.715	0.293	0.219	0.126	0.262	0.01
NO_2	0.767	0.006	0.211	0.81	-0.054	0.838	-0.046
ABT TEMP	-0.111	0.181	0.781	0.421	0.763	0.033	0.882
WIND SPEED	0.059	-0.432	0.542	-0.272	0.864	-0.054	0.879
Eigenvalue	2.131	1.211	1.06	2.646	1.343	2.571	1.554
Variability (%)	30.444	17.297	15.139	37.795	19.183	36.734	22.206
Cumulative %	30.444	47.741	62.88	37.795	56.978	36.734	58.94

a) Cluster 1

The first varifactor (VF1) have a strong positive loadings for CO (0.776), PM_{10} (0.784) and NO_2 (0.767) that explains 30.4% of the total variance in cluster 1. The primary origin of CO in this region could be attributed to the incomplete combustion of fuel in automobiles (motor vehicles, motorcycles, engine boats) [64, 65]. Industrial activities and construction site [66], gases and particles from transport exhaust emission, re-suspension of soil dust and transboundary pollution [67, 68, 69]are the major source of PM_{10} in this area. Emission from industrial activities in Kemaman and Paka-Kertih are the major source of NO_2 [70]. Furthermore, NO_2 is produced when fossils fuel are burnt.

The Third varifactor (VF3) has a strong positive loading for ambient temperature (0.781) explaining about 15.1% of the total variance. During the southwest monsoon season, Terengganu experience high temperature [74], this scenario provides favorable condition for chemical reactions [19] and allows sulfate formation to increase due to faster SO_2 oxidation, which acts as an important precursor for the formation of PM_{10} [59]. NO_2 is also produced when nitrogen in fuel is burnt and when at a very high temperature nitrogen in the air reacts with oxygen [75].

b) Cluster 2

In varimax factor one (VF1), CO (0.826), O_3 (0.756) and NO_2 (0.81) exhibit a strong positive loading which account for 37.8% of the total variance. Large proportion of CO emission can be attributed to fossil fuel combustion in power generation, incomplete combustion of fuel in automobiles, industrial emission, engine boats especially when the engines are not turned properly which result to incomplete combustion [13, 60, 64, 65, 71]. O_3 is a secondary pollutant produced by photochemical oxidation and forms the main component of photochemical smog [76]. Its concentration is produced by O_3 precursors (NOx CO and VOCs) from industrial activities and motor vehicle emission from traffics in Putrajaya the third most populated region and administrative center of Malaysia [42, 77, 78]. NO_2 is largely due to industrial activities as well as emission from traffic. In Malaysia, an analysis of NO_2 emission proves that about 69% of this pollutant is from power stations and industrial activities, 28% from motor vehicles and the remaining 3% makes up other sources [44].

Ambient temperature (0.763) and wind speed (0.864) hold a strong positive loading that explain about 19.2% of the total variance. High temperature due to the tropical climatic characteristics of Malaysia provides favorable condition for nitrogen oxides (NOx) and other VOCs to react, and CO slowly helps to oxidize nitrogen monoxide (NO) to nitrogen dioxide (NO_2) which indirectly accelerates O_3 formation [19, 42, 79, 80]. Wind speed help to transport both primary and secondary pollutants from point source to non-point source [18, 20].

c) Cluster 3

The two varimax factor (VF1, VF2) in this region explains 36.7% and 22.2% of the total variance with a strong positive loadings for CO (0.789), O_3 (0.763), PM_{10} (0.768), NO_2 (0.838), ambient temperature (0.882) and wind speed (0.879). The source of these air pollutants can be strongly attributed to vast industrial activities, high traffic congestion around the clock, biomass burning, heavy construction work for city development as well as re-suspension of soil and road dust in the city of Kuala Lumpur [22, 62, 63].

