FULL TITLE

MULTIVARIATE ANALYSIS OF MONSOON SEASONAL VARIATION AND PREDICTION OF PARTICULATE MATTER EPISODE USING REGRESSION AND HYBRID MODELS

SHORT TITLE

MULTIVARIATE ANALYSIS OF PARTICULATE MATTER CONCENTRATION EPISODE

AMINA NAZIF*, NURUL IZMA MOHAMMED, AMIRHOSSEIN MALAKAHMAD, AND MOTASEM S. ABUALQUMBOZ

Department of Civil and Environmental Engineering, Universiti Teknologi PETRONAS,
32610 Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia.
*aminanazif@yahoo.co.uk

Abstract

Prediction of Particulate Matter (PM_{10}) episode in advance enables for better preparation to avert and reduce the impact of air pollution ahead of time. This is possible with proper understanding of air pollutants and the parameters that influence its pattern. Hence, this study analyzed daily average PM_{10}, temperature (T), humidity (H), wind speed (WS) and wind direction (WD) data for five years (2006-2010), from two industrial air quality monitoring stations. This data was used to evaluate the impact of meteorological parameters and PM_{10} in two peculiar seasons; Southwest Monsoon (SWM) and Northeast Monsoon (NEM) seasons, using Principal Component Analysis (PCA). Subsequently, Lognormal Regression (LR), Multiple Linear Regression (MLR) and Principal Component Regression (PCR) methods were used to forecast next day average PM_{10} concentration level. The PCA result (seasonal variability) showed that peculiar relationship exist between PM_{10} pollutants and meteorological parameters. For the prediction models, the three methods gave significant results in terms of performance indicators. However, PCR had better predictability, having a higher coefficient of determination (R^2) and better performance indicator results than LR and MLR methods. The outcomes of study signify that PCR models can be effectively used as a suitable format in predicting next day average PM_{10} concentration levels.

Keywords; Air pollution; Meteorology; Prediction; Regression

Introduction

Particulate matter with aerodynamic diameter of 10μ and below (PM_{10}) also known as coarse particles, is an air pollutant has characteristics of being both a primary and a secondary pollutant (Harrison et al, 2012). Major sources of anthropogenic PM_{10} are industrial and traffic-related activities (Hörmann et al, 2005; Ul-Saufie et al, 2012b). Additionally, seasonal variation is associated with PM_{10} concentration and can also influence particulate matter (PM) concentration patterns (Kovač-Andrić et al, 2009). In most countries, seasonal variation such as cold and warm seasons tend to affect PM pollution patterns (Kassomenos et al, 2014; Vardoulakis et al, 2008). However, monsoonal season is normally observed in tropical rainforest climate regions of the world. These regions have no distinct cold and warm seasons. The seasonal variation is usually based on windblown patterns and rainfall seasons (Abdullah et al, 2011).

Malaysia is a country with tropical climate features, having seasons of northeast monsoon (November to March), southwest monsoon (May to September), first inter-monsoon (April) and second inter monsoon (October) (DoE,

2010) . The northeast monsoon is usually connected to the wet season and it is linked with long-range transport of air mass from the east coast of Indo-China (Latif et al, 2012). The northeast monsoon season is associated with low PM_{10} concentration levels due to increase in rainfall. It was established that wash out process by rainfall decreases atmospheric aerosols in the atmosphere, leading to lower PM_{10} concentration levels (Azmi et al, 2010). Meanwhile, southwest monsoon season is associated with low rainfall and it is regarded as dry season. In this season, activities such as forest fires and bush burning result in increasing PM_{10} concentration levels.

PM_{10} has enormous health consequences (Ebi et al, 2008; Katsouyanni et al, 2009). Substantially, PM_{10} is associated with internal ailments particularly, lung and cardiovascular diseases, depression, asthma and in extreme cases death (Namdeo et al, 2005; Schwartz, 2001). Therefore, in order to curtail PM_{10} effect, there is the need to predict PM_{10} episode in advance, for better preparedness and impact reduction.

Different statistical analyses have been used previously in forecasting future PM_{10} episode including Artificial Neural Network (ANN) (Nejadkoorki et al, 2012), Quantile Regression (QR) (Ul-Saufie et al, 2012a), Lognormal (Yusof et al, 2010) and Stepwise Regression (SR) methods (Taşpınar et al, 2014). Particularly, Multiple Linear Regressions (MLR) has been used in predicting next day PM_{10} average concentration (Afzali et al, 2014; Slini et al, 2006). MLR has been used in predicting maximum PM_{10} concentrations levels in Turkey using SR (Taşpınar et al, 2014). Hybrid models have been introduced to improve the results of several prediction methods (Taşpınar, 2015). Hybrid models involve using multivariate analysis to improve regression models. Principal component analysis (PCA) has been used previously with ANN (Taşpınar, 2015) and MLR models (Ul-Saufie et al, 2013) to improve models predictability and give better result, but majority of these models need to be tested for multicollinearity and their ability to detect high PM_{10} concentration episode.

The present study was carriedout using meteorological parameters (temperature, humidity, wind speed, and wind direction) and PM_{10} data (January to December from 2006 to 2010) from two industrial air quality monitoring stations in Malaysia. The main aim was to assess prediction capabilities of lognormal regression (LR), multiple linear regression (MLR) and principal component regression (PCR) methods in predicting next day average PM_{10} concentration levels. Also, ability of the models to predict high PM_{10} concentration levels was assessed. Prior to the prediction analyses, the impact of meteorological parameters and PM_{10} in monsoon season was examined using PCA.

Materials and methods

2.1 Study areas

The industrial air quality monitoring stations for this study were located in two different states in Malaysia. These areas comprise of industries and other urban settings. There has been a steady industrial growth in those areas with substantial increase in industrial air pollution sources between 2006 and 2010 (DoE, 2010).

Sarawak

Sarawak is a state located in east Malaysia, with its capital as Kuching. It is positioned at latitude 3.0381° N and longitude 113.7811° E. Kuching city has a population of 325,132 and the metro population is about 684,122. Sarawak has an area coverage of 124,450sq.km, having a population of about 2.4 million people (Department of

Statistics, 2010). With an annual average rainfall of 4,200mm, the average temperature ranges between 19°C to 36°C, average wind speed of 13km/hr and relative humidity ranges from 71% to 97%. The industrial air quality monitoring station in this state is located in Kuching.

Pahang

Pahang is a state in peninsular Malaysia and its state capital is Kuantan, located at latitude 3.7500° N and longitude 102.5000° E, with an area of 36,137sq km and a population of 1.4 million people (Department of Statistics, 2010). Pahang has a temperature range between 23°C to 32°C, average wind speed of 11km/hr, average relative humidity ranges from 71 % to 95 % with an annual rainfall ranging between 2000 and 3000mm. The industrial air quality monitoring station is at Balok Baru.

2.2 Monitoring Records

Daily average PM_{10}, temperature, humidity, wind speed and wind direction data for five years (2006-2010) was used for this study. The data was acquired from the Department of Environment (DoE) Malaysia. For the seasonal analysis, the data was divided into two respective seasons; Southwest Monsoon (SWM) and Northeast Monsoon (NEM) seasons, respectively. The SWM data was from May to September, while NEM was from November to March.

2.3 Methods

2.3.3 Principal Component Analysis (PCA)

PCA reduces large amount of data to principal components which are usually equal to or less than the original data set. PCA maximizes the correlation between the original data set to form data sets that are orthogonal in nature (Abdul-Wahab et al, 2005). Additionally, PCA allows for the identification and observation of variations in the data set. The PCA analysis is shown in Equation 1.

$$PC_i = l_{1i}X_1 + l_{2i}X_2 \ldots . + l_{ni}X_n \quad (1)$$

where, PC_i is the i^{th} principal component and l_{ji} is the loading of the observed variable X_j.

For this analysis, PCA was applied to assess the influence and relationship of meteorological parameters and PM_{10} pollutant in SWM and NEM seasons. Furthermore, PCA was used to produce the PCR models. Daily average PM_{10}, temperature, humidity, wind speed, and wind direction data were used. The PCA was subjected to Varimax rotation analysis to display a well explained result. The principal components (PC) with Eigenvalues of >0.9 and correlation with >0.5 were assigned to be significant and were used for the analyses. Meanwhile, PC1 to 3 were named as S1to 3 for PC's in SWM season and N1 to 3 for PC's in NEM season.

2.3.4 Hybrid Model

Hybrid model in the form of PCR was developed by combining PCA and MLR methods. This would aid in reducing complexity of the models and decrease multicollinearity. Figure 1 shows architecture of the hybrid model.

