APPLICATION OF STEP WISE REGRESSION ANALYSIS IN PREDICTING PARTICULATE MATTER CONCENTRATION EPISODE

AMINA N. 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

Abstract

Particulate Matter is an air pollutant that has resulted in tremendous health effects to the exposed populace. Air quality forecasting is an established process where air pollutants particularly, Particulate Matter (PM_{10}) concentration is predicted in advance, so that adequate measures are implemented to reduce the health effect of PM_{10} to the barest level. The present study used daily average PM_{10} concentration and meteorological parameters (temperature, humidity, wind speed and wind direction) for five years (2006-2010) from three industrial air quality monitoring stations in Malaysia (Balok Baru, Tasek and Paka). Time series plot was used to assess PM_{10} pollution trend in the industrial areas. Additionally, Step Wise Regression (SWR) analysis was used to predict next day PM_{10} concentrations for the three industrial areas. The SWR method was compared with a Persistence model to assess its predictive capabilities. The results for the trend analysis showed that, Balok Baru (BB) had higher PM_{10} concentration levels, having high values in 2006, 2007 and 2009. These values were higher than the Malaysian Ambient Air Quality Guideline (MAAQG) of 150 $\mu g/m^3$. Subsequently, the other two industrial areas Tasek (TK) and Paka (PK) had no record of violating the MAAQG. The results for the SWR analysis had significant R^2 values of 0.64, 0.66 and 0.60, respectively. The model performance results for Variance Inflation Factor (VIF) were less than 5 and Durbin Watson test (DW) had value of 2 for each of the study areas, which were significant. The comparative analysis between SWR and Persistence model showed that the SWR had better capabilities, having lower errors for the BB, TK and PK areas. Using Root Mean Square Error (RMSE), the results showed error differences of 7, 12 and 16%, and higher predictability using Index of Agreement (IA), having a difference of 17, 19 and 16% for BB, TK and PK areas, respectively. The results showed that SWR can be used in predicting PM_{10} next day average concentration, while the extreme event detection results showed that 100 mg/m^3 were better detected than the 150 mg/m^3 bench marked levels.

Keywords: air pollution; particulate matter; daily average forecast; step wise regression analysis; persistence model

Introduction

Particulate Matter (PM) is an air pollutant that has the characteristics of being both a primary (emitted from source) and a secondary pollutant (chemical reaction of precursor pollutants) (Harrison et al., 2012). Increase in anthropogenic activities in regions have resulted in rise in Particulate Matter concentration (Kassomenos et al. 2014). Sources of PM pollution include forest fires, industrial and house hold activities and traffic (Henderson & Johnston, 2012; Hörmann et al. 2005; Kassomenos et al. 2014). In a study conducted by Alvarez et al. 2012, it was stated that machines, engines as well as equipment from plants can release PM pollutants. In addition, other sources of PM can include mechanical abrasion of brakes, tyres and tarmacs from vehicles which are used continuously (Hörmann et al. 2005). Interestingly, non-combustion sources can immensely contribute to an increase in PM pollution (Kassomenos et al. 2014), these sources can include natural sources such as windblown dust (Sharratt & Edgar 2011) as well as marine aerosol (Maggos et al. 2008). Additionally, it has been concluded that meteorological parameters and seasonal variation influences the distinct PM_{10} concentration patterns of different regions in the world (Abdullah et al. 2011; Engler et al. 2012; Latif et al. 2014).

Urban and rural areas are all affected by the increase in PM concentration (Namdeo & Bell, 2005). Several studies conducted in both Asian and European cities have recorded significant PM concentrations in both rural and urban settings (Gioda et al., 2011). Over the years, there has been remarkable focus on PM with 10 micron diameter measurement. PM_{10} has resulted in a lot of health effect including asthma, cardiovascular diseases and mortality (Dennekamp & Abramson, 2011; Schwartz, 2000, 2001). Consequently, in order to curtail the detrimental health effect of PM_{10}, procedures such as early warning measures are considered as alternative PM_{10} management strategy (Nejadkoorki & Baroutian, 2012). PM_{10} forecast, predicts the concentration level in advance so that adequate measures and preparedness can be carried out to reduce the impact of PM_{10} on the receptor (A. Ul-Saufie et al., 2012).

Emphatically, statistical methods have been used to predict PM_{10} hourly, maximum and daily concentration levels (Afzali et al., 2014; Chaloulakou et al., 2003; Taşpınar, 2015). Multiple Linear Regression (MLR) analysis has been widely used to predict daily average PM_{10} levels (Hörmann et al., 2005; Ul-Saufie et al., 2011; Vlachogianni et al., 2011). Additionally, other

methods such as Artificial Neural Network (ANN) (Taşpınar, 2015), Lognormal (Yusof et al., 2010) as well as Quantile Regression (QR) (A. Ul-Saufie et al., 2012) have been applied for PM concentrations prediction. Step Wise Regression (SWR) have also been used previously to predict PM_{10} concentration levels (Henderson et al., 2007; Hosiokangas et al., 1999).

Furthermore, Step Wise Regression (SWR) has been used to enhance the performance of ANN in forecasting fine particulate Matter (Ordieres et al. 2005). Besides, SWR was used to analyse indoor PM_{10} concentration (Lung et al. 2003) as well as forecasting PM_{10} concentration in an Urban setting (Maraziotis et al. 2008). All the previous studies have predicted PM_{10} concentration with significant coefficient of determination and error detection results, but these models do not necessarily have the capability of effectively predicting extreme PM_{10} levels and serving as an alarm notification technique. Hence, this study investigated the use of SWR method to forecast next day PM_{10} daily average concentration in industrial areas and to assess the capability of these method to predict extreme PM_{10} concentration levels.