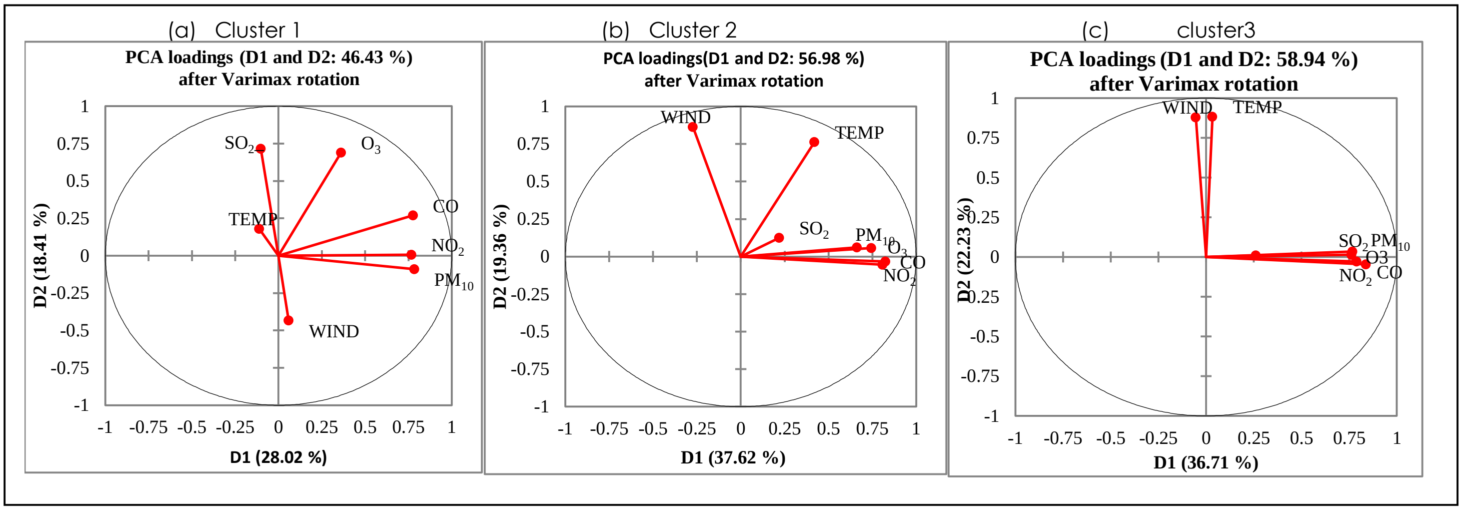


Figure 3 Plot diagram for PCA loading after varimax rotation

The traffic volume of motor vehicles in Kuala Lumpur is about 2,972 vehicles within 16 h of traffic and rises to 28,565 vehicles during the peak hour [15]. Most commercial and industrial activities in Malaysia are found in Kuala Lumpur. This contributes to the emission of O_3 and NO_2 [42]. High temperature provides favorable condition for photochemical oxidation of both O_3 and PM_{10} precursors. Strong positive loading for wind speed (0,879) [21] enable O_3 precursors (CO, NOx, and hydrocarbon) to be transported by multi-scale circulations which increases the concentration of O_3 during the transport period, while PM_{10} are transported from source point to non-source point [18, 20]. Power generation and industrial combustion are examples of point source PM_{10} emission (67]. Bush burning in Indonesia is usually transported into Malaysia (non-point source) by strong wind which pollutes the air [60].

3.3 Multiple Linear Regression for Modeling Air Pollution (MLR)

MLR was applied in order to develop an explicit equation model with low level of complexity, identify the percentage contribution of each pollutant and apportion the spatial mass contribution of each source category.

3.3.1 Propose Equation for the Model

The proposed equations as well as the coefficient of determination (R^2), adjusted coefficient of determination (Adjusted R^2) and the RMSE for each cluster are shown in equations (4(i), (ii), (iii)):

a) Cluster 1

API = 0.78 (CO) + 240.38 (O_3) + 0.177 (PM_{10}) + 50.15 (SO_2) + 25.72 (NO_2) + 0.21 (Ambient Temperature) + 2.25 (Wind speed) + 20.49 (R^2 = 0.51, Adjusted R^2 = 0.51, RMSE = 8.64) (4i)

b) Cluster 2

API = -3.14 (CO) +697.75 (O_3) + 0.10 (PM_{10}) + 6.20 (SO_2) + 10.78 (NO_2) - 1.21 (Ambient Temperature) + 0.42 (Wind Speed) + 47.93 (R^2 = 0.82, Adjusted R^2 = 0.82 RMSE = 7.21) (4ii)

c) Cluster 3

API= -1.70 (CO) + 664.79 (O_3) + 7.97 (PM_{10}) + 74.64 (SO_2) - 50.88 (NO_2) - 3.97 (Ambient Temperature) - 0.40 (wind Speed) + 28.91(R^2 = 0.84 Adjusted R^2 = 0.84 RMSE = 9). (4iii)

From the above equations developed by MLR (4(i), (ii) and (iii)), cluster 3 appears to have the highest R^2 =

0.84, which means that the seven monitored parameters contribute 84% with O_3, PM_{10} and SO_2 having a strong positive influence on the level of air pollution index, while CO, ambient temperature and wind speed negatively influences the API value. Cluster 2 is the second highest contributor to the API value with about 82%. O_3, PM_{10}, SO_2, NO_2 and wind speed positively influence the API, while CO and ambient temperature exhibit a negative influence. The least contributor to the API value is cluster 1 which account for 51%. Although, all the observed parameters in cluster 1 have a positive value of API.

The strong positive influence of O_3(+ 664.79),PM_{10}(+ 7.97) SO_2 (+ 74.64) in cluster 3 at R^2 = 0.84 indicates that only three parameters have strong significant influence. This is as a result of the mainstream economic activities in Kula Lumpur with extensive physical development of infrastructure, emission from traffic congestion, industrialization, and rapid urbanization.

