Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

AIR QUALITY MODELLING USING CHEMOMETRIC TECHNIQUES

A. Azid[1,2,*], N. A. A. Rani[1], M. S. Samsudin[1], S. I. Khalit[1,2], M. B. Gasim[1,3], M. K. A. Kamarudin[3], K. Yunus[4], A. S. M. Saudi[5] and K. M. K. K. Yusof[1]

[1]Faculty of Bioresources and Food Industry, Universiti Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia

[2]UniSZA Science and Medicine Foundation Centre, Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia

[3]East Coast Environmental Research Institute (ESERI), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia

[4]Kulliyyah of Science, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia

[5]Institute of Medical Science and Technology, University of Kuala Lumpur, 43600 Kajang, Selangor, Malaysia

Published online: 08 August 2017

ABSTRACT

The datasets of air quality parameters for three years (2012-2014) were applied. HACA gave the result of three different groups of similarity based on the characteristics of air quality parameters. DA shows all seven parameters (CO, O3, PM10, SO2, NOx, NO and NO2) gave the most significant variables after stepwise backward mode. PCA identifies the major source of air pollution is due to combustion of fossil fuels in motor vehicles and industrial activities. The ANN model shows a better prediction compared to the MLR model with R2 values equal to 0.819 and 0.773 respectively.

Author Correspondence, e-mail: azmanazid@unisza.edu.my

doi: http://dx.doi.org/10.4314/jfas.v9i2s.30

This study presents that the chemometric techniques and modelling become an excellent tool in API assessment, air pollution source identification, apportionment and can be setbacks in designing an API monitoring network for effective air pollution resources management.

Keywords: air pollutant index; chemometric; ANN; MLR.

1. INTRODUCTION

Air pollution is becoming a major environmental issue around the world. Transportations (mobile sources), trans-boundary pollution from neighbouring countries and the industrial activities (stationary sources) are the main sources of air pollution in the world, and Malaysia is one of those countries [1-2]. The effect of air pollution may cause acute and chronic to humans or other living organisms, and cause damage to the natural environment or built environment when enter into the atmosphere [3]. Symptoms such as nose, throat, eye and skin irritation, headache, fatigue, dizziness and difficulty in breathing are general health effect experienced by human due to poor of air quality [4].

The application of chemometric techniques (also known as multivariate techniques) constitutes one of the new and robust statistical methods used by researchers to interpret a large number of data set. It is based on the statistical principle, which involves a simultaneous observation and analysis of more than one variable by simplifying the process within a convenient size. Hydrologist, engineers and environmental scientist are usually faced with a large amount of data, which can be best synthesized using multivariate analysis.

Chemometric techniques such as hierarchical agglomerative cluster analysis (HACA), discriminant analysis (DA) and principal component analysis (PCA) have been verified to be a functional tool, simpler and more easily interpretable results [5]. These methods help to lessen the complexity of databases so that a better understanding and interpretation of air quality data can be achieved successfully for efficient management of the air quality monitoring programs [1-2]. For predicting purposes, artificial neural network (ANN) and multiple linear regression (MLR) were applied due to it has strong capability in predicting the complicated of data, can be trained accurately and gives a better performance compared to other models [6-7].

This study aims to identify the spatial variations in air pollutant index (API) using chemometric techniques. This study also aims to establish the API prediction model using ANN and MLR approaches.

2. EXPERIMENTAL

This study focused on the area of the southern region in Peninsular Malaysia. These areas consist of three parts of states: Johor, Malacca and Negeri Sembilan. These three states have about 27,560 km^2 of total area and widely known as the most developed states in Malaysia. Eight stations were selected for this study-Pasir Gudang (CA0001: N01° 28.225 E103° 53.637), Bukit Rambai (CA0006: N02° 15.510 E102° 10.364), Nilai (CA0010: N02° 49.246 E101° 48.877), Johor Bahru (CA0019: N01° 29.815 E103° 43.617), Bachang (CA0043: N02° 12.789 E102° 14.055), Muar (CA0044: N02° 03.715 E102° 35.587), Seremban (CA0047: N02° 43.418 E101° 58.105) and Tampoi (CA0051: N01° 29.068 E103° 41.064) as can be seen in Fig. 1. All the stations were selected due to the fact that they are in heaviness industrial, residential and commercial areas. The annual mean wind speed of 10m in these areas is 5.1 km/hr. The daily traffic density is moderate to high, which the peak periods found during morning and evening hours.

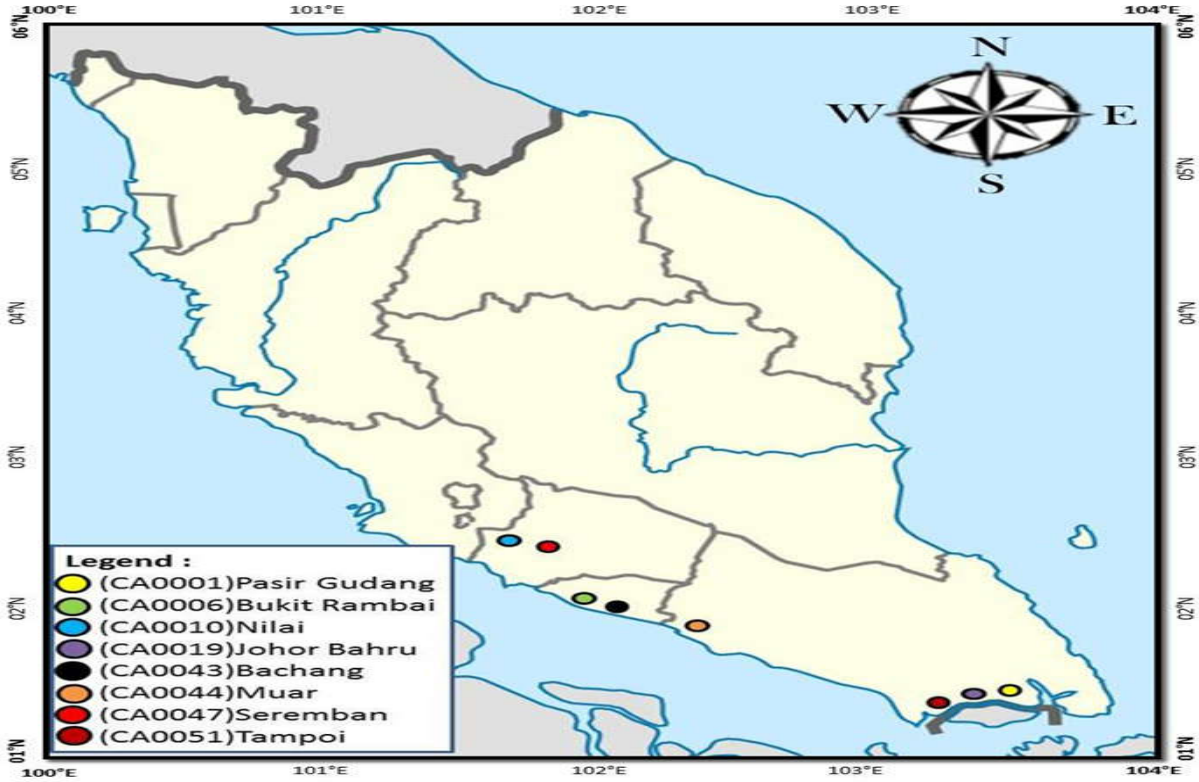

Fig.1. Map of the study area

2.1. Chemometrics Technique

Hierarchical agglomerative cluster analysis (HACA), discriminant analysis (DA), principal component analysis (PCA), multiple linear regression (MLR) and artificial neural network (ANN) were performed on seven parameters to rank their relative significance and to describe their interrelation. XLSTAT2014 software was used to analyse the data with HACA, DA, PCA and MLR while JMP10 software for ANN analysis.