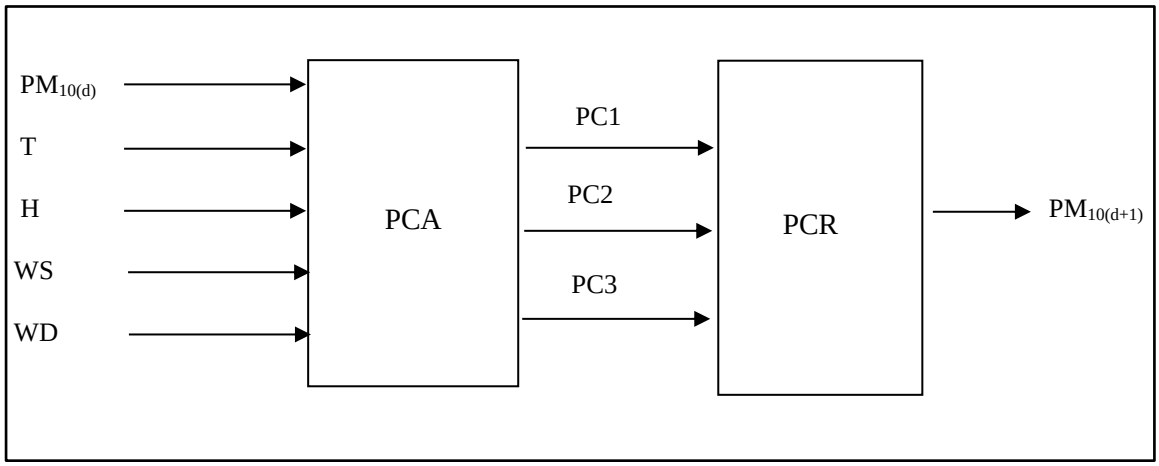

Figure 1 Architecture of Hybrid Model

2.3.1 Lognormal Regression (LR)

The lognormal analysis is a continuous probability distribution of variables, where the logarithm is normally distributed. Hence, if the random variable, X, is log-normally distributed, then Y= In(X). Similarly, if $Y$is normally distributed, then X= exp(Y). It should be noted that a random variable which is log-normally distributed is always in positive real values (Johnson et al, 1994). LR analysis was carried out using Equation 2.

$$Log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \quad (2)$$

where, Log(Y) is the dependent variable, β_0 is the constant coefficient, β_1, β_2, …… β_p are the regression coefficients of the independent variables X_1, X_2,…X_p and ε is the residual error .

2.3.2 Multiple Linear Regression (MLR)

Basically, MLR is used in determining the relationship between a predictant and more than one predictor variables. MLR analysis was carried out using Equation 3.

$$y = b_o + \sum_{i=1}^{n} b_i X_i + \varepsilon \quad (3)$$

where, b_o is the constant, b_i is the regression coefficient, X_i is the independent variables and ε is the error.

All prediction analysis were carried out using daily average PM_{10} (PM_{10}), temperature (T), humidity (H) wind speed (WS) and wind direction (WD). Additionally, previous day PM_{10} ($PM_{10(d-1)}$) data was added to the LR and MLR analysis to give a better model prediction (Afzali et al, 2014; Ul-Saufie et al, 2011).

2.4 Performance indicators

To test model performance and accuracy, several statistical descriptors were used as performance indicators. These indicators include the coefficient of determination (R^2), Adjusted R (Adj R) and Variance Inflation Factor (VIF). Also, Index of Agreement (IA), Prediction Accuracy (PA), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Normalized Absolute Error (NAE) were used based on their applications in previous studies (Hamida et al, 2012; Lu, 2004; Nejadkoorki et al, 2012; Yusof et al, 2010).

Table 1: Performance indicators

Performance Measure	Equation	Description
Coefficient of Determination (R^2)	$R^2= \left[\frac{\sum_{i=1}^{n}(O_i-\bar{O})^2.(P_i-\bar{P})}{n.\sigma_0.\sigma_p}\right]^2$	The R^2 is used to indicate how well model results fit the original data points. It also indicates the similarity between the modelled and the observed concentrations. The R^2 analysis result ranges from 0 to 1, the result with values ≥0.5 is regarded as significant.
Root Mean Square Error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}$	RMSE explains the precision of the model by summarizing the difference between the observed and modelled PM_{10} concentrations. It focuses on assessing the level of errors in the model. Models with low RMSE results are regarded as substantial when compared with other forecasting methods or when compared with other performance indicators.
Mean Absolute Error (MAE)	$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\lvert P_i - O_i\rvert}{n}$	MAE is used to assess the amount of error in a prediction model. When the MAE value is low as compared with other indicators, this indicates a better prediction method. The model with low MAE result as compared to other models is regarded as being significant having fewer errors than the other forecasting models.
Normalized Absolute Error (NAE)	$\mathrm{NAE} = \frac{\sum_{i=1}^{n}\lvert P_i - O_i\rvert}{\sum_{i=1}^{n} O_i}$	NAE is also used to assess error. Results approaching zero indicates a better prediction model. The NAE result is in the range of 0 to 1.
Prediction Accuracy (PA)	$\mathrm{PA}= \frac{\sum_{i=1}^{n}(P_i-\bar{O})^2}{\sum_{i=1}^{n}(O_i-\bar{O})^2}$	PA is used to assess the accuracy of prediction method. The PA results ranges from 0 to 1. When a PA result is closer to 1, it indicates that the method is better for prediction as compared to other methods and performance measures.
Index of Agreement (IA)	$\mathrm{IA} = 1 - \left[\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(\lvert P_i - \bar{O}\rvert - \lvert O_i - \bar{O}\rvert)^2}\right]$	IA is carried out to show the accuracy of the prediction model. The IA result is usually in the range of 0 to 1. The IA result is appropriate when the result is closer to 1.
Variance Inflation Factor (VIF)	$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$	VIF assesses the effect of multicollinearity in a prediction model. It provides a format that measures how much the variance of an estimated regression coefficient is increased because of collinearity. The model having VIF result ≥10 is regarded as having no multicollinearity effect as established by previous studies (Ul-Saufie et al, 2013; Ul-Saufie et al, 2011).

Table 1 contains performance measures, their definitions and equations where n is number of data, O_i is the observed data, $\bar{O}$ is the observed mean, P_i is the predicted data and $\bar{P}$ is the predicted mean, σ_p is the standard

deviation of the predicted values, and σ_o is the standard deviation of the observed values. VIF_i is the variance inflation factor, while, R_i^2 is the coefficient of determination in a regression of the ith predictor on all predictors.

Results and discussion

Descriptive Statistics

The descriptive analysis for Kuching and Balok Baru is shown in Figure 2 and Table 2. Result showed that Year 2006 and 2009 had high daily average PM_{10} concentration levels of 316µg/m^3 and 173µg/m^3 respectively. These levels were higher than the World Health Organization (WHO) guideline of 50µg/m^3 and Malaysian Ambient Air Quality Guideline (MAAQG) of 150µg/m^3. Subsequently, years 2007, 2008, and 2010 had maximum levels at 96µg/m^3, 84µg/m^3, and 53µg/m^3, respectively. All years had minimum levels lower than the MAAQG daily average concentration. Meanwhile, for Balok Baru area years 2006, 2009, and 2010 had maximum concentration levels of 196µg/m^3, 158µg/m^3, and 203µg/m^3, respectively. These values are higher than the WHO guideline limit and the MAAQG of average daily PM_{10} concentration levels. The year 2007 and 2008 had concentration levels at 119µg/m^3 and 131µg/m^3. All minimum PM_{10} concentration levels were lower than the MAAQG.