This study intended to use time series plot to understand the daily average PM_{10} concentration trend in three industrial areas. These PM_{10} trends would be bench marked with the MAAQG ($150\mu g/m^3$) to understand the violations or other wise of the industrial areas. Subsequently, SWR was used to predict next day average PM_{10} concentrations using meteorological parameters (temperature, humidity, wind speed and wind direction) for three industrial areas. Additionally, the SWR analysis was compared with a persistence model to assess its competence. Subsequently, the SWR models were subjected to statistical evaluation to assess its capability in predicting extreme events. These conducts can be used by relevant agencies to assess PM_{10} concentration and to choose a suitable method that would accurately predict PM_{10} concentrations in advance.

2. Materials and Methods

2.1 Study areas

Three Industrial areas were chosen for this analysis as follow:

Pahang which is one of the largest states in peninsular Malaysia. The state capital of Pahang is Kuantan, located at latitude 3.7500° N and longitude 102.5000° E, with a total area of 36,137 sq. km and population of about 1.4 million (Department of Statistics, 2010). Pahang has temperature range between 23ºC to 32ºC, an average wind speed of about 11 km/hr while average relative humidity ranges from 71 to 95% with an annual rainfall ranging between 2000-3000 mm. The industrial air quality monitoring station is situated in Balok Baru, it is located at Latitude E103°22.955 and longitude N03°57.726.

Perak is a state in peninsular Malaysia located at latitude 4.7500ºN and longitude 101.0000ºE. Perak covers an area of 21,035 sq.km with a population of about 2.2 million people (Department of Statistics, 2010). Perak has temperature ranging from 23ºC to 33ºC, with relative humidity of about 82%, with an annual rainfall of about 3,218mm. Additionally, Perak has an average wind speed of about 6 km/hr. One of the industrial air quality monitoring stations in this state is situated at Tasek Ipoh, it is located at latitude E101ºº06.964 and longitude N04º37.781.

Terengganu is located in the north-eastern peninsular and it is bordered in the east by the south china sea, with a coordinate of latitude of 4.7500° N and longitude 103.0000° E. Having a total area of 13,035 sq. km, with a total population of about 1 million people (Department of Statistics, 2010). Terengganu has average temperature between 19ºC to 32ºC, an annual average rainfall of about 3000 mm with wind speed of about 5 km/hr and relative humidity between 73 % to 98 %. The industrial air quality monitoring station in this state is situated at Paka, at latitude N04°35.880 and longitude E103°26.096.

2.2 Monitoring Records

Daily average data for PM_{10} ($\mu g/m^3$), temperature (ºC), relative humidity (%), wind speed (km/hr) and wind direction (degree) for a period of five years (2006-2010) were used to forecast next day PM_{10} concentrations. The data were acquired from the Department of Environment (DoE), Ministry of Natural Resources and Environment of Malaysia.

2.3 Methods

2.3.1 Time Series Trend

Time Series Plots of daily average PM_{10} concentrations for five years for the industrial areas under this study were analysed to show the PM_{10} daily average trend. Figures 1, 2 and 3 show

the time series trend for Balok Baru (BB), Tasek (TK) and Paka (PK). The PM_{10} daily average trend was bench marked with the Malaysian Ambient Air Quality Guideline (MAAQG) limit of 150 µg/m^3 to show the areas that violated the MAAQG or are safely below it. The trend analysis was from year 2006 to 2010.

2.3.2 Step Wise Regression (SWR) Analysis

Step Wise Regression (SWR) Analysis is a step by step approach were insignificant variables are removed from the regression analysis allowing only important variables to be present in the prediction models. Step wise regression can transfer from being a linear regression equation to a multiple linear regression equation (Thomas & Jacko, 2007). For this study forward selection was used, the analysis starts with one predictor variable, testing the selection of each variable using a chosen model. Next, comparison criteria particularly F-test or t-test were done to subsequently adding the predictor variable that would improve the data the most. This selection process was repeated on all the predictor variables until no further improvement is achieved (Thompson, 2001). Step Wise Regression was carried out using multiple linear regressions as shown in equation 1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_3 X_3 + \varepsilon i \quad \text{(eqn.1)}$$

Where Y is the dependent variable (predictant), β_0 is the constant coefficient, $\beta_1, \beta_2, \ldots \beta_p$ are the regression coefficients of the independent variables $X_1, X_2, \ldots X_p$ (predictors) and ε is the residual error.

2.3.3 Persistence Model Analysis

Persistence model is a forecasting analysis that is carried out to assess the predictive capabilities when compared to another model. Persistence Model is shown in equation 2

$$PM_{10}\,(\text{today}) = PM_{10}\,(\text{tomorrow}) \quad \text{(eqn.2)}$$

2.3.4 Performance Indicators

Validation of the forecasting models were carried out to assess the model performance and suitability. These was done using Coefficient of determination (R^2), Adjusted R (AdjR), Variance Inflation factor (VIF), Durbin Watson (DW) Test, Root mean square Error (RMSE), Mean Absolute Error (MAE), Mean Biased Error (MBE) and Index of agreement (IA).