3.3.2 Percentage Contribution of Individual Pollutants in Each Cluster

The standardized coefficient and the leave-one-out cross-validation method were used to model the most significant parameter contributing to the value of API as presented in Figure 4 and 5.From the bar chart in Figure 4, cluster 1 show that PM_{10} is the highest contributor to the API value followed by O_3. Cluster 2 and 3 show dominance in the influence of O_3 to the API value. Similarly, the result for the leave-one-out method (pie chart) in Figure 5 revealed that PM_{10} is the major contributor to the API value to the tune of about 60% followed by O_3 (36%) in cluster 1. About 92% of the API value in cluster 2 is influenced by O_3 while cluster 3 is also influenced by O_3 to a tune of about 97%.

Based on the result from the standardized coefficient and the leave-one-out method, it is statistically evidential that O_3 and PM_{10} are the most significant parameters influencing the value of API in the Peninsular Malaysia. This is largely attributed to the industrial activities in Kemaman, Paka-Kertih and Kuala Lumpur; heavy traffic congestion; biomass burning and construction activities in Kuala Terengganu, Putrajaya and Kuala Lumpur [62, 63, 76].

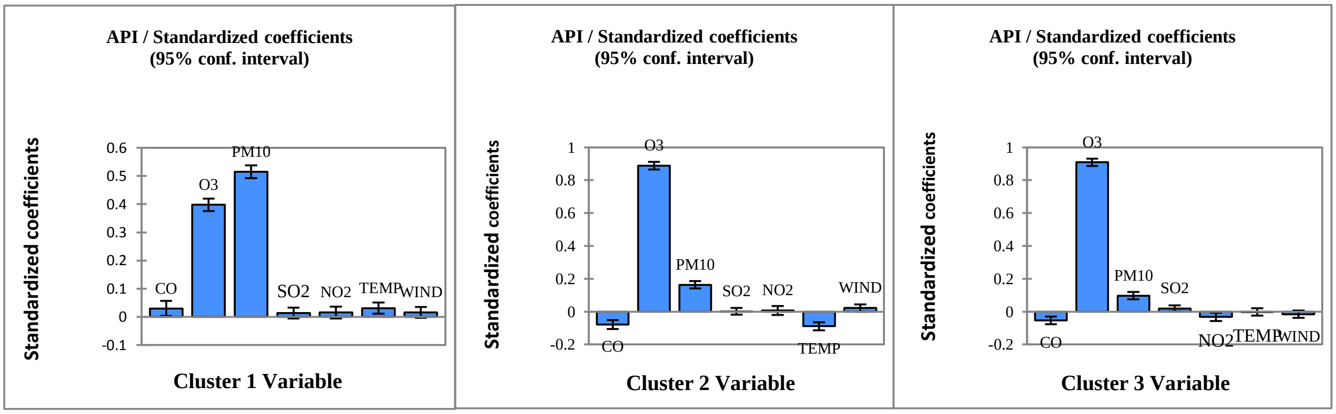


Figure 4 Contribution of each pollutant based on standardized coefficient

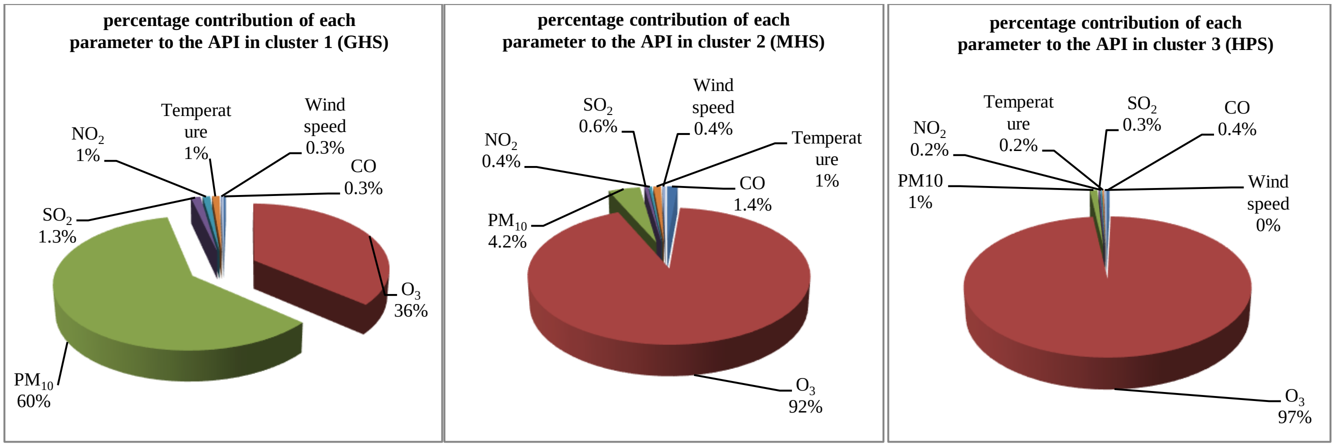


Figure 5 Percentage contribution of pollutants using the leave-one-out method

3.3.3 Source Apportionment of Air Pollution

The percentage mass contribution of each identified possible source category using principal component scores after varimax rotation and MLR model is shown in Figure 6.