2.2. Cluster Analysis

In this study, HACA was used to investigate the spatial of the sampling sites. HACA is a method to classify the objects (cases) into classes (cluster) so that each object is similar to the others within a class, but different from those in other classes with respect to a predetermined selection criterion [16-18]. In HACA, a Euclidean distance (linkage distance) in Ward's method was used for similarity measurement and has proved to be a very efficient method [19-22]. Dendrogram is a result and illustrated based on the most common approach in hierarchical agglomerative clustering [18, 23]. The Dlink/Dmax of Euclidean distance which represents the quotient between the linkage distances shall be divided by the maximal distance [24, 22], and the quotient is multiplied by 100 as a way to standardize the linkage distance

represented by the y-axis [25-27].

2.3. Discriminant Analysis

In this study, DA was employed to determine whether the groups differ with regard to the mean of a variable and used the variable to predict group membership. The purpose of the DA is to determine the variables that discriminate between two or more naturally occurring groups/clusters [22]. The constructs discriminant functions (DFs) which are calculated using Equation (1) [16]:

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

where i is the number of groups (G), ki 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 wj is the weight coefficient assigned by DF analysis (DFA) to a given parameter (pj).

Three groups for spatial analysis (three sampling regions represent the unhealthy site (UHS), moderate healthy site (MHS) and good healthy site (GHS)) which gathered from HACA were selected. The DA was applied to the raw data using the standard, forward stepwise and backward stepwise modes. These data were used to construct DFs to evaluate spatial variations in the ambient air quality. The spatial were the dependent variables, while all the measured parameters constitute the independent variables. In the stepwise forward mode, variable are included step by step starting with the most significant variable until no significant changes were obtained. While, for stepwise backward mode, variables are removed step by step beginning with the least significant variable until no significant changes were obtained.

2.4. Principal Component Analysis

PCA is the most powerful pattern recognition technique and it illustrates the most significant parameters in which it describes the whole data set rendering data reduction with minimum loss of the original information [25-26, 28]. The PCA usually coupled with HACA [22]. The principal component (PC) is expressed as Equation (2):

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

where z is the component score, a is the component loading, x is the measured value of the variable, i is the component number, j is the sample number and m is the total number of

variables.

The PCs are advisable to rotate by varimax rotation due to sometimes it is not readily interpreted [16]. When eigenvalue is more than 1, varimax rotations shall be applied on the PCs [29], in order to obtained new groups of variable (varimax factors, VFs). The number of VFs from varimax rotations is equal to the number of variable in accordance with common features and can include unobservable, hypothetical and latent variables [30]. The VF coefficient which having a correlation of > 0.75 is indicated as "strong", 0.50-0.75 as "moderate" and 0.30-0.49 as "weak" significant factor loading [1, 31]. In this study, PCA was applied to the normalized data sets (seven variables) separately for three different spatial regions, which are UHS, MHS and GHS as delineated by the HACA analysis. The input data matrices (variable x cases) for PCA were 7 x 2979 for UHS, 7 x 891 for MHS and 7 x 4890 for GHS regions.

2.5. Artificial Neural Network

The ANN was developed in the 1950s is a branch of artificial intelligence which aiming at imitating the biological brain architecture [3]. ANN proved its ability to provide better predicting which the results are depending on the use of a large number of inputs [32].

The main objective of designing and building ANN in this study is to determine the significant parameter that was affected by API. The ANN model was developed using seven significant parameters due to all DAs results (normal, stepwise forward and stepwise backward) shows all variables were accepted to build an API. These seven parameters were applied as input selection and API as the output for all models developed, and were divided into training and validation phases of the API prediction model. The data used in this study consists of 8760 observations (data sets). The original values of API were used as a reference. Hence, the ANN shall be compared to the reference model.

The training of the network (Levenberg-Marquardt algorithm) shows significantly faster convergence and able to find better local error minima compared with error back propagation algorithm [33]. The training was carried out based on the early stopping technique and it is interpreted when the measure of the error between the output and the target values of validation reaches a minimum. The root mean square error (RMSE) is the error function used

in the network. In this matter, the early stopping is done so that over training of the network is avoided. If the training is still continued until the performance function is minimized, the network will accept all the noise and will fail to generalize.

In this study, the trial-and-error procedure between one to ten hidden layers in the network structure were examined for ANN modelling in order to approximate any nonlinear function with any level of accuracy and it used to search the best model for prediction of the API. Based on the theoretical studies, a network with a small number of nodes shall probably fail to learn the data, while too many nodes shall fatefully over-fit the training patterns in the network and give a poor generalization performance's result, especially when dealing with noisy data in predicting problems [33].

There are two different criteria that were used in order to evaluate the effectiveness of each network and its ability to make precise prediction [33], which are known as the coefficient of determination (R^2) and the root mean square error (RMSE). The equations of R^2 and RMSE can be referred as Equation (3) and (4):

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{SS_{reg}/n}{SS_{tot}/n} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n}} \tag{4}$$

where, yj stand for the measured value, yj̃ is the estimated value of the dependent variable and n is the number of observations.

The higher the values of R^2 (nearest to 1) and the lower the values of RMSE, the more accurate the predictions are [15-16, 34]. Fig. 2 shows the example of the network structure that was developed for the prediction of the API.

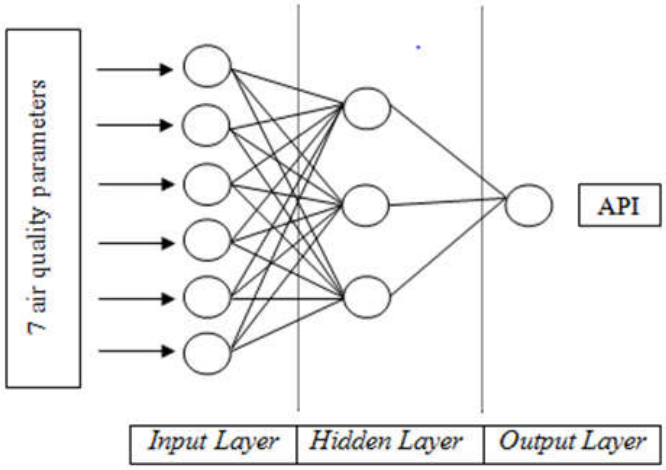

Fig.2. Example of ANN model network structure of this study

2.6. Multiple Liner Regression

The MLR is a modelling technique that widely used to investigate the relationship between two or more independent variables and the dependent variable by fitting a linear equation to observe the data [14, 35]. In this study, it was employed to justify the relationship between the ambient air quality parameters and their impact on the API. This model was applied in this study in order to compare the performance of artificial neural network (ANN) based approaches. The model is a generalization of the simple linear regression model, which every value of the independent variable is associated with a value of the dependent variable. The model is obtained using the Equation (5):

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_{1i2} \ \ldots + \beta_{p-1} x_{p-1} + \varepsilon \quad (5)$$

where Y is the response variable, and there are p – 1 explanatory variable x1, x2, …, xp–1, with p parameters (regression coefficients) β0, β1, β2, …, βp-1 and ε is an error associated with the regression.