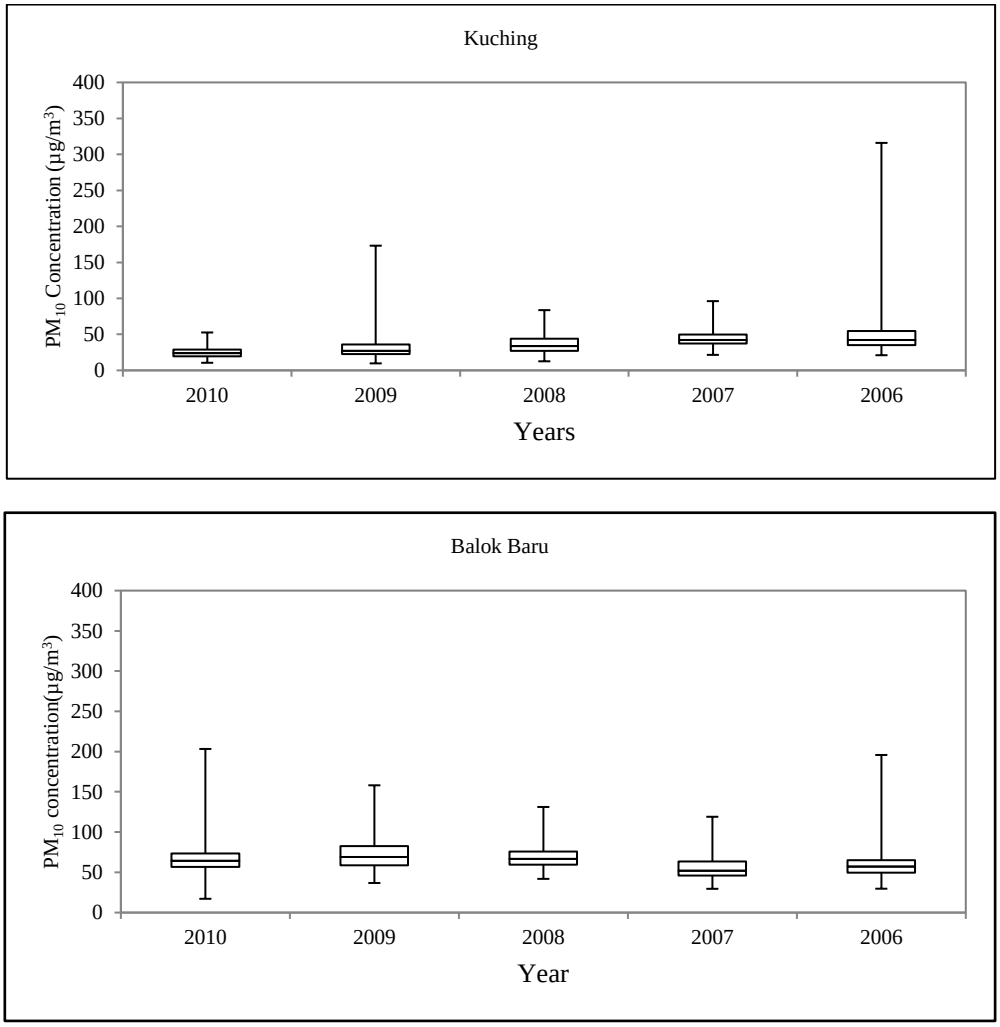

Figure 2 Descriptive Statistics of Kuching and Balok Baru (2006 to 2010)

Table 2: Descriptive statistics of Kuching and Balok Baru from 2006 to 2010

Kuching	PM_{10} ($\mu g/m^3$)	Temperature (^oC)	Humidity (%)	Wind speed (km/hr)	Wind direction (^{o})
Mean	41.77	26.37	82.97	5.12	143.37
Standard Error	0.68	0.04	0.13	0.03	1.75
Median	37.30	26.35	83.25	4.99	151.15
Mode	22.83	26.51	79.88	4.39	202.96
Standard deviation	24.86	1.37	4.77	1.15	64.20
Sample variance	618.20	1.87	22.71	1.32	4121.03
Kurtosis	27.47	-0.02	0.55	6.96	-0.45
Skewness	4.09	-0.07	-0.33	1.55	-0.30
Range	306.30	8.89	31.71	12.32	344.33
Minimum	9.70	21.94	64.5	2.21	4.21
Maximum	316	30.83	96.21	14.53	348.54
Sum	55971.48	35335.08	111184.5	6854.22	192117.5
Count	1340	1340	1340	1340	1340
Balok Baru					
Mean	64.02	26.62	83.89	7.86	183.28
Standard Error	0.49	0.03	0.12	0.06	1.46
Median	60.94	26.75	83.88	7.46	199.88
Mode	67.83	26.52	81.04	7.27	198.29
Standard deviation	18.19	1.14	4.48	2.10	53.92
Sample variance	330.84	1.30	20.05	4.39	2907.43
Kurtosis	5.93	0.88	0.68	2.30	0.87
Skewness	1.63	-0.55	0.23	1.26	-1.14
Range	166.29	9.70	31.83	15.22	303.04
Minimum	29.5	21.83	68.13	3.75	11.54
Maximum	195.79	31.53	99.96	18.98	314.58
Sum	86933.10	36156.07	113917.7	10678.76	248883.5
Count	1358	1358	1358	1358	1358

Time Series Plot

Figure 3 shows the trend analysis for Kuching. For the year 2006, nine violations were recorded for the MAAQG. The highest concentration levels was 316$\mu g/m^3$, others were 171 $\mu g/m^3$, 187 $\mu g/m^3$, 242 $\mu g/m^3$, 250 $\mu g/m^3$ and 168 $\mu g/m^3$. Additionally, 2009 recorded one violation (173$\mu g/m^3$). Apart from industrial activities, this area is hugely affected by haze episodes. This is the reason for very high concentration levels in some years and lower concentration levels in others. There were records of haze in 2006, 2009, and 2010 (DoE, 2006; 2009; 2010; Mutalib et al, 2013), while there were no records of haze pollution in 2007 and 2008 (DoE, 2007; 2008; Mutalib et al, 2013). For 2010, a short spill of haze was recorded in the country but it did not affect Kuching (DoE, 2010). Kuching has been asserted as having PM_{10} pollution issues especially due to haze (Mutalib et al, 2013).

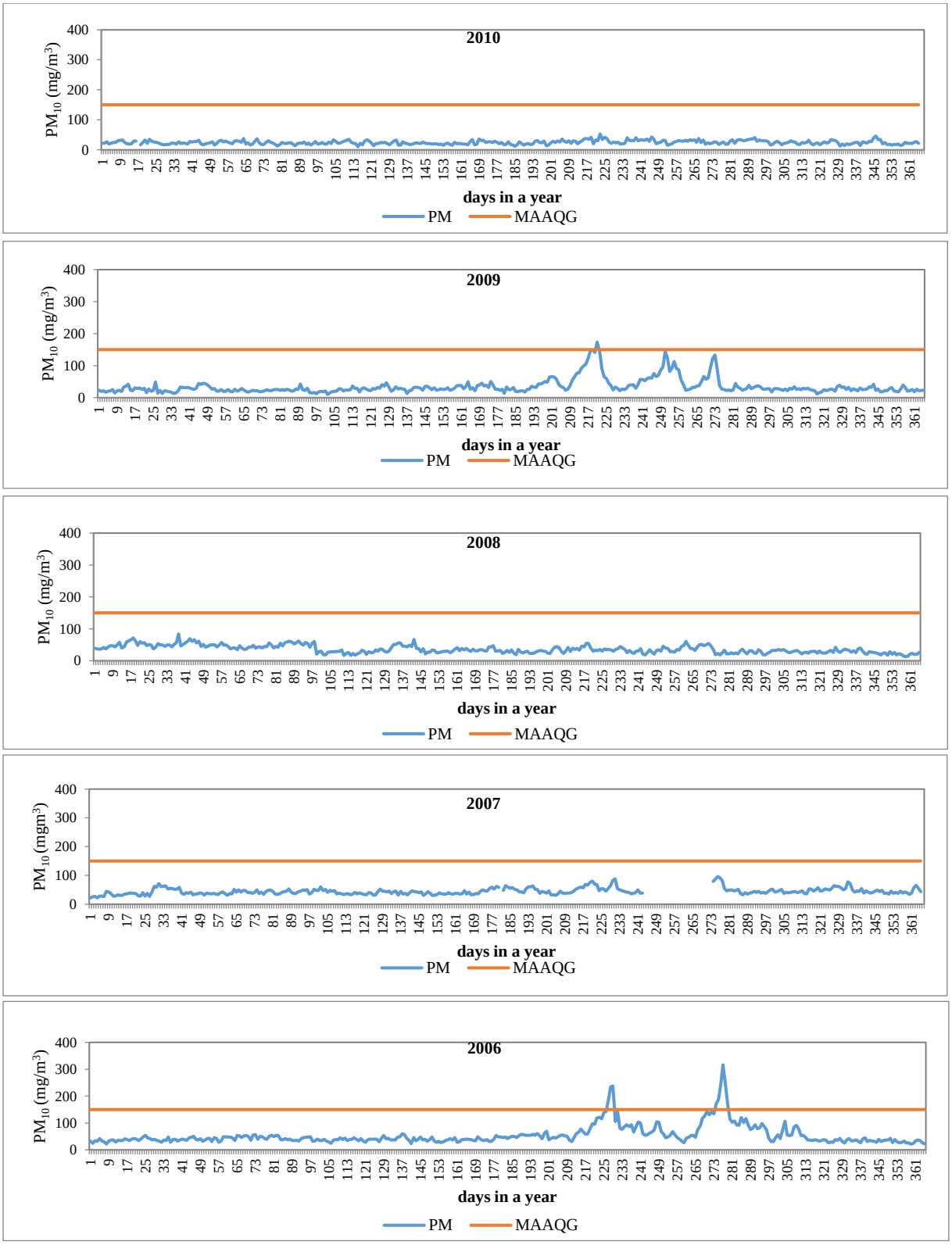

Figure 3 Time series trend for Kuching (2006 to 2010)