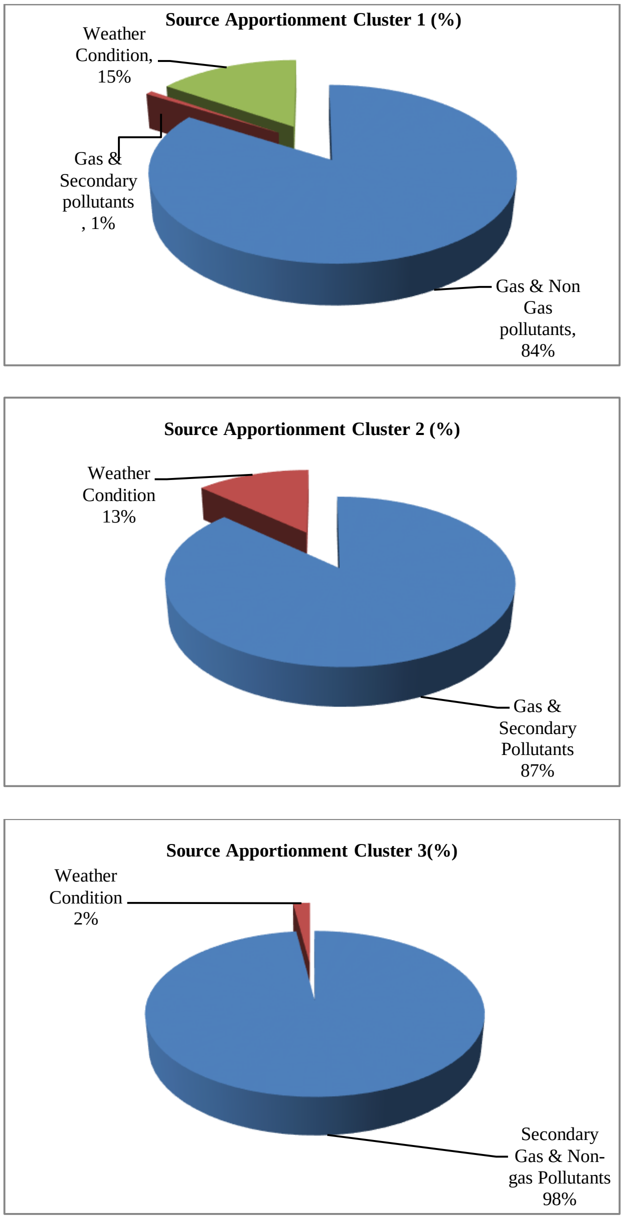


Figure 6 Pie chart for source category apportionment

Cluster 1 is influence to a degree of 84% by gas and non-gas pollutants, 1% by gas and secondary pollutants and 15% by weather condition. In cluster 2, the air pollutant index is influence to a degree of 87% by gas and secondary pollutants and 13% by weather condition. Cluster 3 is influenced by secondary gas and non-gas pollutants to a degree of 98% and 2% by weather condition. However, it is apparent to note that gas pollutants (NO_2 and CO) are the major precursors for the formation of secondary pollutants (O_3 and PM_{10}) [81].

Furthermore, it is important to reiterate that in the event of financial constraint or budget deficit, the government can concentrate in the monitoring and used of these two most significant parameters (O_3 and PM_{10}) in predicting API. This will indeed reduce the cost of equipment's, cost of monitoring redundant stations and pollutants as well as conserve time and energy.

4.0 CONCLUSION

This study has justified the usefulness of chemometric technique in modeling atmospheric air pollution. The result for HACA successfully groups the monitoring stations into three clusters base on their level of similarities and dissimilarities. This classification helps to eliminate unnecessary sampling sites which in turn will reduce the cost and time of monitoring. PCA revealed that the source of origin of these pollutants are from anthropogenically induced emission (automobiles, industries, power plants, transboundary sources and construction sites), which accounts for more than 62%, 56% and 58% of the total variance respectively. The explicit equation model developed by MLR show that cluster 3 has the highest $R^2 = 0.84\%$, followed by cluster 2 with an API value of about 82% and 51% for cluster 3. The percentage contribution of each pollutant using MLR indicates that PM_{10} and O_3 are the most significant parameters influencing the spatial variability of the air pollutant index. These two pollutants can indeed provide a spatial representation of the API within the study area base on the characteristics of the activities in the area. Furthermore, the source profile for each category was successfully apportioned using MLR and principal component scores. The result shows that weather condition, gas, non-gas and secondary gas pollutants play an important role in influencing the level of air quality in the Peninsular Malaysia.