All the ambient air quality parameters (seven parameters) were selected due to significant levels (p) values are below than 0.0001. Then, the R^2 and RMSE were recorded and compared to the ANN model.

3. RESULTS AND DISCUSSION

3.1. Classification of Sampling Station Based On the Air Pollutant Index

From the result, it is evident that the HACA technique is useful in offering reliable

classification of ambient air for the whole region. It also can be used to design future spatial sampling strategies in an optimal manner. However, from time to time, DOE is advisable to make verification in all stations.

This analysis examined about the historical values of the air pollutant index among eight stations along the southern region of Peninsular Malaysia in order to classify the air pollutants index stations based on its similarity level using HACA. This analysis was performed to evaluate the spatial variation among the sampling sites. From the analysis, three groups or clusters of sampling stations were identified (Fig. 3) namely unhealthy site (UHS), moderate healthy site (MHS) and good healthy site (GHS). Fig. 4 shows the three significant regions illustrated by HACA and the potential pollution sources within the study regions.

From the analysis, cluster-I represent MHS which formed by the monitoring sites of CA0001 and CA0019 (Pasir Gudang and Johor Bahru), cluster-II represents UHS of CA0006, CA0010 and CA00051 (Bukit Rambai, Nilai and Tampoi) and cluster-III represents GHS of CA0043, CA0044 and CA0047 (Bachang, Muar and Seremban). This finding implies that for rapid assessment of the air pollutant index region based on three years data, only one station in each cluster is needed in order to represent a reasonably accurate spatial assessment of the ambient air quality for the whole network. The purpose of the CA technique in this study was to reduce the need for numerous sampling stations. Monitoring from three stations that represent three different regions is sufficient.

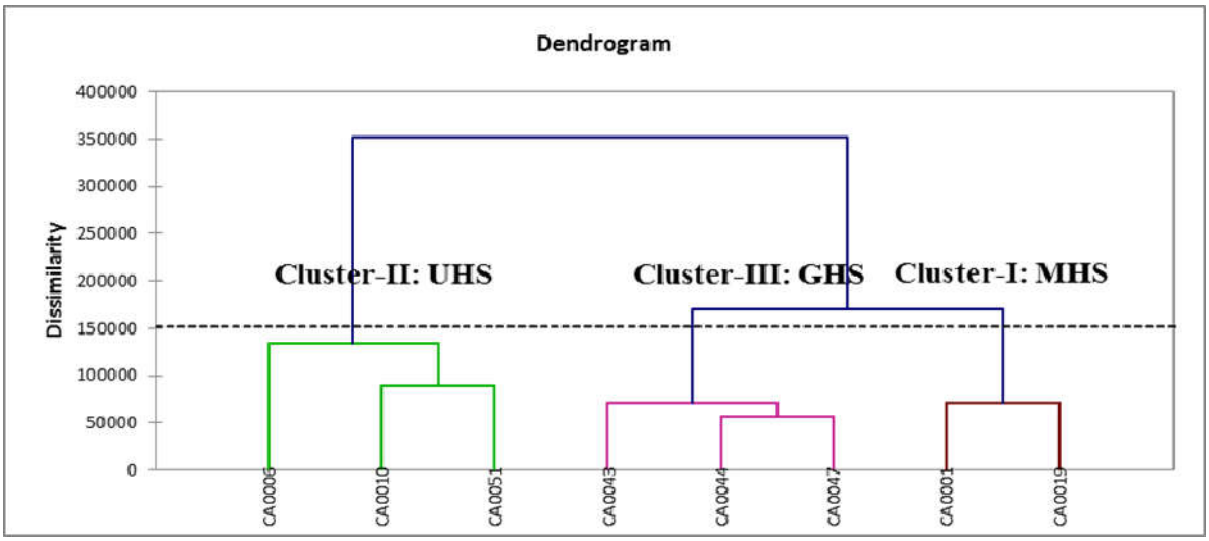

Fig.3. Dendrogram showing different clusters of sampling sites located in the Southern of

