Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 12 (14): 1488-1494, 2012
ISSN 1812-5654 / DOI: 10.3923/jas.2012.1488.1494

Performance of Multiple Linear Regression Model for Long-term PM_{10} Concentration Prediction Based on Gaseous and Meteorological Parameters

[1]Ahmad Zia Ul-Saufie, [2]Ahmad Shukri Yahaya, [2]NorAzam Ramli and [2]Hazrul Abdul Hamid
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara,
13500 Permatang Pauh, Pulau Pinang, Malaysia
[2]Clean Air Research Group, School of Civil Engineering, Engineering Campus,
Universiti Sains Malaysia, 14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia

Abstract: The aim of this study was to investigate performance of Multiple Linear Regression (MLR) method in predicting future (next day, next 2 days and next 3 days) PM_{10} concentration levels in Seberang Perai, Malaysia. The developed model was compared to multiple linear regression models. The model used gaseous (NO_2, SO_2, CO), PM_{10} and meteorological parameters (temperature, relative humidity and wind speed) as predictors. Performance indicators such as Prediction Accuracy (PA), Coefficient of Determination (R^2), Index of Agreement (IA), Normalized Absolute Error (NAE) and Root Mean Square Error (RMSE) were used to measure the accuracy of the models. Performance indicator shows next day (RMSE = 11.211, NAE = 0.124, PA = 0.927, IA = 0.960, R^2 = 0.858,) and next 2-day (RMSE = 14.652, NAE = 0.155, PA = 0.881, IA = 0.925, R^2 = 0.775) and next 3-day (RMSE = 15.611, NAE = 0.167, PA = 0.849, IA = 0.912, R^2 = 0.720). Assessment of model performance indicated that multiple linear regression method can be used for long term PM_{10} concentration prediction with next day for next day.

Key words: Regression models, PM_{10}, long term prediction, performance indicators, Malaysia

INTRODUCTION

Particulate Matter (PM) is one of the air pollutants and the most important in terms of adverse effects on human health. In Malaysia, there are three major sources of air pollution, namely mobile sources, stationary sources and open burning sources (Afroz *et al*., 2003). Several studies about the impacts of PM on human health were published (Alvim-Ferraz *et al*., 2005; Brunekreef and Holgate, 2002; Hoek *et al*., 2002; Kappos *et al*., 2004). PM_{10} concentration is more preferable than SPM for determining air pollution in Malaysia because Air Pollution Index (API) is obtained from the measurement of fine particles which is below 10 μm aero dynamic diameter of particles. Department of Environment, Malaysia established Malaysia Ambient Air Quality Guidelines in 2002 stating daily PM_{10} limit value is 150 μg m^{-3}, while annual PM_{10} value should not exceed 50 μg m^{-3} (Department of Environment Malaysia, 2002). When the PM_{10} concentration level exceed the limit values stated in air quality guidelines, short term and chronic human health problems may occur. Statistical modeling could offer good insights in predicting future air pollution levels (next day, next 2 days and next 3 days).

Multiple linear regression is easy for implementation and calculation. Many researchers used this method as forecasting tool in multiple disciplines. Chaloulakou *et al*. (2003) used this method to investigate the complex relationships between meteorological and time period parameters and forecast future PM_{10} concentrations. In Athens, Grivas and Chololokau (2006) used this method to predict hourly PM_{10} concentrations 24 h in advance and the result showed that multiple regression models can be used to predict PM_{10} 24 h in advance. In Malaysia, Ghazali (2006) used MLR for PM_{10} concentration level prediction and Ul-Saufie *et al*. (2011) compared MLR with feed forward back propagation for PM_{10} concentration prediction. However, both models cannot be used for future prediction.

The aim of this study was to investigate the performance of multiple linear regression method in predicting future (next day, next 2 days and next 3 days) PM_{10} concentration levels in Seberang Perai, Malaysia. Besides, this study also compared performance between

Corresponding Author: Ahmad Zia Ul-Saufie, Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara, 13500 Permatang Pauh, Pulau Pinang, Malaysia Tel: + (60)4-5996227 Fax: + (60)4-5941009

meteorological parameters with gases and meteorological parameter without gases as inputs. This model is useful because it facilitates respective authorities to carry out suitable actions to reduce the impact of air pollution.