Table 1 Performance Indicators for the Prediction Analysis

Performance Indicator	Description	Equation
Coefficient of Determination (R^2)	It is used to signify how well the model results fit to the observed data points.	$R^2 = \left[\frac{\sum_{i=1}^{n}(O_i-\bar{O})^2.(P_i-\bar{P})}{n.\sigma_0.\sigma_p}\right]^2$
Adjusted R	It is a modified version of R^2 and increases only when a new variable improves the model more than would be expected by chance.	$Adjusted\ R = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$
Variance Inflation Factor (VIF)	Provides a format that measures how much the variance of an estimated regression coefficient is increased because of collinearity.	$VIF_i = \frac{1}{1-R_i^2}$
Durbin Watson (DW) Test	It is used to detect the presence of autocorrelation of residuals in a regression analysis.	$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i-1}^{n} e_i^2}$
Root Mean Square Error (RMSE)	Explains the overall accuracy of the model, by summarising the difference between the observed and modelled concentration.	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}$
Mean Absolute Error (MAE)	It is used to measure the amount of error in a prediction model.	$MAE = \frac{\sum_{i=1}^{n}\lvert P_i - O_i\rvert}{n}$
Mean Biased Error (MBE)	It is used to assess the over or under predictability of a model observations.	$MBE = \frac{\sum_{i=1}^{n}(P_i - O_i)}{n}$
Index of Agreement (IA)	It is used to show the overall accuracy of the prediction model. When the IA value is closer to 1, it indicates that the prediction method is good.	$IA = 1 - \left[\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(\lvert P_i - \bar{O}\rvert - \lvert O_i - \bar{O}\rvert)^2}\right]$

Where, *d* is the Durbin-Watson statistical test, *n* represent the number of observations while e_i is the difference between the observed and predicted values, *d* should be between 0 and 4. When *d* = 2 it signifies that there is no autocorrelation in the analysis. Whereas if *d* approaches 0 then this indicates positive autocorrelation while if the value of *d* move towards 4, this indicates that there is negative autocorrelation. *n* is total number of annual measurements, *Pi*

is predicted values, *Oi* is observed values, $\bar{O}$ is the mean observed values, $\bar{P}$ is mean of the predicted values, σ_p is standard deviation of the predicted values, and σ_o is standard deviation of the observed values.

3. Results and discussion

Balok Baru Industrial Area

Based on Figure 1, year 2007 and 2008 had low PM_{10} daily average concentration levels with the highest values at 119 µg/m^3 and 131 µg/m^3, respectively. Year 2006, 2009 and 2010 had high PM_{10} concentration levels. The highest concentration was 196 µg/m^3, while others were 182 µg/m^3, 157 µg/m^3 and 168 µg/m^3. For year 2009, the highest concentration level was at 158 µg/m^3. Meanwhile, for 2010 the highest concentration levels were observed to be 151 µg/m^3 and 203 µg/m^3.

Tasek Industrial Area

From the time series in Figure 2, the trend analysis showed that duration of 2006 to 2010 had low PM_{10} daily average concentration levels, which were lower than the MAAQG of 150µg/m^3. The highest concentrations for 2006-2010 were 128 µg/m^3, 99 µg/m^3, 84 µg/m^3, 93 µg/m^3 and 62 µg/m^3, respectively.

Paka Industrial Area

The trend analysis of Paka area from 2006 to 2010 shows that PM_{10} daily average concentration levels for the area were below the MAAQG (Figure 3). The Maximum PM_{10} daily average concentration levels in duration of 2006 to 2010 were 119 µg/m^3, 73 µg/m^3, 66 µg/m^3, 88 µg/m^3 and 68 µg/m^3, respectively.