Peninsular Malaysia based on the air pollutant index

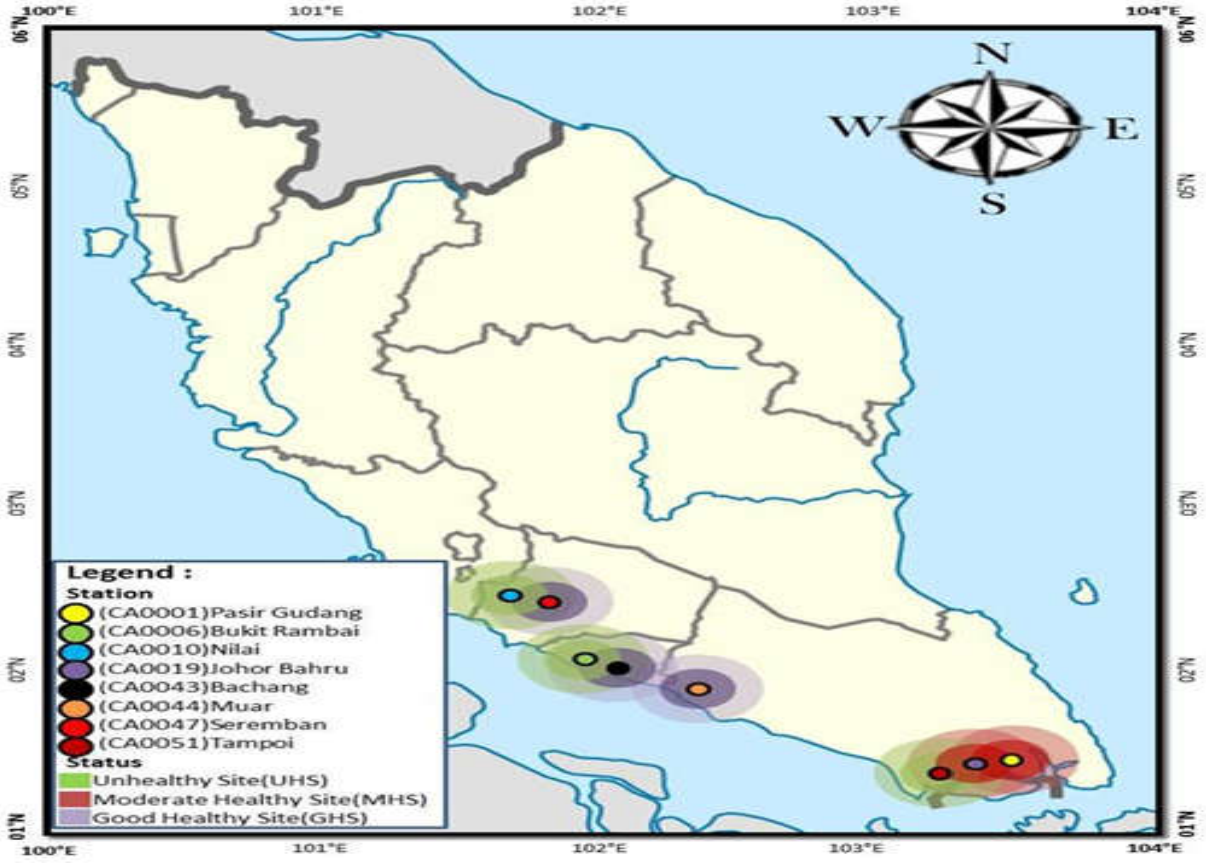

Fig.4. Classification of regions due to air quality status by HACA in the southern region of Peninsular Malaysia

3.2. Spatial Variations of Air Pollutant Index

To study about the spatial variation, DA was applied to the raw data of the southern of Peninsular Malaysia into three main groups (clusters) which has been defined by the HACA. The groups of UHS, MHS and GHS were treated as dependent variables while the ambient air quality parameters were treated as independent variables. In this study, DA was carried out through standard, stepwise forward and stepwise backward methods. The spatial classification accuracy by using standard (69.6%), stepwise forward (91.0%) and stepwise backward (94.4%) mode DFA for seven discriminant variables were identified respectively (Table 1). Seven variables (CO, NO_2, O_3, SO_2, NO_x, NO and PM_{10}) were used due to p-value of all parameters after stepwise forward and backward test have high variation in terms of their spatial distribution. The Wilk's Lambda (Rao's approximation) test for standard mode gave a Lambda value was 0.538 and $p < 0.0001$.

The null hypothesis, H0 states that the mean vectors of the three classes are equal. The alternative hypothesis, Ha state that at least one of the means vector is different from another. Since the computed p-value (< 0.0001) is lower than the significance level alpha = 0.05, one should reject the H0 and accept the Ha. The risk to reject the H0, while it is true is lower than 0.01%. Therefore, the three clusters are indeed different from one another. Then, seven selected variables which gave the most significant by stepwise backward in DA were used for further analysis.

PCA was applied to the data set in order to compare the compositional patterns between the examined ambient air quality parameters and to identify the factors that influence of UHS, MHS and GHS regions. Two PCs were obtained from the GHS region and three PCs from each of the MHS and UHS regions with eigenvalues greater than one and almost 70.2%, 79.3% and 82.3% of the total variance in the data set, respectively. However, component three in the MHS and UHS was accepted due to it has eigenvalue closed to one. The total numbers of VFs were obtained in these regions via FA performed on the PCs. The summary loadings of ambient air quality variables on the varimax-rotated PCs are presented in Table 2.

3.2.1 GHS

Among two VFs for GHS, VF1 comprises about 44.4% of the total variance which it shows strong positive loadings on the NO_x, NO and NO_2. This factor contains dangerous gaseous parameters that are formed naturally from the atmosphere by lightning, plants (agriculture fertilization), soil and water. Besides that, the sources of these gases also may come from the combustion of fossil fuels, motor vehicle exhaust, manufacturing industries and food processing [8-9]. For VF2, the total variance was 25.9% with strong positive loading on O_3 and PM_{10}, which possibly contributed by made and natural sources. The pollution loadings of O_3 are occurred from electrical discharge, electromagnetic radiation and consumer application that related to the oxidization [10]. This can be explained by considering the large number of residents in the area that use home appliances that consists of O_3 compound. While the presence of PM_{10} in GHS region is possibility occur naturally, which originating from urban development areas involving the clearing of land, and living vegetation from agricultural activities. Fossil fuels in vehicles, various industrial processes (timber plant, etc.) and power

plants also gave a significant amount of PM_{10}.

Table 1. Classification matrix for DA of spatial variations in the southern region of Peninsular Malaysia

Sampling Regions	% Correct	Regions Assigned by DA		
		UHS	MHS	GHS
Standard DA Mode (7 Variables)				
UHS	71.3	2341	258	686
MHS	41.6	524	911	755
GHS	86.5	370	72	2843
Total	69.6	3235	1241	4284
Stepwise Forward DA Mode (7 Variables)				
UHS	89.4	2890	80	264
MHS	66.1	85	806	348
GHS	99.8	5	5	4277
Total	91.0	2980	891	4889
Stepwise Backward DA Mode (7 Variables)				
UHS	89.2	2656	77	246
MHS	82	34	731	126
GHS	99.9	2	2	4886
Total	94.4	2692	810	5258

3.2.2. MPS

In the case of MPS, the first VF1 explains 43.3% of the total variance and it has strong positive loadings on CO, NO and NO_x. This factor could be considered by chemical components of various anthropogenic activities such as industrial, domestic, commercial and agricultural activities in the study area. VF2 explains 22.1% of the total variance and shows strong positive loading on PM_{10}, which related to industrial activities such as flour mill plant especially in Pasir Gudang area. VF3 explains 13.8% of the variance and have positive loading on SO_2, which are related to volcanoes activity from neighbouring countries and in various industrial processes such as combustion of coal and petroleum compound. Besides that, especially in Pasir Gudang,

there have more than 300 manufacturing companies and the world's largest edible oil tankage facility, which constitute a point source of pollution.

Table 2. Loadings of environmental variables on the varimax-rotated PCs for ambient air quality data collected from GHS, MHS and LPS of the Southern region of Peninsular Malaysia

Variables	GHS		MHS			UHS		
	VF1	VF2	VF1	VF2	VF3	VF1	VF2	VF3
CO	0.622	0.545	**0.775**	0.319	-0.324	0.410	**0.851**	-0.107
O_3	-0.443	**0.740**	-0.499	0.694	0.054	-0.557	0.553	0.226
PM_{10}	0.136	**0.869**	0.287	**0.751**	0.138	-0.184	**0.909**	0.189
SO_2	0.340	0.439	-0.086	0.068	**0.983**	0.117	0.072	**0.956**
NO_x	**0.977**	0.047	**0.961**	0.133	-0.029	**0.968**	0.051	0.125
NO	**0.886**	-0.197	**0.943**	-0.106	-0.028	**0.905**	-0.056	0.014
NO_2	**0.722**	0.322	0.355	0.631	-0.015	0.680	0.220	0.271
Eigenvalue	3.105	1.811	3.034	1.550	0.968	2.783	2.024	0.955
Variability (%)	44.352	25.875	43.341	22.147	13.824	39.754	28.915	13.647
Cumulative %	44.352	70.227	43.341	65.488	79.312	39.754	68.669	82.316