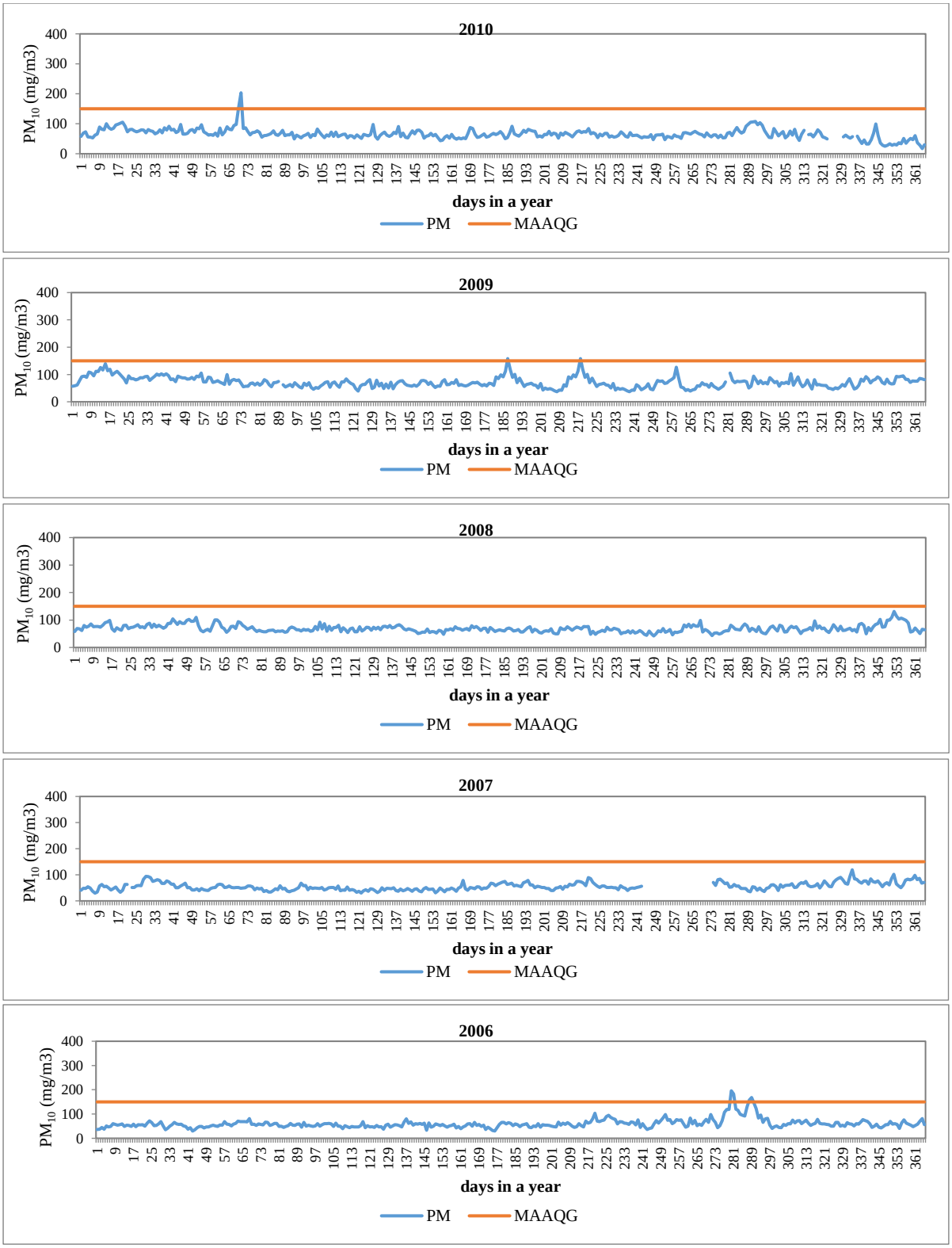

Figure 4 Time series trend for Balok Baru (2006 to 2010)

Figure 4 shows the trend analysis of Balok Baru. Year 2006, 2009, and 2010 had high PM_{10} concentration levels. In 2006, there were four violations recorded (MAAQG) and the highest concentration was 196 µg/m^3. Others were 182 µg/m^3, 157 µg/m^3 and 168 µg/m^3. In 2009, only one violation was recorded with concentration level at

158 µg/m^3. Meanwhile, for 2010 there were two violations (151 µg/m^3 and 203 µg/m^3). This area has been attributed to having high PM_{10} concentration due to industrial activities (Azid et al, 2014). Balok Baru industrial area was recorded as having substantial industrial air pollution sources and considerable haze pollution effect (DoE, 2006; 2010).

Principal Component Analysis

Monsoon Seasonal Variation Analysis

The PCA seasonal analysis result is as shown in Table 3 and Figure 5. The Varimax rotated result showed that a cumulative percentage of 83% was obtained in the SWM season. S1 had an Eigenvalue of 2.24 and a variance percentage of 36% having a significant value of >0.8 for temperature (positive relationship) and humidity (negative relationship).

Table 3: Varimax Rotated Result of Kuching

Variables	**Southwest Monsoon**				**Northeast Monsoon**		
	S1	S2	S3		N1	N2	N3
$PM_{10(d)}$	0.222	0.055	**-0.971**		0.092	-0.015	**0.994**
Temperature	**0.970**	0.032	-0.142		**0.976**	0.050	0.021
Humidity	**-0.896**	-0.185	0.229		**-0.946**	-0.159	-0.133
Wind speed	0.111	**0.992**	-0.052		0.127	**0.992**	-0.015
Wind direction	0.033	0.023	-0.068		0.056	-0.003	-0.051
Eigen value	2.2446	1.0032	0.9190		2.0421	1.1016	0.9644
variance	36.1%	20.5%	20.4%		37.5%	20%	20%
Cumulative	44.9%	65.0%	83.3%		40.8%	62.9%	82.2%

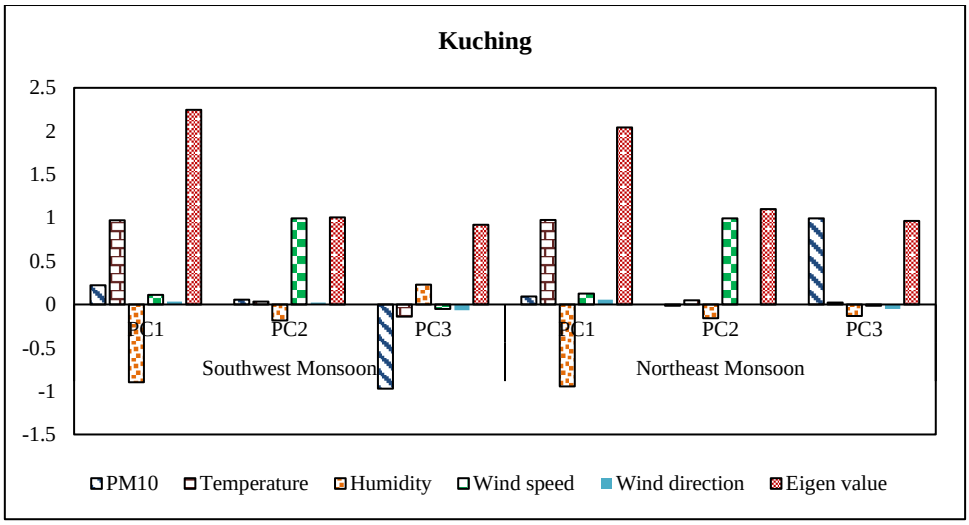

Figure 5 PCA Varimax Rotated Result for Seasonal Variation for Kuching (2006-2010)

Additionally, S2 and S3 had a variance of 20% each, with wind speed having a significant correlation value of 0.99. $PM_{10(d)}$ had a significant correlation result of -0.97 emphasizing a negative relationship in S3. Meanwhile, NEM season had a cumulative percentage of 82 %. N1 had an Eigenvalue of 2.04 with a variance percentage of 38%, temperature and humidity had significant correlation value of >0.9 each having positive and negative relationships, respectively. Furthermore, N2 and N3 had a variance result of 20% each. For N2 wind speed was significant, having a positive correlation value while PM_{10} had a positive relationship in N3. The results indicated that as temperature increases humidity decreases. This establishes existence of an inverse relationship which is

similarly resulted in by previous studies (Azmi et al, 2010; Dominick et al, 2012). Results show that PM_{10} has direct relationship with temperature, but has an inverse relationship with humidity in both seasons. This states that PM_{10} concentration increases with increasing temperature and decreasing humidity. Similar results were established by other reseachers (Kassomenos et al, 2014; Vardoulakis et al, 2008). Also, it was established that Kuching has substantial PM_{10} pollution due to favourable atmospheric conditions, particularly, humidity and temperature.