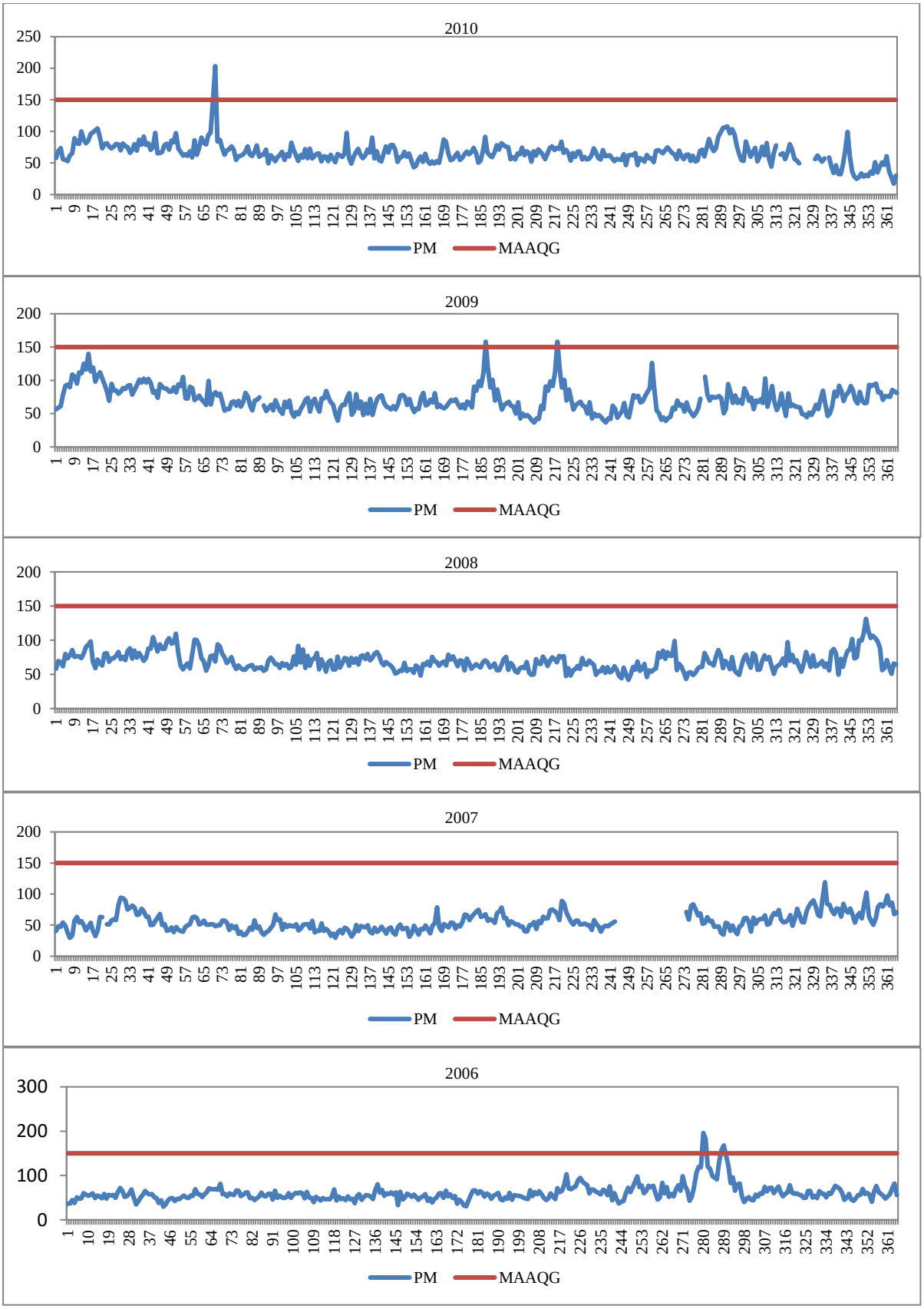

Figure 1. Trend Analysis of PM_{10} concentrations (2006-2010) for Balok Baru Area