3.2.3. UHS

Finally, for the UHS region, the VF1 explains about 39.8% of the total variance and has strong positive loadings on NO_x and NO. The existence of NO_x and NO are related to the influence of anthropogenic and natural sources such as related to the clearing of land especially when biomass is burnt, combustion of fossil fuels, motor vehicle exhaust and various manufacturing industries. VF2 and VF3 explain 28.9% and 13.7% respectively of the total variance and shows strong positive loading on CO, PM_{10} and SO_2. Vehicle engines operate at high temperatures, which emitted in the exhaust fumes may cause the presence of CO in the UHS region. While the strong positive loading on PM_{10} and SO_2 are suspected to originate from a variety of mobile and stationary sources such as combustion of coal and oil, wood stoves, power plants, diesel truck, etc. due to industrial activities. Volcanoes, dust storms and

combustion of forest and grassland from neighbouring countries are also considering the contributor of the pollutants (CO, PM_{10} and SO_2) in the area.

3.3. The Application of API for Air Quality Classification

The DOE was released the API readings for the first time by the government of Malaysia since 1997. Currently, there are 51 automatic stations located throughout Malaysia. Five air pollutant parameters namely CO, NO_2, O_3, SO_2 and PM_{10} were used to determine the air quality status and to classify the ambient air based on the API. Although, the API is an effective tool for air monitoring, the relevance in other particular parameters such as NO_x, and NO which are not listed in Malaysia's API calculation was tested together in this study. From the study, these two parameters show the influence and relevance to the API monitoring. The findings from this study indicate that all of the parameters tested exhibit positive correlation for all VFs.

3.4. Predicting the API Data Using ANN

Table 3 illustrates the structures of constructing networks and their performance level. The optimum trained ANN structures were selected according to the minimum of RMSE and maximum R^2 values of the test set. Ten of hidden nodes were examined for best ANN structure and the graph of R^2 and RMSE were plotted (Fig. 5). From the findings, the model of ANN 9 was selected due to the R^2 and RMSE values of the test set are equal to 0.819 and 6.424 respectively (training phase). While, for validation phase, the R^2 and RMSE values were 0.796 and 6.764 respectively. Training and validation prediction performance of the best performance network (ANN 9) was shown in Fig. 6.

Table 3. ANN structure optimization

Models	No. of Hidden Neurons Layer	Training R^2	Training RMSE	Validation R^2	Validation RMSE
ANN 1	1	0.79	6.919	0.767	7.231
ANN 2	2	0.798	6.779	0.78	7.024
ANN 3	3	0.801	6.737	0.776	7.089
ANN 4	4	0.803	6.692	0.785	6.948
ANN 5	5	0.8	6.743	0.781	7.012
ANN 6	6	0.804	6.682	0.784	6.967
ANN 7	7	0.807	6.631	0.784	6.954
ANN 8	8	0.805	6.658	0.78	7.017
ANN 9	**9**	**0.819**	**6.424**	**0.796**	**6.764**
ANN 10	10	0.803	6.701	0.779	7.04

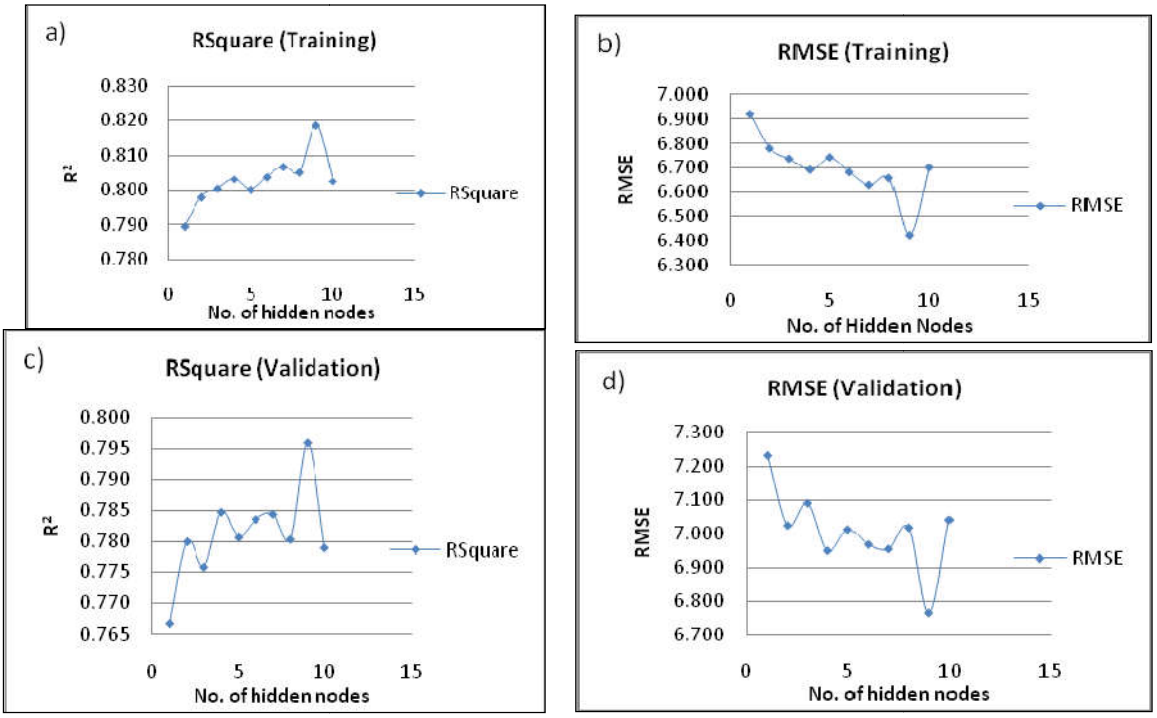

Fig.5. The performance of each ANN model obtained by: a)-b) R^2 and RMSE values (training), and c)-d) R^2 and RMSE values (validation)