Acknowledgement

The authors wish to extend their profound gratitude to the Department of Environment Malaysia for making available the data used for this study. East Coast Environmental Research Institute (ESERI) is highly appreciated for the wonderful assistance it provided in putting this work to reality.

References

[1] Jacob, D. J. & Winner, D. A. 2009. Effect of Climate Change on Air Quality. *Atmospheric Environment*. 43(1): 51-63.

[2] Rao, B. V., Ramjee, E. Reddy, V. V. 2014. Spatial and Temporal Determination of the Air Quality at Major Transportation Regions of the Hyderabad Metropolitant City: India. *International Journal of Engineering Research and Apps*. 4: 9-14.

[3] Salma, I., Chi, X., & Maenhaut, W. 2004. Elemental and Organic Carbon in Urban Canyon and Background Environments in Budapest, Hungary. *Atmospheric Environment*. 38(1): 27-36.

[4] Monks, P. S., Grainer, C. *et al.* 2009. Atmospheric Composition Changes-global and Regional Air Quality. *Atmospheric Environment*. 43: 4268-5350.

[5] MacNee, W., & Donaldson, K. 2003. Mechanism of Lung Injury Caused by PM_{10} and Ultrafine Particles with Special Reference to COPD. *European Respiratory Journal*. 21(40 suppl): 47s-51s.

[6] Ilyas, S. Z., Khattak, A. I., Nasir, S. M., Qurashi, T., &Durrani, R. 2010. Air Pollution Assessment in Urban Areas and Its Impact on Human Health in the City of Quetta, Pakistan. *Clean Technologies and Environmental Policy*. 12(3): 291-299.

[7] Ballester, F., Llop, S., Estarlich, M., Esplugues, A.,Rebagliato, M., & Iñiguez, C. 2010. Preterm Birth and Exposure to Air Pollutants During Pregnancy. *Environmental Research 110*. 8: 778-785.

[8] World Health Organization (WHO) 2002. World Health Report: Reducing Risk, Promoting Healthy Life. Geneva, Switzerland.

[9] Seinfled, J. H., Pandis, S. N. 1998. *Atmospheric Chemistry and Physics*. Wiley, Hobooken.

[10] Mutalib, S. N. S. A., Juahir, H., Azid, A., Sharif, S. M., Latif, M. T., Aris, A. Z., Zain, S. M., Dominick, D. 2013. Spatial and Temporal Air Quality Pattern Recognition Using Environmetric Techniques: A Case Study in Malaysia. *Environmental Science, Processes & Impacts*. 1-2.

[11] Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C. &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(8): 1-14.

[12] Jamal, H. H., Pillay, M. S., Zailina, H., Shamsul, B. S., Sinha, K., Zaman Huri, Z., Khew, S. L., Mazrura, S., Ambu, S., Rahimah, A., Ruzita, M. S. 2004. A Study of Health Impact & Risk Assessment of Urban Air Pollution in Klang Valley. UKM Pakarunding Sdn. Bhd., Malaysia, Kuala Lumpur.

[13] Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. 2012. Spatial Assessment of Air Quality Patterns in Malaysia Using Multivariate Analysis. *Atmospheric Environment*. 60: 172-181.

[14] Mohamad, J., Kiggundu, A. T. 2007. The Rise of the Private Car in Kuala Lumpur, Malaysia-Assessing The Policy Options. *IATSS Res.* 31(1): 69-77.

[15] Ministry of Transport Malaysia (MOT). 2010. Malaysia Transport Statistics Report. Putrajaya, Malaysia.

[16] Malaysia Meteorological Department (MMD). 2009. General Climate of Malaysia. Kuala Lumpur: Ministry of Science, Technology and Innovation.

[17] Malaysia Meteorological Department (MMD). 2010. General Climate of Malaysia. Kuala Lumpur: Ministry of Science, Technology and Innovation.

[18] Dawson, J. P., Adams, P. J., Pandis, S. N. 2007. Sensitivity of PM2.5 to climate in the Eastern U.S: A Modeling Case Study. *Atmos. Chem. Phys*. 7: 4295-4309.

[19] Kovac-Andric, E., Brana, J., Gvozdic, V. 2009. Impact of Meteorological Factors on Ozone Concentrations Modeled by Time Series and Multivariate Statistical Methods. *Ecological Informatics.* 4(2): 117-122.

[20] Kleeman, M. J. 2007. A Preliminary Assessment of the Sensitivity of Air Quality in California to Global Change. *Climatic Change*. 87: S273-S292.