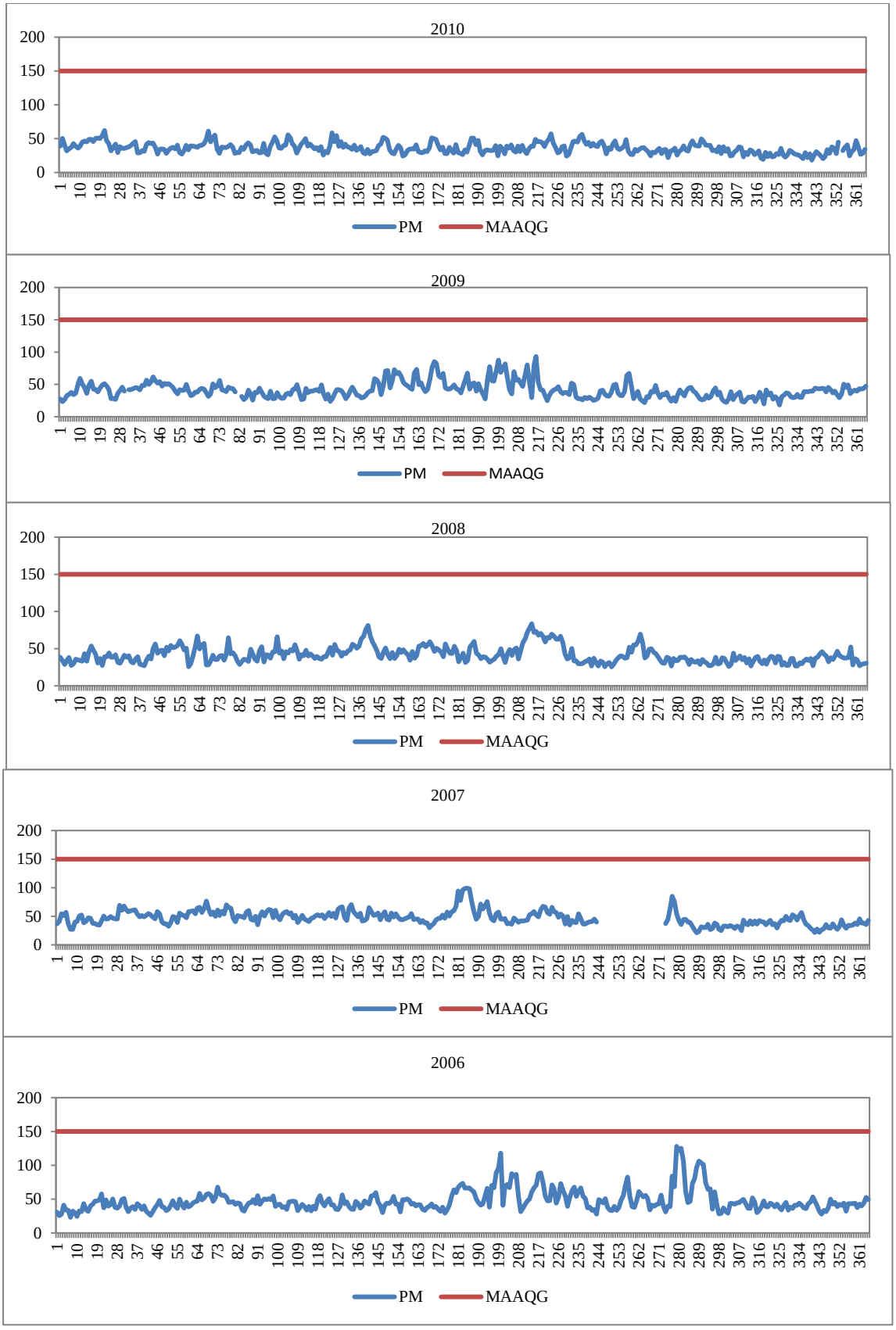

Figure 2. Trend Analysis of PM_{10} concentrations (2006-2010) for Tasek Area

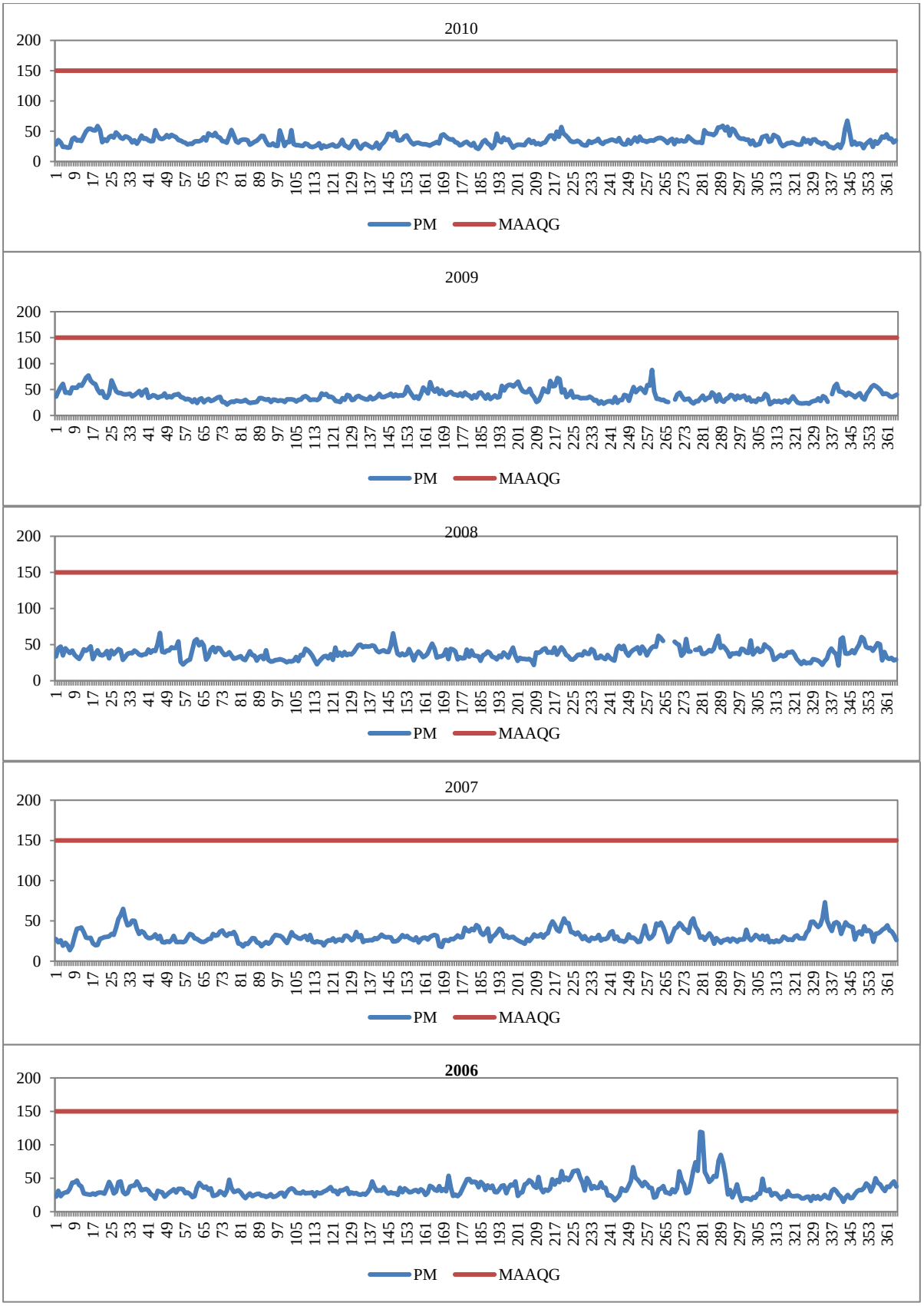

Figure 3. Trend Analysis of PM_{10} concentrations (2006-2010) for Paka Area

Table 2. Step Wise Regression Analysis

Study area	SWR equations	R^2	AdjR
Balok Baru	PM_{10}=13.156+0.795$PM_{10(d-1)}$	63.15%	63.13%
	$PM_{10=}$21.147+0.770 $PM_{10(d-1)}$-0.035WD	64.17%	64.17%
	$PM_{10=}$6.678+0.773 $PM_{10(d-1)}$-0.0333WD+0.52T	64.27%	64.20%
	PM_{10}=95.3-0.09WD-0.50T	10.00%	10.00%
Tasek	$PM_{10=}$9.433+0.7807 $PM_{10(d-1)}$	60.81%	60.79%
	PM_{10}=52.207+0.731 $PM_{10(d-1)}$+2.29T	64.31%	64.26%
	PM_{10}=52.450+0.736P $PM_{10(d-1)}$+2.46T-0.72WS	64.81%	64.74%
	PM_{10}=7.327+0.718 $PM_{10(d-1)}$+1.33T-0.98WS-0.339H	65.39%	65.29%
	PM_{10}=7.688+0.712$PM_{10(d-1)}$+1.38T-0.99WS-0.321H-0.0140WD	65.60%	65.49%
	PM_{10}=92.46+1.27T-1.04WS-0.92H-0.041WD	21.46%	21.26%
Paka	PM_{10}=8.344+0.761 $PM_{10(d-1)}$	58.27%	58.24%
	PM_{10}=5.826+0.736 $PM_{10(d-1)}$+0.714WS	59.87%	59.82%
	PM_{10}=2.135+0.737 $PM_{10(d-1)}$+0.732WS + 0.137T	59.97%	59.88%
	PM_{10}=3.686+0.736$PM_{10(d-1)}$+0.632WS+0.140T-0.0055WD	60.03%	59.91%
	PM_{10}=29.68+1.36WS+0.028T-0.0093WD	10.00%	10.00%