Table 4 and Figure 6 displays the analysis for Balok Baru. The Varimax rotated result showed that in the SWM season, S1 and S2 had Eigenvalues of 2.01 and 1.02, respectively. Having a cumulative percentage of 61% and variance percentage of 20% each. S1 and S2 had an influence of wind speed and PM_{10}, both having a significant positive correlation.

Table 4 Varimax Rotated Result of Balok Baru

Variables	**Southwest Monsoon**			**Northeast Monsoon**	
	S1	S2		N1	N2
$PM_{10(d)}$	0.049	**0.998**		0.041	0.117
Temperature	0.113	-0.014		**-0.956**	-0.031
Humidity	-0.182	-0.033		0.344	-0.167
Wind speed	**0.968**	0.053		0.030	**0.937**
Wind direction	-0.129	-0.024		0.108	-0.330
Eigen value	2.01	1.02		2.34	1.27
Variance	20.0%	20.0%		20.9%	20.6%
Cumulative	40.2%	60.6%		46.8%	72.1%

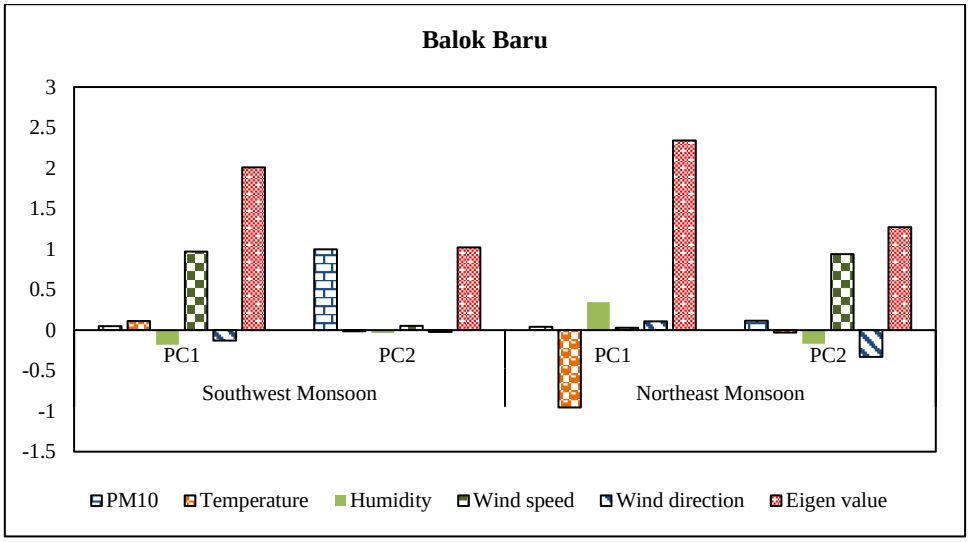

Figure 6 PCA Varimax Rotated Result for Seasonal Variation of Balok Baru (2006-2010)

For NEM season, N1 and N2 had Eigenvalues of 2.34 and 1.27 with a cumulative percentage of 72% and a variance of 20% each. N1 had the significant influence of temperature negatively, while N2 had an influence of wind speed having a positive correlation. The result suggested that apart from the sources of PM_{10} pollution in this area, some PM_{10} pollutants are transboundary in nature. In addition, temperature can assist in the chemical formation of air pollutants, while increasing wind speed and wind direction resulted in dilution of air pollutants. Meanwhile, humidity has the ability to absorb air pollutants, which establishes the assertion of previous studies

(Azmi et al, 2010; Kassomenos et al, 2014; Kozawa et al, 2012). However, an inverse relationship can occur between PM_{10} and temperature, especially on colder days corresponding to the reduction in dispersion of pollutants due to stable atmospheric conditions (Vardoulakis et al, 2008).

Principal Component Regression (PCR)

PCA was carried out to obtain the variables to be used in the PCR analysis. The Varimax rotated result for Kuching is shown in Table 5 and Figure 7.

Table 5 Principal Component Analysis of Kuching (2006-2010)

	PC1	**PC2**	**PC3**
$PM_{10(d)}$	0.182	**0.982**	0.017
Temperature	**0.973**	0.098	0.049
Humidity	**-0.928**	-0.204	0.166
Wind speed	0.123	0.017	**-0.992**
Wind direction	0.053	0.037	-0.007
Variance %	37.2%	20.4%	20.3%
Eigen value	2.18	0.99	0.92
Cumulative	43.7%	63.7%	82.3%

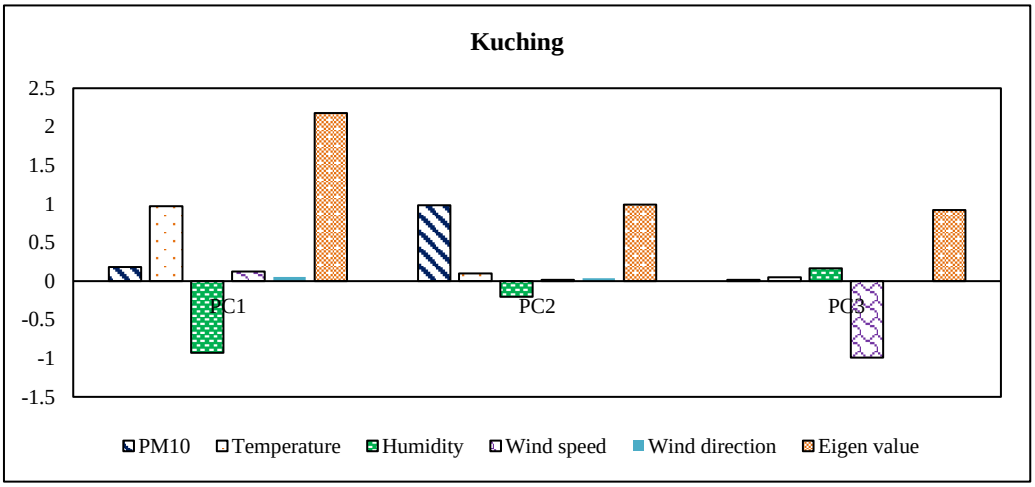

Figure 7 PCA Varimax Rotated Result Kuching (2006-2010)

PCA for Kuching had Eigenvalue of 2.18, 0.99 and 0.92 associated with PC1, PC2, and PC3, respectively. The cumulative variation of the 3 PCs was 82% and the variance percentages were 37% and 20%, respectively. Subsequently, the 3 PCs were subjected to regression analysis to form PCR models. The PCR model was used to predict next day PM_{10} concentration using $PM_{10(d+1)}$ as the predictant and PC1, PC2 and PC3 as the predictors.

Table 6 Principal Component Analysis of Balok Baru (2006-2010)

	PC1	PC2
$PM_{10(d)}$	0.007	-0.135
Temperature	**-0.973**	-0.047
Humidity	0.247	0.207
Wind direction	0.049	**0.970**
Eigen value	1.79	1.08
Variance %	25.3%	25.1%
Cumulative	44.7%	71.6%

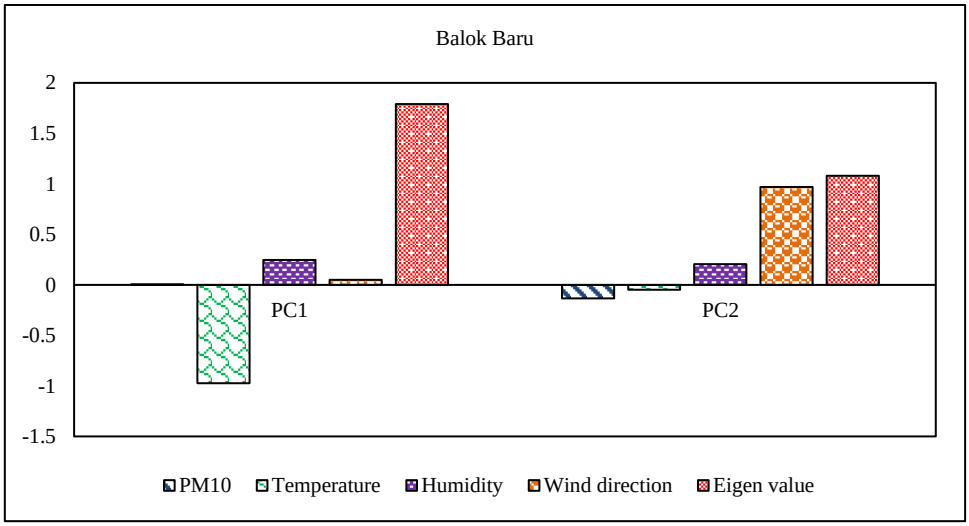

Figure 8 PCA Varimax Rotated Result Balok Baru Area (2006-2010)

The Varimax rotation result for the Balok Baru is shown in Figure 8 and Table 6. Four variables (wind speed was removed from the analysis to improve the model) were used for the PCA. The result showed that two PCs had Eigenvalue above 0.90. PC1 and PC2 had a cumulative percentage of 72% and the variance percentage of PC1 and PC2 was 25% each.