[21] Wang, L., Li, J., Lu, H., Gu, Z., Rioual, P., Hao, Q. & Chu, G. 2012. The East Asian Winter Monsoon Over the Last 15,000 Years: Its Links to High-Latitudes and Tropical Climate Systems and Complex Correlation to the Summer Monsoon. *Quaternary Science Review.32*: 131-142.

[22] Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. 2010. Trend and Status of Air Quality at Three Different Monitoring Stations in The Klang Valley, Malaysia. *Air Quality, Atmosphere & Health.* 3(1): 53-64.

[23] Razak, M. I. M., Ahmad, I., Bujang, I., Talib, A. H., & Ibrahim, Z. 2013. Economics of Air Pollution in Malaysia. *Int J Humanit Soc Sci*. 3(13): 173-177.

[24] Department of Environment Malaysia (DOE). 2012. Malaysia Environmental Quality Report. Ministry of Science, Technology and Environment, Kuala Lumpur.

[25] Department of Environment Malaysia (DOE). 1997. A Guide to Air Pollution Index In Malaysia (API). Ministry of Science, Technology and Environment, Kuala Lumpur.

[26] Wu, E. M., Kuo, S., 2013. A Study of the Use of a Statistical Analysis Model to Monitor Air Pollution Status in an Air Quality-Total Control District. *Atmosphere.* 4: 349-364.

[27] Arhami, M., Kamali, N., & Rajabi, M. M. 2013. Predicting Hourly Air Pollutant Levels Using Artificial Neural Networks Coupled with Uncertainty Analysis by Monte Carlo Simulations. *Environmental Science and Pollution Research.* 20(7): 4777-4789.

[28] Simeonov, V., Einax, J.W., Stanimirova, I., Kraft, J. 2002 Environmental Modelling and Interpretation of River Water Monitoring Data. *Analytical and Bio Analytical Chemistry.* 374: 898-905.

[29] Juahir, H., Zain, S. M., Yusoff, M. K., Hanidza, T. T., Armi, A. M., Toriman, M. E., & Mokhtar, M. 2011. Spatial water quality Assessment of Langat River Basin (Malaysia) Using Environmetric Techniques. *Environmental Monitoring and Assessment*. 173(1-4): 625-641.

[30] Samsudin, M. S., Juahir, H., Zain, S. M., & Adnan, N. H. 2011. Surface River Water Quality Interpretation Using Environmetric Techniques: Case study at Perlis River Basin, Malaysia. *International Journal of Environmental Protection*.

[31] Al-Odaini, N. A., Zakaria, M. P., Zali, M. A., Juahir, H., Yaziz, M. I., & Surif, S. 2012. Application of Chemometrics in Understanding the Spatial Distribution of Human Pharmaceuticals in Surface Water. *Environmental Monitoring and Assessment.* 184(11): 6735-6748.

[32] Smith, D. G. 1990. A Better Water Quality Indexing System for Rivers and Streams. *Water Research.* 24(10): 1237-1244.

[33] Satheeshkumar, P., Khan, A. B. 2012 Identification of Mangrove Water Quality by Multivariate Statistical Analysis Methods in Pondicherry Coast. *India Environ Monit Assess.* 184(6): 3761-3774. DOI 10.1007/s10661-011-2222-4.

[34] Deepulal, P. M., Sujatha, C. H., & George, R. 2012. Chemometric Study on the Trace Metal Accumulation in the Sediments of the Cochin Estuary—Southwest Coast of India. *Environmental Monitoring and Assessment*. 184(10): 6261-6279.

[35] Wahid, N. B. A., Latif, M. T., &Suratman, S. 2013. Composition and Source Apportionment of Surfactants in Atmospheric Aerosols of Urban and Semi-Urban Areas in Malaysia. *Chemosphere*. 91(11): 1508-1516.

[36] Mustaffa, N. I. H., Latif, M. T., Ali, M. M., & Khan, M. F. 2014. Source Apportionment of Surfactants in Marine Aerosols at Different Locations Along the Malacca Straits. *Environmental Science and Pollution Research.* 21(10): 6590-6602.

[37] Thurston, G. D Spengler, J. D. 1985. A Quantitative Assessment of Source Contributions to Inhalable Particulate Matter Pollution in Metropolitan Boston. *Atmospheric Environment*. 19: 9-25.

[38] Thurston, G. D., K. Ito, T. Mar, W. F. Christensen, D. J. Eatough, R. C. Henry, E. Kim, F. Laden, R. Lall, T. V. Larson, H. Liu, L. Neas, J. Pinto, M. Stolzel, H. Suh, P. K. Hopke. 2005. Dec. Workgroup Report: Workshop on Source Apportionment of Particulate Matter Health Effects-Intercomparison of Results and Implications. *Environ. Health Perspect*. 113(12): 1768-1774.