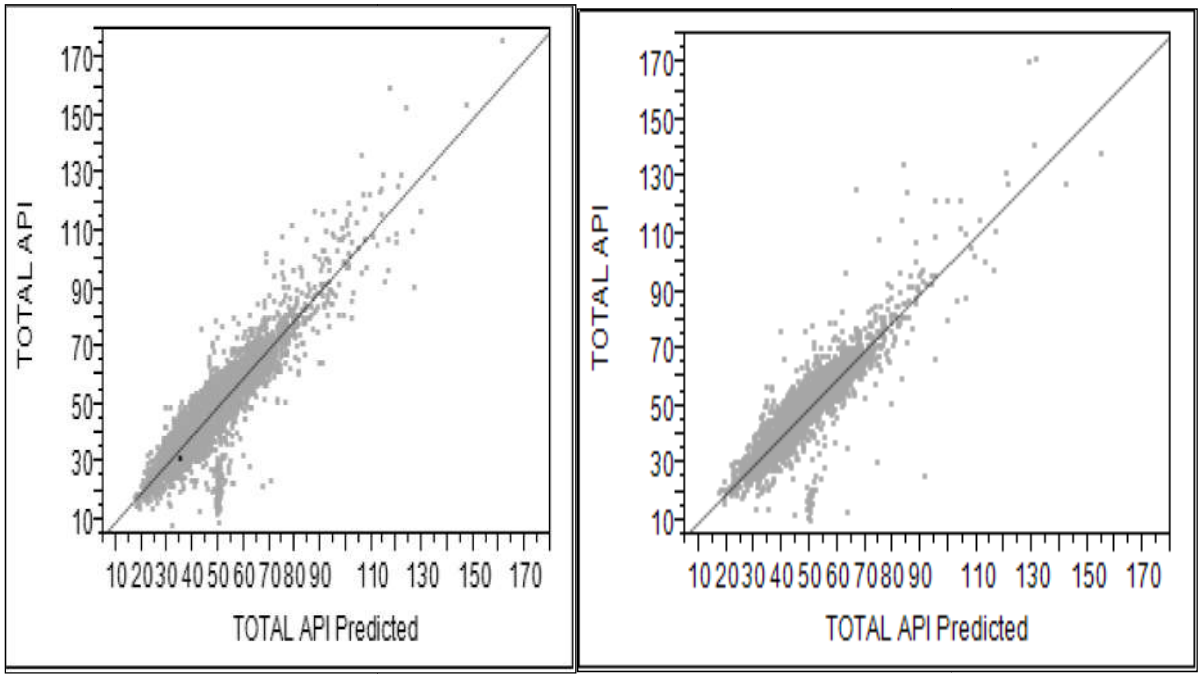

Fig.6. The prediction performance during (a) training and (b) validation phases

3.5. Predicting the API Data Using MLR

MLR model was developed to describe the behavior of the variables. It is based on a linear least-square fitting process and required a trace element or property to be determined for each source [11]. In this study, seven ambient air pollutant parameters and MLR were combined together in order to identify potential air pollution sources in the study area. Basically, MLR was applied in this study to explain the relationship between the source apportionment from seven variable parameters and their correlation to API values. Seven variables (independent variable) were chosen due to all DAs (standard, stepwise forward and stepwise backward) results shows acceptable to build an API (dependent variable).

Source apportionment can be used to estimate the contribution of identified sources to the concentrations of each parameter [12]. To evaluate the model performance, coefficient of determination (R^2) essentially be gathered [13]. However, R^2 values only provide information about how well it performs on external data. RMSE measure residue error in which it is given estimation of the mean difference between observed and modelled value of the API. The nearest R^2 value to one and the smallest value of RMSE, the better model shall be performed [14]. From the finding, the values of R^2 and RMSE were 0.773 and 7.172 respectively from the goodness of fit statistics (Table 4).

Table 4. Summary of regression of the API's variable

Goodness of Fit Statistics	
Observations	8760
Sum of weights	8760
DF	8752
R^2	0.773
Adjusted R^2	0.773
MSE	51.44
RMSE	7.172

Fig. 7 depicts the standardize coefficients of independent variable for linear regression model of the API. The finding shows that PM_{10} account as the highest pollution contributes to ambient air quality in the Southern Region of Peninsular Malaysia, while for the next main contributor was O_3 that may come from the anthropogenic activities. The third contributors were NO_x, NO and NO_2 which it may come from natural and anthropogenic activities along the vicinity area. Meanwhile, the SO_2 shows no or less influence to the API surrounding the study area. The negative standardized coefficient of the independent variable was CO, which owing to negative correlation to the API values.

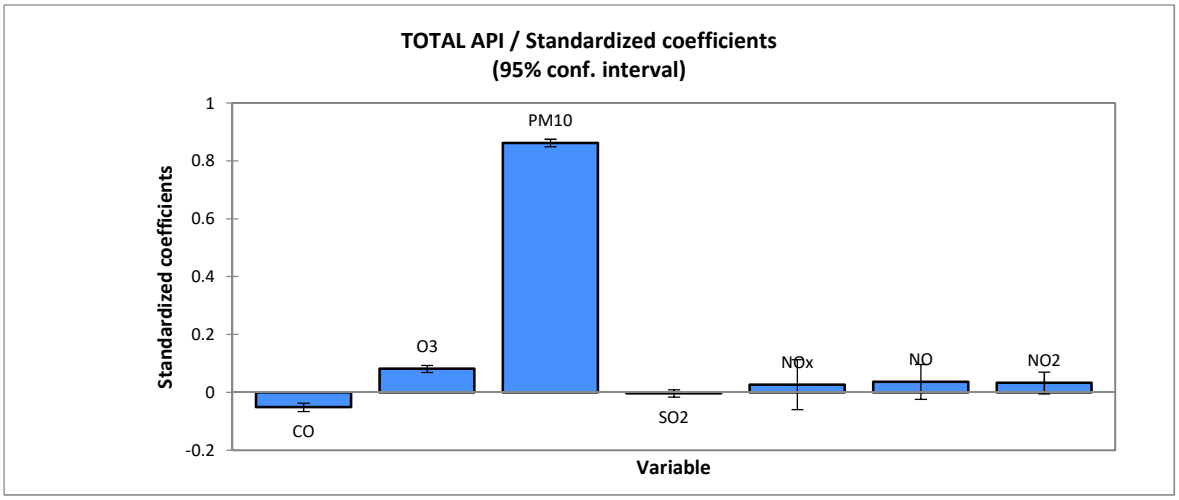

Fig.7. Bar chart of standardized coefficient for each

Fig. 8 shows the residual analysis of the observed and predicted API using the MLR model. The findings show the deficiency of the model, which the data set indicates a great difference in the range of -10 to 10. The verification of the model was influenced by the outer

observation as in Fig. 9. It reveals the graph of calculating API (predicted) versus API (actual), which from the actual API indicates that some of the observations were out of the lower and upper boundary range (95% of the confidence interval). The purpose of plotting the graph was proved that the MLR model able to use for the API prediction due to the great difference between predicted API and calculated API.