Lognormal Regression Analysis

Using lognormal regression, the models are displayed in Table 7. The analysis showed that the R^2 value for Kuching and Balok Baru were 0.78 and 0.63, respectively. The R^2 result was significant since it is >0.5, the analysis of variance (ANOVA) was significant at 99% confidence interval ($p<0.001$), showing the analysis is good.

Multiple Linear Regression Analysis

The MLR analysis showed that the R^2 values for Kuching and Balok Baru were 0.84 and 0.64 respectively. Both areas had significant R^2 values, having values > 0.5. From Table 7, for the prediction models, PM_{10} was the predictant while the predictors were, previous day PM_{10} concentration level ($PM_{10(d-1)}$), Temperature (T), Humidity (H), Wind Speed (WS), and Wind direction (WD).

Table 7 Prediction models and Performance indicators

Areas	Methods	Prediction Models	R^2	AdjR
Kuching	LR	PM_{10}= 8.00+0.83(log$PM_{10(d-1)}$)-0.37(logT)-1.33(logH)-0.16(logWS)-0.0092(logWD)	0.78	0.78
	MLR	PM_{10} = 98.67 + 0.87 ($PM_{10(d-1)}$) - 0.91(T) - 0.80(H) - 0.64 (WS) – 0.0004(WD)	0.84	0.83
	PCR	PM_{10} = 74.8 + 1.76 PC1 + 0.466 PC3	0.93	0.93
Balok Baru	LR	PM_{10}= 0.53+0.75(log $PM_{10(d-1)}$)+0.28(logT)-0.03(logH)+0.001(logWS)-0.06(logWD)	0.63	0.63
	MLR	PM_{10} = 8.63+0.77($PM_{10(d-1)}$)+0.49(T)-0.03(H)+0.11(WS)–0.03 (WD)	0.64	0.64
	PCR	PM_{10} = -25.5 + 1.05 PC1 + 1.84 PC2	0.90	0.90

All prediction models had significant R^2 and *Adj* R values, showing that LR, MLR, and PCR can be used to predict PM_{10} next day average concentration level with reputable result. Similar results were achieved in previous studies (Taşpınar et al, 2014; Ul-Saufie et al, 2013).

For the VIF analysis, the PCR models intially had a VIF result of 13-71 which was very high, hence, stepwise principal component regression was carried out to reduce the VIF value. Consequently, for Kuching, PC1 and PC3 were used for the model and the VIF value was reduced to 1.23. Meanwhile, PC1 and PC2 were used for the Balok Baru PCR model achieving a VIF result of 9.02 (Table 8). Additionally, the VIF results for both MLR and LR models were lower than 10, establishing that there was no multicollinearity problem in all the prediction models. The result for the performance indictaors is as shown in Table 8.

Table 8 Performance Indicators

Areas	Methods	RMSE	NAE	MAE	PA	IA	VIF
Kuching	LR	10.21	0.1549	6.4755	0.2946	0.9626	1.03-4.91
	MLR	10.09	0.1565	6.5415	0.8354	0.9532	1.11-2.06
	PCR	6.72	0.1262	5.2748	0.9269	0.9807	1.23
Balok Baru	LR	11.01	0.1253	8.0248	0.5881	0.8733	1.09-1.75
	MLR	10.87	0.1251	8.0129	0.6429	0.8815	1.09-1.89
	PCR	5.89	0.0705	4.5149	0.8949	0.9717	9.02

The performance indicator results established significant performance, which are comparable to previous studies using MLR, ANN and SR (Chaloulakou et al, 2003; Ul-Saufie et al, 2012b; Ul-Saufie et al, 2013). However, PCR models had better results for all the performance indicators.

Table 9 Analysis of Variance (ANOVA)

Area	Methods	F-value	P-value	Area	Methods	F-value	P-value
Kuching	LR	923	0.001	**Balok Baru**	LR	427	0.001
	MLR	1352	0.001		MLR	486	0.001
	PCR	8478	0.001		PCR	5771	0.001

Table 9 displayed the ANOVA result. The sum of squares showed a total variability for all the models. For the F-value, the results were between 923-8478 (Kuching) and 427-5771 (Balok Baru). All the F-values for the three models were greater than the critical values for the two study areas. This signifies that the outputs are not related at random and the models can significantly predict next day PM_{10} concentration level. A p-value of <0.001 was achieved, emphasizing that all models were good at 99% confidence interval. Overall, based on the ANOVA results, all methods are significantly reliable as confirmed by previous studies (Azmi et al, 2010; Ul-Saufie et al, 2012b).

The models ability to detect high daily average concentration for PM_{10} was also conducted using the MAAQG (150μg/m^3) and a newly proposed guideline value of 100μg/m^3. The POD, CSI, and TrueSS analysis showed that PCR had better detection capacity for high PM_{10} concentration levels than the other two methods as shown in Table 10. The FAR analysis established that the PCR model had less ability to give false alarm predictions than MLR models. However, similar FAR capabilities were established between PCR and LR especially in Kuching. Additionally, the POD analysis showed that PCR had a better detection capacity than LR, especially for 150μg/m^3 detection. MLR had a slightly better POD than the other two models having a slight difference of 1% between MLR and PCR for the 100μg/m^3 benchmark. The CSI analysis showed that PCR had better correspondence than the other methods in both benchmark levels. Lastly, TrueSS result showed that PCR had better result for the 150μg/m^3 detection than the other two methods, but PCR and MLR had similar abilities in the 100μg/m^3 benchmark in Kuching.

Table 10 Statistical Evaluation for prediction of high PM_{10} levels (150μg/m^3 and 100μg/m^3)

	Index	**Equation**	**Kuching**		**Balok Baru**	
			150 μg/m^3	**100 μg/m^3**	**150 μg/m^3**	**100 μg/m^3**
LR	Probability of Detection (POD)	A/(A+B)*	0.58	0.70	-	0.26
	False Alarm Rate (FAR)	C/(C+A)	0.13	0.03	-	0.22
	Critical Success Index (CSI)	A/(A+B+C)	0.53	0.68	-	0.24
	TrueSS	A/(A+B)+D/(D+C) -1	0.58	0.70	-	0.26
MLR	Probability of Detection (POD)	A/(A+B)	0.64	0.76	0.14	0.49
	False Alarm Rate (FAR)	C/(C+A)	0.30	0.09	0.50	0.19
	Critical Success Index (CSI)	A/(A+B+C)	0.50	0.70	0.13	0.44
	TrueSS	A/(A+B)+D/(D+C) -1	0.63	0.75	0.14	0.49
PCR	Probability of Detection (POD)	A/(A+B)	0.91	0.75	0.50	0.85
	False Alarm Rate (FAR)	C/(C+A)	0.09	0.03	0.25	0.10
	Critical Success Index (CSI)	A/(A+B+C)	0.83	0.73	0.43	0.78
	TrueSS	A/(A+B)+D/(D+C) -1	0.91	0.75	0.50	0.85

*A= observed and predicted exceedances, B= observed but not predicted, C= Predicted but not observed, D= None exceedances

Furthermore, the LR and MLR showed that detection for 100μg/m^3 was better than the 150μg/m^3 benchmark. This asserts the outcome of previous studies emphasizing that as the benchmark level decreases the performance of the prediction model improves (Slini et al., 2002; Chaloulakou et al., 2003). The PCR result showed a better detection of the 150μg/m^3 than the 100μg/m^3 benchmark for the Kuching, but better detection was observed for the 100μg/m^3 than 150μg/m^3 benchmark in Balok Baru.

Validation

To execute a real PCR model, the predicted values can be calculated using the original matrix component (PCA coefficient) of the PCA result as shown in Table 11.