Table 2 shows various forecasting models, in addition to displaying the significance of the variables and the performance capacity of the models. The forecasting models include following; the dependent variable Y is PM_{10}, while the independent variables are previous day PM_{10} concentration ($PM_{10(d-1)}$), Temperature (T), Humidity (H), Wind Speed (WS) as well as Wind Direction (WD).

For Balok Baru area four equation models were established with different R^2 and adj R values. The highest significance was based on R^2 value of 64.27% and adj R value of 64.20%, which was the third prediction equation, having only three predictors; $PM_{10(d-1)}$, WS and T as the significant predictor variables. Additionally, meteorological parameters and previous day PM_{10} concentration accounted for 10% and 63% respectively of the daily average PM_{10} variation in the Balok Baru area.

For the Tasek area, six different equation models were established with the fifth prediction equation having the highest R^2 value of 65.60% and adj R values of 65.49%, having all the five predictor variables; $PM_{10(d-1)}$, T, WS, H as well as WD as the significant predictor variable. Additionally, meteorological parameters and previous day PM_{10} concentrations accounted for 21% and 60% variability of daily average PM_{10} concentration in this area.

Subsequently, for Paka area, there were five different equation models, with the fourth prediction model having the highest R^2 value of 60.03% and adj R of 59.91%, having four predictor variables; $PM_{10(d-1)}$, WS, T and WD. Additionally, meteorological parameters and previous day PM_{10} concentrations accounted for 10% and 60% variability of daily average PM_{10} concentration the Paka area.

The prediction models that had the highest R^2 values of 0.64, 0.66 and 0.60 respectively as well as highest adj R values of 0.64, 0.66 and 0.60 respectively, were regarded as the best forecasting models for this study and are shown in Table 3. These R^2 values are agreeable to previous study using SWR (Taşpınar & Bozkurt, 2014). The R^2 result for this study were similar to the R^2 result obtained in previous studies using ANN (Afzali et al., 2014; Ul-Saufie et al., 2013).

Table 3. Model Performance

Study Area	Best prediction models	R^2	AdjR	Range VIF	Durbin-Watson
Balok Baru	$PM_{10=}6.678+0.773\ PM_{10(d-1)}-0.0333WD+0.52T$	0.64	0.64	1.088-1.886	2
Tasek	$PM_{10}=7.688+0.712PM_{10(d-1)}+1.38T-0.99WS-0.321H-0.0140WD$	0.66	0.66	1.042-3.100	2
Paka	$PM_{10}=3.686+0.736PM_{10(d-1)}+0.632WS+0.140T-0.0055WD$	0.60	0.60	1.016-1.646	2

The prediction models for each of the industrial areas were subjected to other model performance measurement known as variance inflation factor (VIF) and Durbin Watsons (DW) test. Result for the VIF shows that, all the study areas had values below 5. This indicates that all the prediction equations had no problem of collinearity. Additionally, the DW test result showed that all prediction models have no problem of autocorrelation, as all DW values achieved at 2. These result are comparable with the results obtained from previous studies (A. Z. Ul-Saufie et al., 2012; Ul-Saufie et al., 2013). Table 4 shows the ANOVA of the best forecasting models.

Table 4. ANOVA analysis

	Result for ANOVA					
Stations		**Df**	**Sum of Squares**	**Mean square**	**F-value**	**Significance**
Balok Baru	Regression	3	288562	96187	812	P<0.001
	Residual	1354	160393	118		
	Total	1357	448955			
Tasek	Regression	5	180975	36195	572	P<0.001
	Residual	1500	94898	63		
	Total	1505	275872			
Paka	Regression	4	87041	21760	529	P<0.001
	Residual	1408	57960	41		
	Total	1412	145001			

The results of ANOVA analysis showed that the F-value for Balok Baru, Tasek and Paka industrial areas were 812, 572 and 529, respectively. This signifies that all models are able to

predict next day PM_{10} concentration significantly as the observed values were greater than the critical values of F (5.42, 4.13, 4.64) for each area of study. In overall, based on the ANOVA results all models are significantly reliable and the regression output is not at random. The findings are similar to those achieved in previous studies (Azmi et al., 2010; A. Z. Ul-Saufie et al., 2012).

Table 5. Comparative Analysis between Step Wise Regression and Persistence Model

	Balok Baru		**Tasek**		**Paka**	
	SWR Model	**Persistence Model**	**SWR Model**	**Persistence Model**	**SWR Model**	**Persistence Model**
SDEV O	18.189	18.188	13.833	13.538	10.134	10.137
SDEV P	14.584	18.174	11.218	13.536	7.847	10.136
AVE O	64.000	64.000	43.000	43.000	35.000	35.000
AVE P	64.000	64.000	43.000	43.000	35.000	35.000
RMSE	10.873	11.684	7.941	8.998	6.407	6.967
MAE	8.009	8.683	5.810	5.956	4.382	4.633
NAE	0.125	0.136	0.136	0.139	0.125	0.132
MBE	0.130	-0.014	-0.071	-0.002	0.009	-0.002
IA	0.881	0.729	0.887	0.719	0.862	0.721