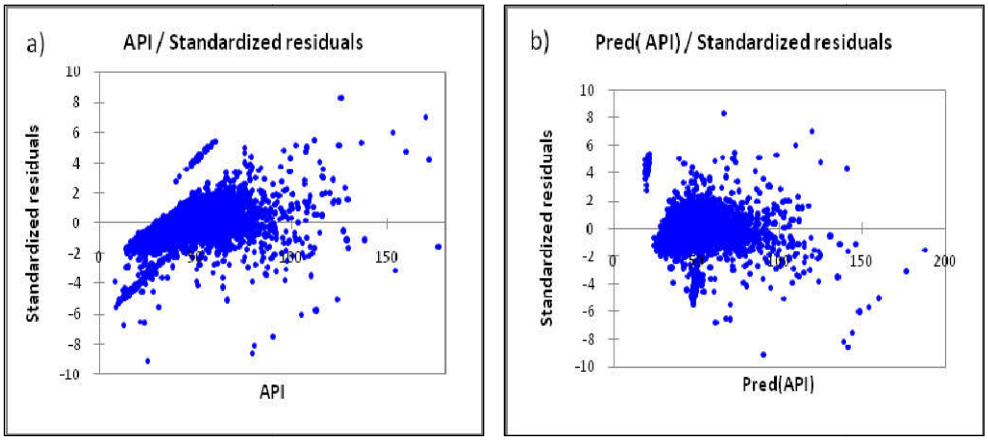

Fig.8. Scatter plot diagram of standardized residuals for: a) actual API, and b) predicted API

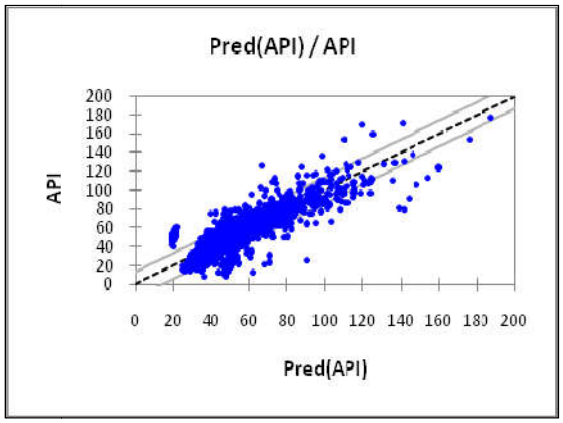

Fig.9. Scatter plot diagram of predicted air pollutant index versus actual air pollutant index

3.6. Comparison's Performances between ANN and MLR Models

Performance indicators (comprises R^2 and RMSE) for ANN and MLR were used to compare for the API prediction in the southern region of Peninsular Malaysia. Table 5 lists the performance indicator values. According to [15], the nearest value of R^2 to a value of 1 and the smallest value of RMSE indicates the model has a high correlation of variables. Based on the finding, the R^2 and RMSE for MLR indicate 0.773 and 7.172 respectively. While for ANN, the values were 0.819 and 6.424 respectively. Hence, it can be concluded that the ANN model should provide a better prediction than MLR in term of the API prediction. Fig. 10 and Fig. 11

show the comparison graphs of actual API, ANN and MLR.

Table 5. Performance indicator between ANN 9 and MLR models

Performance Indicator	ANN	MLR
R^2	0.819	0.773
RMSE	6.424	7.172

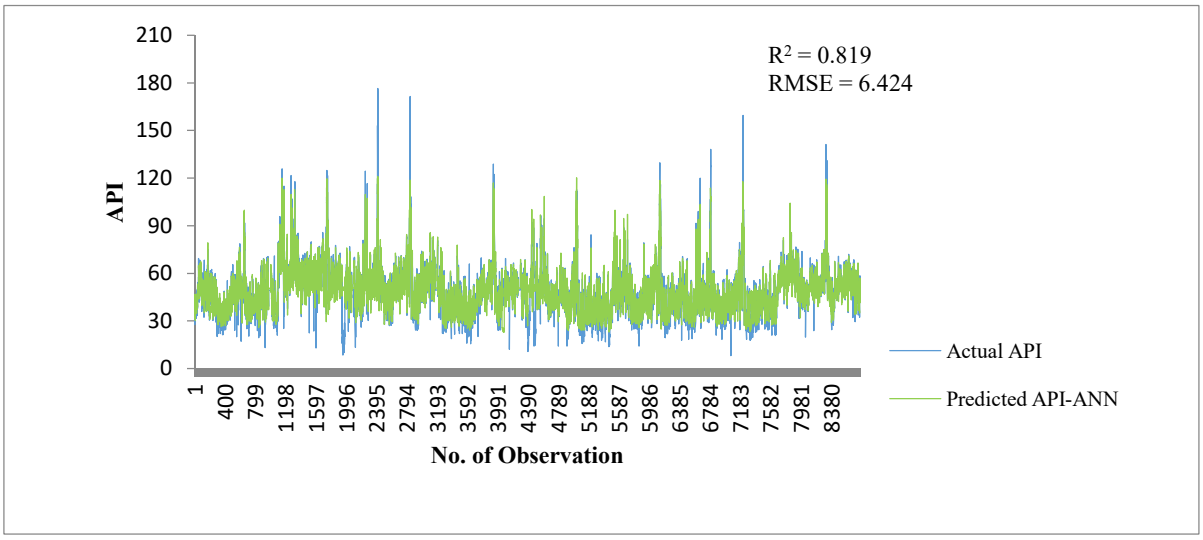

Fig.10. The results for: API-ANN together with their actual and predicted values

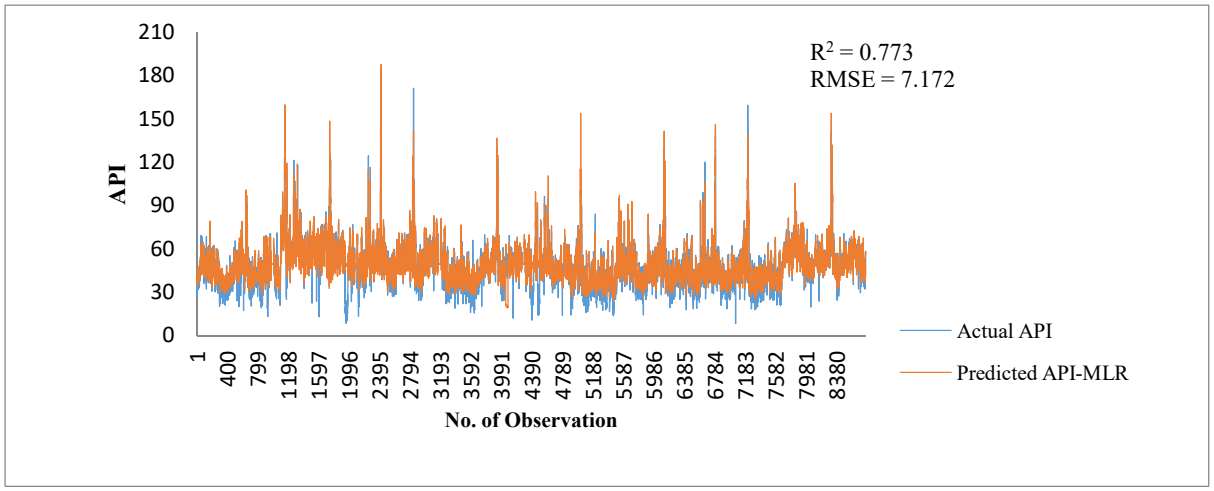

Fig.11. The results for: MLR together with their actual and predicted values

4. CONCLUSION

Air pollutant index monitoring programs generate multidimensional information that needs chemometric techniques for analysis and interpretation of data. Based on the conducted