Table 11 Matrix Component for PCR models

Variables	Kuching		Balok Baru	
	KPC1	KPC3	BPC1	BPC2
$PM_{10(d)}$	0.3655	0.4012	0.3149	0.3313
Temperature	0.6089	0.1144	0.2690	-0.7556
Humidity	-0.6402	-0.0715	-0.4972	0.4004
Wind speed	0.2597	-0.7303	0.5232	0.3249
Wind direction	0.1352	-0.5361	-0.5563	-0.2311

The PC's were named as KPC1 and 3 for Kuching, BPC1 and 2 for Balok Baru. This would be multiplied with each of the variables. Subsequently, the results would be inserted into the PCR models to predict PM_{10} concentration levels.

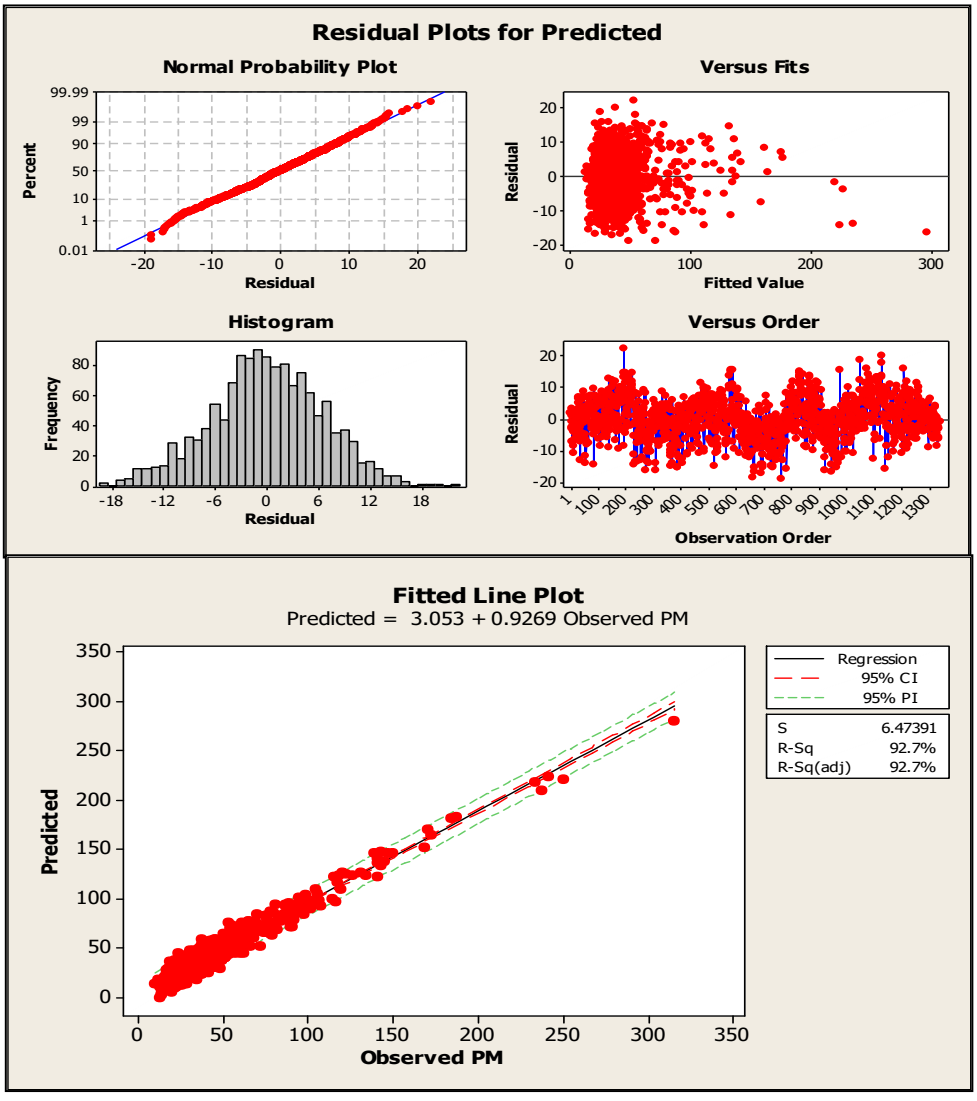

Figure 9 Residual and Fitted line Plots for Kuching

Figures 9 and 10 displayed the residual and fitted line plots. The plots showed good agreement between the predicted and observed values using the PCR models for both Kuching and Balok Baru. The results are comparable with previous study conducted by Ul-Saufie et al. (2013).

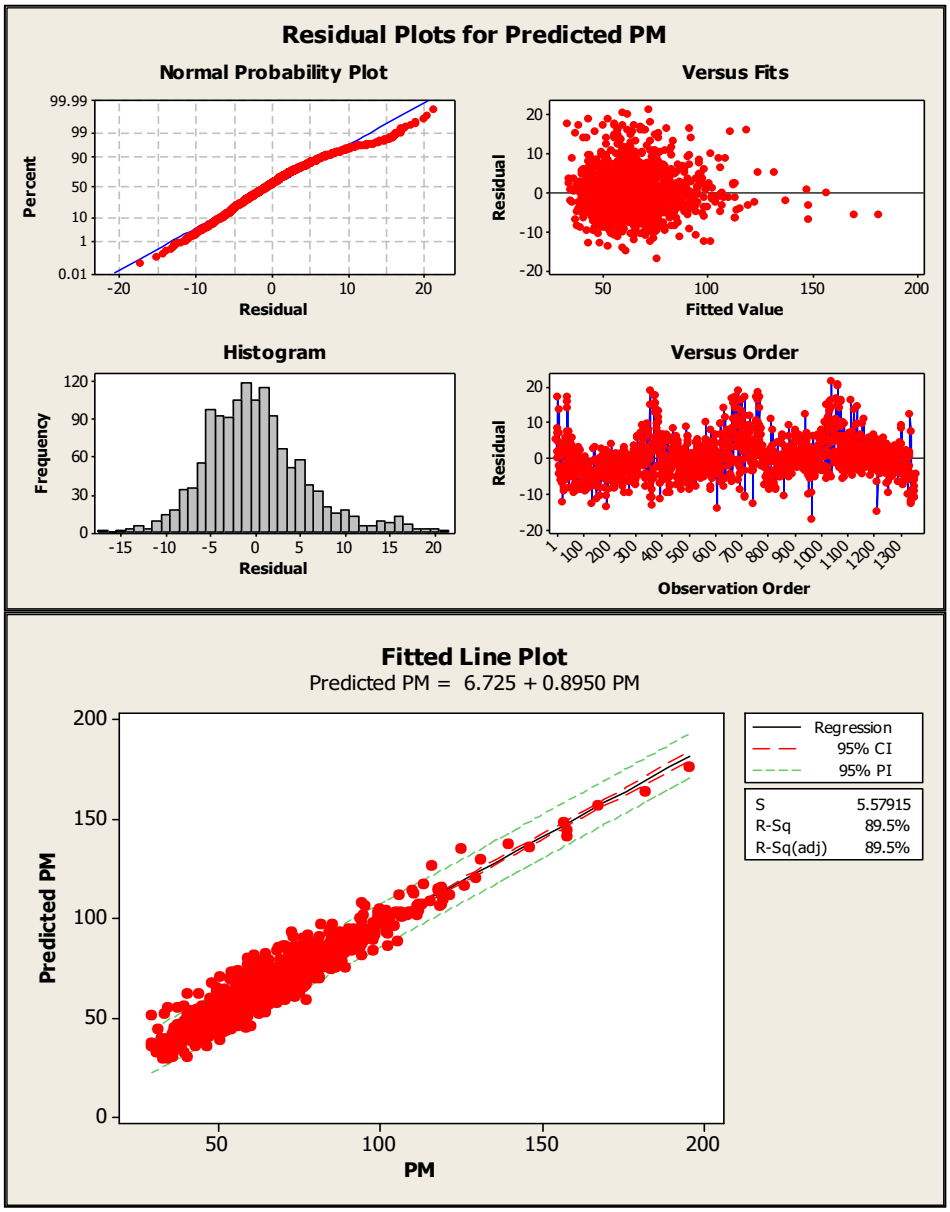

Figure 10 Residual and fitted line Plot for Balok Baru

Conclusion

Daily average PM_{10}, temperature (T), humidity (H), wind speed (WS) and wind direction (WD) data were analysed for five years (2006-2010), from two industrial air quality monitoring stations. Based on seasonal variation analysis using PCA Kuching had similar significant variables for both SWM and NEM seasons having significant correlation for humidity, temperature, wind speed and PM_{10}. Meanwhile, for Balok Baru, the significant variables were wind speed and PM_{10} in the SWM, while temperature and wind speed were more significant in the NEM

season. Furthermore, the prediction model using LR, MLR, and PCR established significant results in terms of the ability to predict next day average PM_{10} concentration levels, with all models having R^2 between 0.60 to 0.93. However, the PCR model was a better method, having higher predictability and lower error levels as established in the performance indicator analysis. No multicollinearity problem was established for LR and MLR models. However, step wise PCR was used to reduce the multicollinearity problem of the PCR models and favourable results were achieved. It was determined that PCR models had better ability to detect high PM_{10} concentration levels and achieve low false alarm rate than LR and MLR methods. The PCR prediction method can be used in predicting PM_{10} concentration for better air quality management strategies and sustainable development planning.