MATERIALS AND METHODS

Site description: Seberang Perai, Pulau Pinang monitoring site is located at Taman Inderawasih (05°23.4704'N, 100°23.1977'E), at the north part of Peninsular Malaysia. This site is just a few kilometers from industrial area and surrounded by busy roads. Annual hourly observations for PM_{10} in Seberang Prai, Pulau Pinang, Malaysia from January 2004 to December 2007 were selected for PM_{10} concentration level prediction. The hourly observations were transformed into daily data by taking the average PM_{10} concentration level for each day. The chosen variables such as Relative Humidity (RH), Wind Speed (WS), nitrogen dioxide (NO_2), Temperature (T), carbon monoxide (CO), sulphur dioxide (SO_2) and previous day PM_{10} were selected to study their influences on PM_{10} concentration. The wind over country generally variable and light. Wind flow patterns can be described by four seasons namely north east monsoon known as wet seasons (November to March), transitional period (April to May), South-west monsoon, knows as dry seasons (June to September) and another transitional period (October to November). Average values for the chosen variables were 6.5 m sec^{-1} (WS), 28°C (T), 75.35% (RH), 0.0061 ppm (SO_2), 0.01334 ppm (NO_2), 0.4963 ppm (CO) and 67.24 μg m^{-3} (PM_{10}).

Multiple linear regression: Multiple linear regression is one of the modeling techniques used to investigate the relationship between a dependent variable and several independent variables. In multiple linear regression model, the error term denoted by ε is assumed to be normally distributed with mean 0 and constant variance σ. ε is also assumed to be uncorrelated.

We assume that the multiple linear regression model have k independent variables and there are n observations. Thus the regression model can be written as (Kovac-Andric *et al*., 2009):

$$y = b_0+b_1x_1+b_2x_2+b_3x_3+\ldots+b_{1k}x_{1k}+\varepsilon \text{ with } i = 1, \ldots, n \quad (1)$$

where, b_i are the regression coefficients, x_i are independent variables and ε is error associated with the regression. To estimate the value of the parameters, the least squares method was used.

VIF or variance inflation will be used for study effect of multicollinearity on the variance of estimated regression coefficients. The VIF is given by:

$$VIF_i = \frac{1}{1-R_i^2} \quad (2)$$

where, VIF_i is the variance inflation factor associated with the ith predictor and R^2_i is the multiple coefficient of determination in a regression of the ith predictor on all other predictors.

The Durbin-Watson (DW) statistic tests for autocorrelation of residuals. This test important to check that model assumptions is satisfied. The DW statistic is given by:

$$d = \frac{\sum_{i=1}^{n}(\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n}\hat{e}_i^2} \quad (3)$$

where, n n is number of observations, $\hat{e}_i = y_i-\hat{y}_i$ (y_i = observed values and $\hat{y}_i$ is predicted values. d is Durbin-Watson statistics and always between 0-4. A value d = 2 indicates no autocorrelation in the data, if values toward 0 indicates positive auto correlation and values approaching 4 indicates negative autocorrelation.

Performance indicators: Performance indicators were used to evaluate the goodness of fit for the MLR for future PM_{10} concentration prediction in Seberang Prai, Pulau Pinang. Performance indicators used to determine the best method in predicting PM_{10} concentration are NAE, RMSE, IA, PA and coefficient of determination (R^2) (Table 1).

Table 1: Performance indicators (Ul-Saufie *et al*., 2011)

Performance indicators	Equation	Description
Mean absolute error (MAE)	$MAE = \frac{\sum_{i=1}^{n}\lvert P_i - O_i\rvert}{n}$	MAE value closer to zero indicates better method
Normalized absolute error (NAE)	$NAE = \frac{\sum_{i=1}^{n}Abs(P_i - O_i)}{\sum_{i=1}^{n}O_i}$	NAE value closer to zero indicates better method
Index of agreement (IA)	$IA = 1-\left[\frac{\sum_{i=1}^{n}(P-O_i)^2}{\sum_{i=1}^{n}(\lvert P_i-\bar{O}\rvert+\lvert O_i-\bar{O}\rvert)^2}\right]$	IA value closer to 1 indicates better method
Prediction accuracy (PA)	$PA = \frac{\sum_{i=1}^{n}(P_i-\bar{O})^2}{\sum_{i=1}^{n}(O_i-\bar{O})^2}$	PA value closer to 1 indicates better method
Coefficient of determination (R^2)	$R^2 = \left[\frac{\sum_{i=1}^{n}(P_i-\bar{P})(O_i-\bar{O})}{n.S_{pred}.S_{obs}}\right]^2$	R^2 value closer to 1 indicates better method

RESULTS AND DISCUSSION

Multiple linear regression models were developed with 1428 (next day), 1427 (next 2 days) and 1426 (next 3 days) sets of data (average daily data from January 2004 to December 2007) using SPSS version 19.0. These years were selected due to limitation to access the data. Table 2 showed the summary model for PM_{10} concentration predictions based on gases and meteorological parameters. The result showed that all three future PM_{10} concentration prediction models showed no problems with multicollinearity as the value for Variance Inflation Factor (VIF) was lower than 10. Durbin Watson statistic showed that the summary model did not have any autocorrelation problem for next day (DW = 2.117), next 2 days (DW = 1.160) and next 3 days (DW = 1.043). Table 3 also showed the summary model predicting PM_{10} concentration based on meteorological parameters without gases. The result showed that the model did not imply multicollinearity (VIF = 1.257-1.870) and autocorrelation problem (DW = 0.900-2.152) with R^2 greater than 0.6.

