Application of Principal Component Analysis and Multiple Linear regression for Air Pollution Modeling in Selected Monitoring Stations in Malaysia

Hamza Ahmad Isiyaka, Ekhwan Mohd Toriman and Hafizan Juahir
East Coast Environmental Research Institute (ESERI), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu Malaysia.

Key Words
Multiple linear regression, Principal component analysis, Box and whisker plots, Air pollution modeling, Geographical information system.

ABSTRACT
This study was carried out to identify the major sources of air pollution and determine the level at which each pollutants contributes to the air pollution index (API). Principal component analysis attributes the pollution source to anthropogenic induced emission (industries, power plants, motor vehicle) accounting for more than 38% and 20% of the total variance. Multiple linear regression (MLR) was used to develop an explicit equation with less complexity and identify the level at which each pollutant contributes to the air pollution index. MLR, box and whisker plots revealed that PM$_{10}$, SO$_2$ and NO$_2$ are the most significant parameters affecting the level of pollution. Geographical information system (GIS) was also applied to map out the air quality monitoring sites within the study area. This study simplifies the complexity and dynamic nature of air pollution by providing fundamental information for decision making and policy implementation by stakeholders.

© 2015 Elixir All rights reserved.

Introduction
Atmospheric air pollution is a fundamental challenge that requires an urgent attention due to its ability to alter the general ecosystem balance and discomfort biotic and abiotic components [1, 2]. Despite the effort put by many stakeholders to control and limit the level of pollutant emission, its complexity from point and non-point source make it difficult to curb [1]. [3] suggested that the correlation between the sources and resulting pollutants are not instantaneous. However, it has been identified that rapid industrial expansion, vehicular emission, an increase in urban population and power generation are the major factors contributing to urban air pollution [4].

Air is a composition of gasses held by the force of gravity that act as the spacesuit of the biosphere [5]. Its ability to reduce temperature extremes, conserve heat, the source of oxygen, carbon dioxide and absorption of ultraviolet solar radiation render it a basic necessity for existence [6, 7]. Air pollution is a mixture of primary particles emitted from different sources and secondary particles from aerosols formed by chemical reactions [8, 9]. The atmospheric condition of an area is polluted when gasses and aerosols accumulate in a high concentration above the threshold set as a standard [10].

Lately, research conducted by [11, 12] revealed that motor vehicle emission contributes about 82% of the total pollution load in Malaysia. Other pollutants are emitted from a stationary source (industrial activities, power stations, and construction sites) and transboundary pollutants especially during the Sumatra bush burning in Indonesia with the worst episode in 1997 [13, 14]. A report by the department of environment shows that the air quality status in Malaysia based on air pollution index (API) fall within the range of good and moderate from 2008 to 2011. In 2008, the status of good air quality fall around 59%, 55.6% in 2009, 63% in 2010 and 55% in 2011 [13].

However, the rapid expansion in the economic settings of Malaysia in terms of industrial activities, construction of modern city centers, expansion in mechanized system of agriculture and tourism have affected the quality of air in both micro and macro scale [14, 15].

Furthermore, an understanding of the regional variability in the characteristics of air pollutants requires an integration of samples from monitoring sites and robust statistical techniques [17]. Application of PCA and MLR model have been widely used by researchers [2, 14, 18, 19, 20, 21, 22, 23, 24] to model the dynamic characteristic of environmental pollution. This technique can reduce the dimensionality of large data sets, identify the major pollution source and model the percentage contribution of individual pollutants [18, 24].

The objectives of this study are to understand the complex source of air pollution, determine the contribution of each pollutant and develop an explicit equation model with the low level of complexity.

Materials and Methods
Study area
The study area comprises of five air quality monitoring sites under the supervision of the Department of Environment Malaysia (DOE). The spatial location of these monitoring sites can provide a synoptic view of the air quality characteristics within the study area.