[39] Jaafar, S. A., Latif, M. T., Chian, C. W., Han, W. S., Wahid, N. B. A., Razak, I. S.& Tahir, N. M. 2014. Surfactants in the Sea-Surface Microlayer and Atmospheric Aerosol Around the Southern Region of Peninsular Malaysia. *Marine pollution bulletin*. 84(1): 35-43.

[40] Chatterjee, S., Price, B. 1999. Regression Analysis by Example. third ed. Wiley, Chichester.

[41] Petrie, A., Sabin, C. 2012. Medical Statistics at a Glance. Blackwell Science, Oxford. Querol, X., Alastuey, A., Lopez-Soler, A., Plana, F., Mantilla, E., Juan, R., Ruiz, C. R., La Orden, A. Characterisation of Atmospheric Particulates Around a Coal Fired Power Station. *Int. J. Coal Geol*. 40: 175-188.

[42] Department of Statistic Malaysia. 2010. Basic Population Characteristics by Administrative Districts. 2009 Report.

[43] Department of Environment Malaysia (DOE). 2007. Malaysia Environmental Quality Report, Ministry of Science, Technology and Environment, Kuala Lumpur.

[44] Department of Statistic Malaysia. 2010. Basic Population Characteristics by Administrative Districts. 2009 Report.

[45] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. 2004. Methods for Imputation of Missing Values in Air Quality Data Sets. *Atmospheric Environment*. 38(18): 2895-2907.

[46] Wang, Y. B., Liu, C. W., Liao, P. Y., & Lee, J. J. 2014. Spatial Pattern Assessment of River Water Quality: Implications of Reducing the Number of Monitoring Stations and Chemical Parameters. *Environmental Monitoring and Assessment.* 186(3): 1781-1792.

[47] Deepulal, P. M., Sujatha, C. H., & George, R. 2012. Chemometric Study on the Trace Metal Accumulation in the Sediments of the Cochin Estuary—Southwest Coast of India. *Environmental Monitoring and Assessment*. 184(10): 6261-6279.

[48] Yang, Y., Liu, Z., Chen, F., Wu, S., Zhang, L., Kang, M., & Li, J. 2014. Assessment of Trace Element Contamination in Sediment Cores from the Pearl River and Estuary, South China: Geochemical and Multivariate Analysis Approaches. *Environmental Monitoring and Assessment*. 186(12): 8089-8107.

[49] Zhang, X., Jiang, H., & Zhang, Y. 2013. Spatial Distribution and Source Identification of Persistent Pollutants in Marine Sediments of Hong Kong. *Environmental Monitoring and Assessment*. 185(6): 4693-4704.

[50] Lau, J., Hung, W. T., Cheung, C. S. 2009. Interpretation of Air Quality in Relation to Monitoring Station's Surrounding. *Atmospheric Environmetric*. 43: 769-777.

[51] Singh, K. P., Malik, A., Sinha, S. 2005. Water Quality Assessment and Apportionment of Pollution Sources of Gomti River (India) Using Multivariate Statistical Techniques E A Case Study. *Analytica Chimica Acta.* 538: 355-374.

[52] Shrestha, S., & 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(4): 464-475.

[53] Osman, R., Saim, N., Juahir, H., & Abdullah, M. P. 2012. Chemometric Application in Identifying Sources of Organic Contaminants in Langat River Basin. *Environmental Monitoring and Assessment.* 184(2): 1001-1014.

[54] Brumelis, G., Lapina, L., Nikodemus, O., Tabors, G. 2000. Use of an Artificial Model of Moitoring Data to Aid Interpretation of Principal Component Analysis. *Environmental Modeling & Software*. 15 (8): 775-763.

[55] Love, D., Hallbauer, D., Amos, A., & Hranova, R. 2004. Factor Analysis as a Tool in Groundwater Quality Management: Two Southern African Case Studies. *Physics and Chemistry of the Earth, Parts A/B/C*. 29(15): 1135-1143.

[56] Liu, C. W., Lin, K. H., & Kuo, Y. M. 2003. Application of Factor Analysis in the Assessment of Groundwater Quality in a Blackfoot Disease Area in Taiwan. *Science of the Total Environment*. 313(1): 77-89.

[57] Shi, G. L., Zeng, F., LI, X., Feng, Y. C., Wang, Y. Q., Liu, G. X., Zhu, T. 2011. Estimated Contributions and Uncertainties of PCA/MLR -CMB Results: Source Apportionment For Synthetic and Ambient Datasets. *Atmos. Environ*. 45: 2811-2819.

[59] Ilten, N., & Selici, A. T. 2008. Investigating the Impacts of Some Meteorological Parameters on Air Pollution in Balikesir, Turkey. *Environmental Monitoring and Assessment.* 140(1-3): 267-277.

[60] Department of Environment Malaysia (DOE). 2005. Malaysia Environmental Quality Report, Ministry of Science, Technology and Environment, Kuala Lumpur.

[61] Department of Environment Malaysia (DOE). 2010. Malaysia Environmental Quality Report, Ministry of Science, Technology and Environment, Kuala Lumpur.