SWR and persistence model were compared as shown in Table 5.The standard deviation (SDEV) of the observed data was higher than that of the predicted data. This can be considered as typical because the SWR model attempts to approximate an average behaviour. This was the case in previous study (Slini et al., 2002). For the error assessment RMSE, MAE and NAE were used. The results showed that for all the study areas, the SWR model had lower error than the persistence model indicating a good agreement between the residual values, thus a better capability of the SWR. Additionally, the MBE result using SWR model for Balok Baru and Paka areas slightly over predicted the PM_{10} daily average concentration. While Tasek area slightly under predicted the concentration levels. For the persistence models, all areas slightly under predicted the average concentration levels. Overall, the MBE result shows that all models had low residual errors, with both methods having values approaching zero. The IA result

showed that both models had good predicting capacity, having values approaching 1, but SWR had better predicting skills than the persistence model, having values > 0.8.

Table 6. Statistical Evaluation for prediction of high PM_{10} levels (150mg/m^3 and 100mg/m^3)

Index	**Equation**	**Balok Baru**		**Tasek**	
		150 mg/m^3	**100 mg/m^3**	**150 mg/m^3**	**100 mg/m^3**
Probability of Detection (POD)	A/(A+B)	0.33	0.44	0.16	0.44
False Alarm Rate (FAR)	C/(C+A)	0.50	0.25	0.50	0.25
Critical Success Index (CSI)	A/(A+B+C)	0.17	0.38	0.17	0.44

A= observed and predicted exceedances, B= observed but not predicted, C= Predicted but not observed,

Subsequently, the statistical evaluation of the SWR model to predict high concentration levels is shown in table 6. The models ability to test for high daily average concentration of PM_{10} was also carried out. These was done to assess the MAAQG of 150 mg/m^3 and the newly proposed MAAQG of PM_{10} which is 100 mg/m^3. The result for the Probability of Detection (POD) showed that the 100 mg/m^3 bench mark level would be better detected than the 150 mg/m^3 for two areas in this study . Paka area was excluded in the analysis as a result of no record of observed and predicted exceedances (A), as well as no record of predicted but not observed (C) values in the area for both bench marked levels. The False Alarm Rate (FAR) analysis result showed higher false alarm rate for 150 mg/m^3 than the 100 mg/m^3 bench mark. This states that the SWR can better detect alarm rates for lower bench mark levels. Furthermore, the Critical Success Index (CSI) analysis also showed that the 100 mg/m^3 would be better detected than the 150 mg/m^3 bench mark levels using SWR methods. This asserts the outcomes of previous studies emphasizing that as the bench mark level decreases the performance of the prediction model improves (Slini et al., 2002; Chaloulakou et al., 2003).

Conclusion

The daily average PM_{10} concentrations from year 2006 to 2010 for two of the study areas (Tasek and Paka) were low, having values below the MAAQG daily average limit of 150 $\mu g/m^3$. Subsequently, for the Balok Baru industrial area, there were some violations in years 2006, 2009 and 2010. The findings indicate that, even though all three stations are meant for industrial settings, the level of pollutions and activities that contribute to PM_{10} concentration increment varies.

For the Prediction models, the analysis showed that SWR had significant R^2 values for all the study areas. Apparently, all the SWR models had R^2 values ≥ 0.6, at significance level of $p<0.001$ as well as considerable ANOVA results. The SWR analysis showed that the combination of previous day PM_{10} daily average concentrations and meteorological parameters as predictors would account for higher percentage of daily average PM_{10} variation. Additionally, the results of comparative analysis between the SWR and Persistence model showed that, the SWR had better prediction abilities than the Persistence model. Furthermore, the SWR model had the capability to predict daily average PM_{10} concentration levels using meteorological parameters and previous day PM_{10} concentration with significant results. Subsequently, the statistical evaluation to assess the ability for the SWR model to predict extreme PM_{10} event showed that, the model has a reduced capability in predicting bench marked levels >150 mg/m^3, but better capabilities in detecting bench marked levels of 100 mg/m^3 PM_{10} concentration levels in these areas.

Overall, this study shows the complexity in statistical atmospheric analysis for forecasting PM_{10} daily average concentration and its distinctness in peculiar areas. Further analysis can be carried out in extremely high concentrated areas to understand the capabilities of SWR model. Additionally, other precursor pollutants and meteorological parameters could be added to the models to assess the proficiency of the SWR model, which are intended for air quality management strategy, designed in favour of better implementation programs to assess future PM_{10} concentrations episodes.

Reference

Abdullah, N., Shuhaimi, S., Toh, Y., Shafee, A., & Maznorizan, M. (2011). The Study of Seasonal Variation of PM10 Concentration in Peninsula, Sabah and Sarawak. *Malaysian Meteorological Department*(9).

Afzali, A., Rashid, M., Sabariah, B., & Ramli, M. (2014). *PM10 Pollution: Its Prediction and Meteorological Influence in PasirGudang, Johor.* Paper presented at the IOP Conference Series: Earth and Environmental Science.

Alvarez, R. A., & Paranhos, E. (2012). Air pollution issues associated. *Air Pollution Issues Associated." EM Feature June*.

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Quality, Atmosphere & Health, 3*(1), 53-64.

Chaloulakou, A., Grivas, G., & Spyrellis, N. (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.

Dennekamp, M., & Abramson, M. J. (2011). The effects of bushfire smoke on respiratory health. *Respirology, 16*(2), 198-209.