These stations comprises of Pasir Gudang, Johor (ST01) located in the southern Peninsular Malaysia; Kemaman, Terengganu (ST02) is situated in the eastern Peninsular Malaysia; Perai, Pulau Pinang (ST03) in the Northern Peninsular Malaysia; Sibiu Sarawak (4) and Tawu Sabah (ST05) are situated in the eastern Malaysia. The location of the sampling sites were mapped out using the ArcGIS version 10.22 software comprising of the latitude and longitude of each sampling station in Fig. 1 and Table 1. GIS is a computer-based mapping and information retrieval system based on a unique geographical location. Air quality parameters (carbon monoxide (CO), ozone (O$_3$), particulate matter (PM$_{10}$), sulfur dioxide (SO$_2$) and nitrogen dioxide (NO$_2$)) observed over a period of five years (2003-2007) were sourced from DOE in the form of hourly
reading. The data were converted to daily observations, and all missing values were estimated using the nearest neighbor in the XLSTAT 2014 add-in software environment. The total number of missing values was around 3%. The nearest neighbor allows an estimation of unknown values using the known values at neighboring locations [12]. Equation 1 can be used to estimate missing values:

\[
y = y_i; \text{ if } x \leq x_1 + \left( \frac{x_2 - x_1}{2} \right) \\
y = y_1; \text{ if } x \geq x_1 + \left( \frac{x_2 - x_1}{2} \right)
\]  

Where \( y \) represents the interpolate, \( x \) is the time point of the interpolate, \( y_i \) and \( x_i \) are the coordinates of the starting point of the gap and \( y_2 \) and \( x_2 \) are the end points of the gaps.

Although, the entire analysis in this study were done using the XLSTAT 2014 add-in software. This software can handle flexible complex data sets; its application is user-friendly and very flexible. The software have been explored by several researchers to model different environmental issues ranging from water pollution to atmospheric air modeling [12,14,21,25].

![Figure 1. Location of monitoring site using ArcGIS](image)

**Principal component analysis (PCA)**

Ambient air pollution monitoring usually generates huge data, and PCA can reduce the dimensionality of observed parameters with minimal loss of the original variables [25]. This can be achieved by an unsupervised elimination of the less significant observations and retaining the most important observations called principal components [23,26]. PCA is a pattern recognition technique used to identify the possible source of pollution [12].

Furthermore, in order to provide a simple and accurate representation of the source of pollution, it is imperative to rotate the principal components by varimax rotation with an eigenvalue greater than one [25; 27]. The varimax rotation produces new groups of variables called varimax factors (VFs) that are equal to the numbers of parameters explaining the source of pollution, only factor loading with a high variability can be used as a base for interpretation.

This accounted for more than 38% (VF1) and 20% (VF2) of the total variance in the data set as displayed in the PCA loading plot in Fig. 2.

However, in order to identify the most significant parameters explaining the source of pollution, only factor loading greater than 0.70 were selected for interpretation. This is because strong factors represent strong positive loading that can influence the air quality characteristics of an area. According to [4] factor loading with a high variability can be used as a base for source identification while those factors from 0.5-0.3 represent a weaker influence on the source of air pollution. The scree plot diagram explaining the cut-off point for the eigenvalue is also displayed in Fig. 3.

**Multiple linear regression model (MLR)**

MLR is a modeling technique that can predict the variability that exists between a dependent and an independent variable [29,30]. It is used to form explicit equations that are less complex [14]. The model is embedded with \( k \) independent variables and \( n \) observation. Thus, the regress model can be represented as [31]:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i \quad \text{Eq} \quad (3)
\]

Where \( i = 1, \ldots, n, \beta_0, \beta_1 \) and \( \beta_k \) are regression coefficient, \( x_j \) and \( x_k \) are independent variables and \( \epsilon \) is error associated with the regression.

Using this method, the contribution of each parameter were predicted base on the coefficient of determination (R²), adjusted coefficient of determination (R²) and Root mean square error (RMSE) [32]. To achieve this, the parameters were introduced to the linear model as independent variables with API as the dependent variables [14].

The XLSTAT 2014 add-in software was applied in order to understand the underlying statistical composition of each observed parameter in the entire data sets as described in Table 2 below.

**Results and Discussion**

**Descriptive statistics**

The descriptive statistics of the observed parameters comprising of the total number of observation, minimum and maximum values, media, mean, variance and standard deviation are shown in Table 2.

**Principal component analysis for source identification**

PCA was applied in this study in order to understand the pattern and possible sources of pollution based on the activities within the study area. Two varifactors were obtained for the varimax rotation with an eigenvalue greater than one. This accounted for more than 38% (VF1) and 20% (VF2) of the total variance in the data set as described in the PCA loading plot in Fig. 2.

![Figure 2. PCA loading plot after varimax rotation](image)
Figure 3. Scree plot diagram for PCA loading

Furthermore, a description of the factor loading, eigenvalue, cumulative percentage and total variance for the entire datasets is shown in Table 3. The factor loadings highlighted in bold represent those with a high level of variability in the source of pollution.