[62] Abdullah, A. M., Abu Samah, M. A., & Jun, T. Y. 2012. An Overview of the Air Pollution Trend in Klang Valley, Malaysia. *Open Environ Sci*. 6: 13-19.

[63] Sara, Y. Y., Rashid, M., Chuah, T. G., Suhaimi, M., & Mohamed, N. N. 2013. Characteristics of Airborne Pm2.5 and Pm2.5-10 in the Urban Environment of Kuala Lumpur. *Advanced Materials Research.* 620: 502-510.

[64] Janssen, N. A., van Vliet, P. H., Aarts, F., Harssema, H., & Brunekreef, B. 2001. Assessment of Exposure to Traffic Related Air Pollution of Children Attending Schools Near Motorways. *Atmospheric Environment*. 35(22): 3875-3884.

[65] Morawska, L., Jayaratne, E. R., Mengersen, K., Jamriska, M., & Thomas, S. 2002. Differences in Airborne Particle and Gaseous Concentrations in Urban Air Between Weekdays and Weekends. *Atmospheric Environment.* 36(27): 4375-4383.

[66] Pandey, S. K., Tripathi, B. D., & Mishra, V. K. 2008. Dust Deposition in a Sub-Tropical Opencast Coalmine Area, India. *Journal of Environmental Management*. 86(1): 132-138.

[67] Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., &Tager, I. 2004. Air Pollution and Cardiovascular Disease a Statement for Healthcare Professionals from the Expert Panel on Population And Prevention Science of the American Heart Association. *Circulation.* 109(21): 2655-2671.

[68] Arsene, C., Olariu, R. I., & Mihalopoulos, N. 2007. Chemical Composition of Rainwater in the Northeastern Romania, Iasi Region (2003–2006). *Atmospheric Environment.* 41(40): 9452-9467.

[69] Yusof, N. F. F. M., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A., & al Madhoun, W. 2010. Monsoonal Differences and Probability Distribution of PM10 Concentration. *Environmental Monitoring and Assessment*. 163(1-4): 655-667.

[70] Ghazali, N. A., Ramli, N. A., & Yahaya, A. S. 2009. A Study to Investigate and Model the Transformation of Nitrogen Dioxide into Ozone Using Time Series Plot. *European Journal of Scientific Research*. 37(2): 192-205.

[71] Intergovernmental Panel on Climate Change (IIPC). 1999. IIPC Special Report: Aviation and the Global Atmosphere, Summary for Policy Makers. IIPC - XV/DOC 99.

[72] Propecus, F., Lonel, I., Belegante, L., Lontis, N., Cebrucean V. 2011. Direct Measurement and Numerical Simulation Issues in Airport Air Quality. *International Journal of Energy & Environment.* 3(5): 410-417.

[73] Scheelhaase, J. D., Grimme, W. G. 2007. Emissions Trading for International Aviation, Estimation of the Economic Impact on Selected European Airlines. *Journal of Air Transport Management*. 13: 253-263.

[74] Malaysia Meteorological Department (MMD). 2012. General Climate of Malaysia Kuala Lumpur: Ministry of Science, Technology and Innovation.

[75] Brunekreef, B., Holgate, S. T. 2012. Air Pollution and Health Lancet. 360: 1233-1242.

[76] Banan, N., Latif, M. T., Juneng, L., & Ahamad, F. 2013. Characteristics of Surface Ozone Concentrations at Stations with Different Backgrounds in the Malaysian Peninsula. *Aerosol and Air Quality Research.* 13(3): 1090-1106.

[77] Mittal, M. L., Hess, P. G., Jain, S. L., Arya, B. C., & Sharma, C. 2007. Surface Ozone in the Indian Region. *Atmospheric Environment.* 41(31): 6572-6584.

[78] Sadanaga, Y., Sengen, M., Takenaka, N., & Bandow, H. 2012. Analyses of the Ozone Weekend Effect in Tokyo, Japan: Regime of Oxidant (O3+ NO2) Production. *Aerosol Air Qual. Res*. 12: 161-168.

[79] Atkinson, R. 2000. Atmospheric Chemistry of VOCs and NOx. *Atmospheric Environment*. 34: 2063-2101.

[80] Bauer, S. E., & Langmann, B. 2002. Analysis of a Summer Smog Episode in the Berlin-Brandenburg Region with a Nested Atmosphere-Chemistry Model. *Atmospheric Chemistry and Physics.* 2(4): 259-270.

[81] Latif, M. T., Azmi, S. Z., Noor, A. D. M., Ismail, A. S., Johny, Z., Idrus, S., & Mokhtar, M. B. 2011. The Impact of Urban Growth on Regional Air Quality Surrounding the Langat River Basin, Malaysia. *The Environmentalist*. 31(3): 315-324.