Department of Statistics, M. (2010). Population and housing census of Malaysia 2010. Malaysia: Department of Statistics, MAlaysia.

Engler, C., Birmili, W., Spindler, G., & Wiedensohler, A. (2012). Analysis of exceedances in the daily PM 10 mass concentration (50 µg m− 3) at a roadside station in Leipzig, Germany. *Atmospheric Chemistry and Physics, 12*(21), 10107-10123.

Gioda, A., Amaral, B. S., Monteiro, I. L. G., & Saint'Pierre, T. D. (2011). Chemical composition, sources, solubility, and transport of aerosol trace elements in a tropical region. *Journal of Environmental Monitoring, 13*(8), 2134-2142.

Harrison, R. M., Laxen, D., Moorcroft, S., & Laxen, K. (2012). Processes affecting concentrations of fine particulate matter (PM 2.5) in the UK atmosphere. *Atmospheric Environment, 46*, 115-124.

Henderson, S. B., Beckerman, B., Jerrett, M., & Brauer, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental science & technology, 41*(7), 2422-2428.

Henderson, S. B., & Johnston, F. H. (2012). Measures of forest fire smoke exposure and their associations with respiratory health outcomes. *Current opinion in allergy and clinical immunology, 12*(3), 221-227.

Hörmann, S., Pfeiler, B., & Stadlober, E. (2005). Analysis and prediction of particulate matter PM10 for the winter season in Graz. *Austrian Journal of Statistics, 34*(4), 307-326.

Hosiokangas, J., Ruuskanen, J., & Pekkanen, J. (1999). Effects of soil dust episodes and mixed fuel sources on source apportionment of PM10 particles in Kuopio, Finland. *Atmospheric Environment, 33*(23), 3821-3829.

Kassomenos, P., Vardoulakis, S., Chaloulakou, A., Paschalidou, A., Grivas, G., Borge, R., & Lumbreras, J. (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.

Latif, M. T., Dominick, D., Ahamad, F., Khan, M. F., Juneng, L., Hamzah, F. M., & Nadzir, M. S. (2014). Long term assessment of air quality from a background station on the Malaysian Peninsula. *Sci Total Environ, 482-483*, 336-348. doi: 10.1016/j.scitotenv.2014.02.132

Lung, S. C., Kao, M. C., & Hu, S. C. (2003). Contribution of incense burning to indoor PM10 and particle-bound polycyclic aromatic hydrocarbons under two ventilation conditions. *Indoor Air, 13*(2), 194-199.

Maggos, T., Michopoulos, J., Flocas, H., Asimakopoulos, D., & Vasilakos, C. (2008). Ions species size distribution in particulate matter associated with VOCs and meteorological conditions over an urban region. *Chemosphere, 72*(3), 496-503.

Maraziotis, E., Sarotis, L., Marazioti, C., & Marazioti, P. (2008). Statistical analysis of inhalable (PM10) and fine particles (PM2. 5) concentrations in urban region of Patras, Greece. *Global NEST Journal, 10*(2), 123-131.

Namdeo, A., & Bell, M. (2005). Characteristics and health implications of fine and coarse particulates at roadside, urban background and rural sites in UK. *Environ Int, 31*(4), 565-573.

Nejadkoorki, F., & Baroutian, S. (2012). Forecasting extreme PM10 concentrations using artificial neural networks. *Int. J. Environ. Res, 6*(1), 277-284.

Ordieres, J., Vergara, E., Capuz, R., & Salazar, R. (2005). Neural network prediction model for fine particulate matter (PM 2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). *Environmental Modelling & Software, 20*(5), 547-559.

Schwartz, J. (2000). Assessing confounding, effect modification, and thresholds in the association between ambient particles and daily deaths. *Environmental health perspectives, 108*(6), 563.

Schwartz, J. (2001). Air pollution and blood markers of cardiovascular risk. *Environmental health perspectives, 109*(Suppl 3), 405.

Sharratt, B., & Edgar, R. (2011). Implications of changing PM10 air quality standards on Pacific Northwest communities affected by windblown dust. *Atmospheric Environment, 45*(27), 4626-4630.

Slini, T., Karatzas, K., & Papadopoulos, A. (2002). Regression analysis and urban air quality forecasting: An application for the city of Athens. *Global Nest, 4*(2-3), 153-162.

Taşpınar, F. (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).

Taşpınar, F., & Bozkurt, Z. (2014). Application of Artificial Neural Networks and Regression Models in the Prediction of Daily Maximum PM10 Concentration in Düzce, Turkey.

Thomas, S., & Jacko, R. B. (2007). Model for forecasting expressway fine particulate matter and carbon monoxide concentration: application of regression and neural network models. *Journal of the Air & Waste Management Association, 57*(4), 480-488.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *The Journal of Experimental Education, 70*(1), 80-93.

Ul-Saufie, A., Yahya, A., Ramli, N., & Hamid, H. (2012). *Future PM10 Concentration Prediction Using Quantile Regression Models.* Paper presented at the International Conference on Environmental and Agriculture Engineering, IACSIT Press, Singapore.

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N., & Hamid, H. A. (2012). 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.

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A., Rosaida, N., & Hamid, H. A. (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.

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

Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., & Kukkonen, J. (2011). Evaluation of a multiple regression model for the forecasting of the concentrations of NO x and PM 10 in Athens and Helsinki. *Science of the total environment, 409*(8), 1559-1571.

Yusof, N. F. F. M., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A., & al Madhoun, W. (2010). Monsoonal differences and probability distribution of PM10 concentration. *Environmental monitoring and assessment, 163*(1-4), 655-667.