In Table 3, the first varimax factor (VF1) has a strong positive loading for PM$_{10}$ (0.809), SO$_2$ (0.730) and NO$_2$ (0.817). The source of PM$_{10}$ can be attributed to industrial and construction activities, vehicular emission, soil dust resuspension and bush burning within the study area [33,34]. A report by the Malaysia Ministry of Transport [35] revealed that the number of registered vehicles in Malaysia is on a rapid increase from 934,367 (4.42%) in 2004 to 1,160,082 as at 2010. This increase contributes immensely to the emission of PM$_{10}$ within the study area. PM$_{10}$ can also be produced by photochemical oxidation of its precursors as secondary pollutants under favorable atmospheric condition [36]. SO$_2$ is produced by industrial emission, coal fire plants, as well as emission from heavy diesel engines, buses, and lorries [14,37].

NO$_2$ is mostly produced from fossil fuel burning especially in industries and reaction of nitrogen with oxygen in the air at a very high temperature [6]. A study conducted by [38] indicates that about 69% of NO$_2$ is emitted from power stations and industrial activities while motor vehicles emission account for 28%, and the remaining 3% makes up other sources.

VF2 has a strong positive loading for O$_3$, which is the primary component of smog formed by photochemical oxidation of its precursor (NOx CO and VOCs) [39]. According to [39], a high concentration of O$_3$ exceeding 40ppbv can limit the life span and stress tolerance of forest trees and crops. Its primary source of origin is from motor vehicle emission in high traffic areas.

**Air pollution modeling using multiple linear regression**

An explicit equation with a small level of complexity was developed to explain the level of pollution. The standardized coefficient that represents the individual contribution of pollutants is described in Fig. 4. However, the box and whisker plots in Fig. 6 were used to ascertain the contribution of individual pollutants to the level of the air pollution index within the study area. Furthermore, the equation developed for the model comprising of each pollutant, the coefficient of determination ($R^2$) and the root mean square error (RMSE) can be written in equation (4) as:

$$\text{API} = 7.8504 \times \text{(CO)} + 0.4391 + \text{(O3)} + 0.5389 \times \text{(PM10)} + 496.300 \times \text{(SO2)} + 50.8519 \times \text{(NO2)} + 16.6158$$

$$R^2 = 0.741, \quad \text{RMSE} = 7.924 \ (4)$$

From the above equation (4) it is glaring that all the observed parameters have a positive influence on the API value with an $R^2 = 74\%$ and RMSE = 7.9. Although PM$_{10}$ appears to have more significant impact followed by SO$_2$ compared with other parameters. The high concentration of PM$_{10}$ and SO$_2$ can be linked with the industrial activities, power plants, emission from vehicles, construction sites and resuspension of soil dust [14,32]. Extensive petrochemical activities and Liquefied natural gas exploration by PETRONAS can be found in Kertih and Sarawak. Johor is known for its extensive industrial activities spanning from primary production to more advanced industries. Heavy Traffic can also be found in Kuala Terengganu, Johor and other city centers of the study area [15].

Figure 4. Standardized coefficient for individual Contribution of pollutants

The scatter plot represented in Fig. 5 explains the prediction performance of the MLR model. It provides a visual representation of the relationship between the response and predictor variables.

Figure 5. Scatter plot for predicted API
Atmospheric air pollution exhibits a complex and dynamic characteristics induced by anthropogenic activities from point source and non-point sources. PCA and MLR were applied to simplify these complexity and provide fundamental information for decision making and policy implementation by stakeholders. In this regard, the source of pollution within the study area were identified using PCA that account for more than 38% and 20% of the total variance in the data sets. The sources of pollution were attributed to anthropogenic induced activities (Industrial activities, power plants, motor vehicle emission, construction sites, biomass burning). MLR model was used to predict the level of pollution to the tune of 74% as well as develop an explicit equation model with less complexity. Based on MLR, box and whisker plots, PM$_{10}$, SO$_2$ and NO$_2$ are the most influencing parameters within the study area. GIS was also used to map out the location of each monitoring stations in order to provide a spatial representation of monitoring sites in the environment.

### Acknowledgement

The authors wish to express their special thanks to the East Coast Environmental Research Institutes (ESERI) for their consistent and constructive advices during this work. Our profound gratitude goes to the Department of Environment Malaysia (DOE) for providing all the data needed for this study.

### References