Acknowledgement

Appreciation goes to Universiti Teknologi PETRONAS for making this study possible. Additional gratitude goes to the Department of Environment (DOE) Malaysia for providing the data used for this study.

References

S.A. Abdul-Wahab, C.S. Bakheit, & S.M. Al-Alawi. (2005). Principal Component and Multiple Regression Analysis in Modelling of Ground-Level Ozone and Factors Affecting Its Concentrations. *Environmental Modelling & Software, 20*(10), 1263-1271.

N. Abdullah, S. Shuhaimi, Y. Toh, A. Shafee, & M. Maznorizan. (2011). The Study of Seasonal Variation of Pm10 Concentration in Peninsula, Sabah and Sarawak. *Malaysian Meteorological Department*(9).

A. Afzali, M. Rashid, B. Sabariah, & M. Ramli. (2014). *Pm10 Pollution: Its Prediction and Meteorological Influence in Pasirgudang, Johor*. Paper presented at the IOP Conference Series: Earth and Environmental Science.

A. Azid, H. Juahir, M.E. Toriman, A. Endut, M.K.A. Kamarudin, M.N.A. Rahman, . . . K. Yunus. (2014). Source Apportionment of Air Pollution: A Case Study in Malaysia. *Jurnal Teknologi, 72*(1).

S.Z. Azmi, M.T. Latif, A.S. Ismail, L. Juneng, & A.A. Jemain. (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.

A. Chaloulakou, G. Grivas, & N. Spyrellis. (2003). Neural Network and Multiple Regression Models for Pm10 Prediction in Athens: A Comparative Assessment. *Journal of the Air & Waste Management Association, 53*(10), 1183-1190.

D.o.E. (2006). Malaysia Environmental Quality Report. Malaysia: Ministry of Natural Resources and Environment.

D.o.E. (2007). Malaysia Environmental Quality Report. Malaysia: Ministry of Natural Resources.

D.o.E. (2008). Malaysia Environmental Quality Report. Malaysia: Ministry of Natural Resources and Environment.

D.o.E. (2009). Malaysia Environmental Quality Report 2009. Malaysia: Ministry of Natural Resources and Environment.

D.o.E. (2010). Annual Report on Malaysia Environmental Quality 2010. Malaysia: Ministry of Science, Technology and Environment,.

M. Department of Statistics. (2010). Population and Housing Census of Malaysia 2010. Malaysia: Department of Statistics, MAlaysia.

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

K.L. Ebi, & G. McGregor. (2008). Climate Change, Tropospheric Ozone and Particulate Matter, and Health Impacts. *Environ Health Perspect, 116*(11), 1449-1455.

H.A. Hamida, A.S. Yahayab, N.A. Ramlib, & A.Z. Ul-Saufie. (2012). Performance of Parameter Estimator for the Two-Parameter and Three-Parameter Gamma Distribution in Pm10 Data Modelling. *International Journal of Engineering and Technology, 2*(4), 637-643.

R.M. Harrison, D. Laxen, S. Moorcroft, & K. Laxen. (2012). Processes Affecting Concentrations of Fine Particulate Matter (Pm 2.5) in the Uk Atmosphere. *Atmospheric environment, 46*, 115-124.

S. Hörmann, B. Pfeiler, & E. Stadlober. (2005). Analysis and Prediction of Particulate Matter Pm10 for the Winter Season in Graz. *Austrian Journal of Statistics, 34*(4), 307-326.

N. Johnson, S. Kotz, & N. Balakrishnan. (1994). Lognormal Distributions. Continuous Univariate Distributions (Vol. 1): John Wiley & Sons.

P. Kassomenos, S. Vardoulakis, A. Chaloulakou, A. Paschalidou, G. Grivas, R. Borge, & J. Lumbreras. (2014). Study of Pm 10 and Pm 2.5 Levels in Three European Cities: Analysis of Intra and Inter Urban Variations. *Atmospheric environment, 87*, 153-163.

K. Katsouyanni, J.M. Samet, H. Anderson, R. Atkinson, A. Le Tertre, S. Medina, . . . D. Krewski. (2009). Air Pollution and Health: A European and North American Approach (Aphena). *Research report (Health Effects Institute)*(142), 5-90.

E. Kovač-Andrić, J. Brana, & V. Gvozdić. (2009). Impact of Meteorological Factors on Ozone Concentrations Modelled by Time Series Analysis and Multivariate Statistical Methods. *Ecological Informatics, 4*(2), 117-122.

K.H. Kozawa, A.M. Winer, & S.A. Fruin. (2012). Ultrafine Particle Size Distributions near Freeways: Effects of Differing Wind Directions on Exposure. *Atmospheric environment, 63*, 250-260.

M.T. Latif, L.S. Huey, & L. Juneng. (2012). Variations of Surface Ozone Concentration across the Klang Valley, Malaysia. *Atmospheric environment, 61*, 434-445.

H.-C. Lu. (2004). Estimating the Emission Source Reduction of Pm 10 in Central Taiwan. *Chemosphere, 54*(7), 805-814.

S.N.S.A. Mutalib, H. Juahir, A. Azid, S.M. Sharif, M.T. Latif, A.Z. Aris, . . . D. Dominick. (2013). Spatial and Temporal Air Quality Pattern Recognition Using Environmetric Techniques: A Case Study in Malaysia. *Environmental Science: Processes & Impacts, 15*(9), 1717-1728.

A. Namdeo, & M. Bell. (2005). Characteristics and Health Implications of Fine and Coarse Particulates at Roadside, Urban Background and Rural Sites in Uk. *Environment International, 31*(4), 565-573.

F. Nejadkoorki, & S. Baroutian. (2012). Forecasting Extreme Pm10 Concentrations Using Artificial Neural Networks. *Int. J. Environ. Res, 6*(1), 277-284.

J. Schwartz. (2001). Air Pollution and Blood Markers of Cardiovascular Risk. *Environmental health perspectives, 109*(Suppl 3), 405.

T. Slini, A. Kaprara, K. Karatzas, & N. Moussiopoulos. (2006). Pm 10 Forecasting for Thessaloniki, Greece. *Environmental Modelling & Software, 21*(4), 559-565.

F. Taşpınar. (2015). Improving Artificial Neural Network Model Predictions of Daily Average Pm10 Concentrations by Applying Pca and Implementing Seasonal Models. *Journal of the Air & Waste Management Association*(just-accepted).

F. Taşpınar, & Z. Bozkurt. (2014). Application of Artificial Neural Networks and Regression Models in the Prediction of Daily Maximum Pm10 Concentration in Düzce, Turkey.

A. Ul-Saufie, A. Yahya, N. Ramli, & H. Hamid. (2012a). *Future Pm10 Concentration Prediction Using Quantile Regression Models.* Paper presented at the International Conference on Environmental and Agriculture Engineering, IACSIT Press, Singapore.

A.Z. Ul-Saufie, A.S. Yahaya, N. Ramli, & H.A. Hamid. (2012b). Performance of Multiple Linear Regression Model for Long-Term Pm^ Sub 10^ Concentration Prediction Based on Gaseous and Meteorological Parameters. *Journal of Applied Sciences, 12*(14), 1488.

A.Z. Ul-Saufie, A.S. Yahaya, N.A. Ramli, N. Rosaida, & H.A. Hamid. (2013). Future Daily Pm 10 Concentrations Prediction by Combining Regression Models and Feedforward Backpropagation Models with Principle Component Analysis (Pca). *Atmospheric environment, 77*, 621-630.

A.Z. Ul-Saufie, A.S. Yahya, N.A. Ramli, & H.A. Hamid. (2011). 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, 1*(4).

S. Vardoulakis, & P. Kassomenos. (2008). Sources and Factors Affecting Pm 10 Levels in Two European Cities: Implications for Local Air Quality Management. *Atmospheric environment, 42*(17), 3949-3963.

N.F.F.M. Yusof, N.A. Ramli, A.S. Yahaya, N. Sansuddin, N.A. Ghazali, & W. al Madhoun. (2010). Monsoonal Differences and Probability Distribution of Pm10 Concentration. *Environmental Monitoring and Assessment, 163*(1-4), 655-667.