analysis, HACA grouped the study area (eight stations) into three clusters and this information can be used for reducing the number of sampling sites without missing important information. DA used for the data reduction. The findings of DA in this study provide information that all parameters (CO, O_3, PM_{10}, SO_2, NO_x, NO and NO_2) are the most significant variables and gave 94.4% correct (after stepwise backward DA mode). For PCA analysis, this method helped in identifying the sources or factors that responsible for API variations (which are mainly related to the combustion of fossil fuels, motor vehicle exhaust and natural sources such as volcanoes activity from neighbouring countries) in three different regions. Subsequently, the receptor modelling in ANN and MLR of the API (API-ANN and API-MLR) provides apportionment by various sources in respective regions contributing to the air pollutants. The API-ANN model shows a better prediction compared to the API–MLR model. Thus, the application of chemometric techniques and modelling gave an excellent exploratory tool in API assessment, identification and apportionment of pollution sources and interpretation of complex dataset in order to understand their spatial variations in the study area. The information from this study can be setbacks for DOE or other public agencies in designing the monitoring network for effective management of air pollution resources and minimize time and cost without losing any important information.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] Azid A, Juahir H, Toriman M E, Kamarudin M K, Saudi A S, Hasnam C N, Aziz N A, Azaman F, Latif M T, Zainuddin S F, Osman M R. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. Water, Air, and Soil Pollution, 2014, 225(8)1-14

[2] Mutalib S N, Juahir H, Azid A, Sharif S M, Latif M T, Aris A Z, Zain S M, Dominick D. Spatial and temporal air quality pattern recognition using environmetric techniques: A case

study in Malaysia. Environmental Science: Processes and Impacts, 2013, 15(9):1717-1728

[3] Moustris K P, Ziomas I C, Paliatsos A G. 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, and Soil Pollution, 2010, 209(1-4):29-43

[4] Xie H, Ma F, Bai Q. Prediction of indoor air quality using artificial neural networks. In 5th International Conference on Natural Computation, 2009, pp. 412-418

[5] Mazlum N, Özer A, Mazlum S. Interpretation of water quality data by principal components analysis. Turkish Journal of Engineering and Environmental Sciences, 1999, 23(1):19-26

[6] Kamal M M, Jailani R, Shauri R L A. Prediction of ambient air quality based on neural network technique. In 4th Student Conference on Research and Development, 2006, pp. 115-119

[7] Karatzas K D, Kaltsatos S. Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. Simulation Modelling Practice and Theory, 2007, 15(10):1310-1319

[8] Motallebi N, Flocchini R G, Myrup L O, Cahill T A. A principal component analysis of visibility and air pollution in six California cities. Atmósfera, 1990, 3(2):127-141

[9] Levine J S, Augustsson T R, Andersont I C, Hoell Jr J M. Tropospheric sources of NO_x: Lighting and biology. Atmospheric Environment, 1984, 18(9):1797-1804

[10] Koike K, Nifuku M, Izumi K, Nakamura S, Fujiwara S, Horiguchi S. Explosion properties of highly concentrated ozone gas. Journal of Loss Preventive in the Process Industries, 2005, 18(4-6):465-468

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

[12] Simeonov V, Stratis J A, Samara C, Zachariadis G, Voutsa D, Anthemidis A, Sofoniou M, Kouimtzis T. Assessment of the surface water quality in Northern Greece. Water Research, 2003, 37(17):4119-4224

[13] Pearson K. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London, 1896, 187:253-318

[14] Pai T Y, Sung P J, Lin C Y, Leu H G, Shieh Y R, Chang S C, Chen S W, Jou J J. Predicting hourly ozone concentration in Dali area of Taichung Country based on multiple linear regression method. International Journal of Applied Science and Engineering, 2009, 7(2):127-132

[15] Aertsen W, Kinta V, Orshovena J, Özkan K, Muysa B. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. Ecological Modelling. 2010, 221(8):1119-1130

[16] Azid A, Juahir H, Toriman M E, Endut A, Kamarudin M K, Rahman M N, Hasnam C N, Saudi A S, Yunus K. Source apportionment of air pollution: A case study in Malaysia. Jurnal Teknologi. 2015, 72(1):83-88

[17] Massart D. L., Kaufman L. The interpretation of analytical data by the use of cluster analysis. New York: Wiley, 1983

[18] McKenna Jr J E. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environmental Modelling and Software, 2003, 18(2):205-220

[19] Willet P. Similarity and clustering in chemical information systems. New York: Wiley, 1987

[20] Adams M J. The principles of multivariate data analysis. In P. R. Ashurst, & M.J. Dennis (Eds.), Analytical methods of food authentication. London: Blackie Academic and Professional, 1998, pp. 308-336

[21] Kellner R, Mermet J M, Otto M, Widmer H M. Analytical chemistry. Weinheim: Wiley-VCH, 1998

[22] Juahir H, Zain S M, Yusoff M K, Hanidza T T, Armi A M, Toriman M E, Mokhtar M. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. Environmental Monitoring and Assessment, 2011, 173(1-4):625-641

[23] Forina M, Armanino C, Raggio V. Clustering with dendrograms on interpretation variables. Analytica Chimica Acta, 2002, 454(1):13-19

[24] Dominick D, Juahir H, Latif M T, Zain SM, Aris AZ. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmospheric Environment, 2012, 60:172-181

[25] Singh K P, Malik A, Mohan D, Sinha S. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)-A case study. Water Research, 2004, 38(18):3980-3992

[26] Singh K P, Malik A, Sinha S. Water quality assessment and apportionment of pollution sources of Gomti River (India) using multivariate statistical techniques-A case study. Analytica Chimica Acta, 2005, 538(1):355-374.

[27] Shrestha S, Kazama F. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin. Japan. Environmental Modelling and Software, 2007, 22(4):464-475.

[28] Kannel P R, Lee S, Kanel S R, Khan S P. Chemometric application in classification and assessment of monitoring locations of an urban river system. Analytica Chimica Acta, 2007 582(2):390-399

[29] Kim J. O., Mueller C. W. Introduction to factor analysis: What it is and how to do it. Quantitative applications in the social science series. California: Sage Publications, 1987

[30] Vega M, Pardo R, Barrato E, Deban L. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Research, 1998, 32(12):3581-3592

[31] Liu C W, Lin K H, Kuo Y M. Application of factor analysis in the assessment of groundwater quality in a black foot disease area in Taiwan. Science of the Total Environment, 2003, 313(1):77-89

[32] Chaloulakou A, Grivas G, Spyrellis N. Neural network and multiple regression model for PM10 prediction in Athens: A comparative assessment. Journal of the Air and Waste Management Association, 2003, 53(10):1183-1190

[33] Nasir M F M, Juahir H, Roslan N, Mohd I, Shafie N A, Ramli N. Artificial neural networks combined with sensitivity analysis as a prediction model for water quality index in Juru River, Malaysia. International Journal of Environmental Protection, 2011, 1(3):1-8

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

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

How to cite this article:
Azid A, Rani NAA, Samsudin MS, Khalit SI, Gasim MB, Kamarudin MKA Yunus K, Saudi ASM, Yusof KMKK.Air quality modeling using chemometric techniques. J. Fundam. Appl. Sci., 2017, 9*(2S), 443-466.*