PM_{10} level decreased during strong wind events because the strong wind dispersed the PM_{10} away. Negative correlation between temperature and PM_{10} was due to no significant temperature fluctuation in Malaysia (24-32°C). Similar results were found by Yusof *et al.* (2008). SO_2 had positive correlation with PM_{10} because most SO_2 in the area came from petrol fueled vehicle motor emissions. Besides that, SO_2 also came from industrial activities processing materials containing sulfur. For NO_2 and CO, the main sources for these two gases are diesel fueled vehicle emission. Our findings reflected negative correlation between these two gases and PM_{10} because there was less diesel fueled vehicle emission in this area.

Analysis of Variance (ANOVA) was conducted to test whether the models were significantly better at predicting the outcomes than using a mean. Table 4 showed the result for ANOVA (gases and meteorological parameters as input). The results indicated that observed values of F were 1243.152 (next day), 701.940 (next 2 days) and 503.147 (next 3 days) where the critical values $F_{0.05, 7, 1420}$, $F_{0.05, 7, 1419}$ and $F_{0.05, 7, 1419}$ were less than 2.103. From this result, all regression models were useful as predictors because the observed F ratios were four or five times greater than the critical values of F. Besides, it also indicated that the model significantly improved our capability to predict PM_{10} concentration. Similar conclusion was found in respect of applying meteorological parameters as inputs as shown in Table 5.

One of the assumptions for MLR was residuals (or errors) were normally distributed with zero mean and constant variances. Residual analysis was very important in determining the adequacy of the statistical model. If the error showed any pattern, the model was considered as not taking care of all the systematic information. Figure 1 and 2 showed that the residuals were normally distributed with zero mean for the models. Figure 3 and 4

Table 2: Model summary of PM_{10} based on meteorological parameters with gaseous

Models	R^2	Range of VIF	Durbin-Watson
Next day			
$PM_{10, t+1}$ = 12.50+0.95 PM_{10} -0.14 WS-0.1T+ 0.05 RH-589.91 NO_2-9.30 CO+172.06 SO_2	0.858	1.368-2.101	2.117
Next 2 days			
$PM_{10, t+2}$ = 39.9+0.9 PM_{10}-0.2 WS-0.8 T+ 0.1 RH-1046.9 NO_2-18.2 CO+305.2 SO_2	0.775	1.369-2.103	1.160
Next 3 days			
$PM_{10, t+3}$ = 53.3+0.9 PM_{10}-0.76 WS-1.1 T+ 0.2 RH-1298.9 NO_2-23.9 CO+433.7 SO_2	0.720	1.361-1.571	1.043

Table 3: Model summary of PM_{10} based on meteorological parameters without gaseous

Models	R^2	Range of VIF	Durbin-Watson
Next day			
$PM_{10,t+1}$ = -20.0412+0.1118 RH+ 0.9124 PM_{10}+0.4064 T+0.9256 WS	0.851	1.257-1.869	2.152
Next 2 days			
$PM_{10,t+2}$ = -19.9386+0.2156 RH+ 0.8549 PM_{10}+ 0.1801 T+1.2776 WS	0.748	1.257-1.870	1.165
Next 3 days			
$PM_{10,t+3}$ = -23.6811+0.3170 RH+ 0.8091 PM10+0.1740 T+1.1775 WS	0.676	1.261-1.763	0.900

Table 4: Result for ANOVA, gaseous and meteorological parameters as input

Model	Sum of squares	df	Mean square	F-value	Significance
Next day					
Regression	1071039	7	153005.5	1243.15	p<0.001
Residual	174771.7	1420	123.079		
Total	1245811	1427			
Next 2 days					
Regression	966611.6	7	138087.3	701.94	p<0.001
Residual	279149	1419	196.722		
Total	1245761	1426			
Next 3 days					
Regression	886100.8	7	126585.8	503.147	p<0.001
Residual	356752.1	1418	251.588		
Total	1242853	1425			

Table 5: Result for ANOVA, meteorological parameters as input

Model	Sum of squares	df	Mean square	F-value	Significance
Next day					
Regression	1061199.0	7	153005.5	1243.15	p<0.001
Residual	184612.1	1420	123.079		
Total	1245811.0	1427			
Next 2 days					
Regression	932788.2	7	138087.3	701.94	p<0.001
Residual	312972.4	1419	196.722		
Total	1245761.0	1426			
Next 3 days					
Regression	854337.5	7	126585.8	503.147	p<0.001
Residual	388515.3	1418	251.588		
Total	1242853.0	1425			

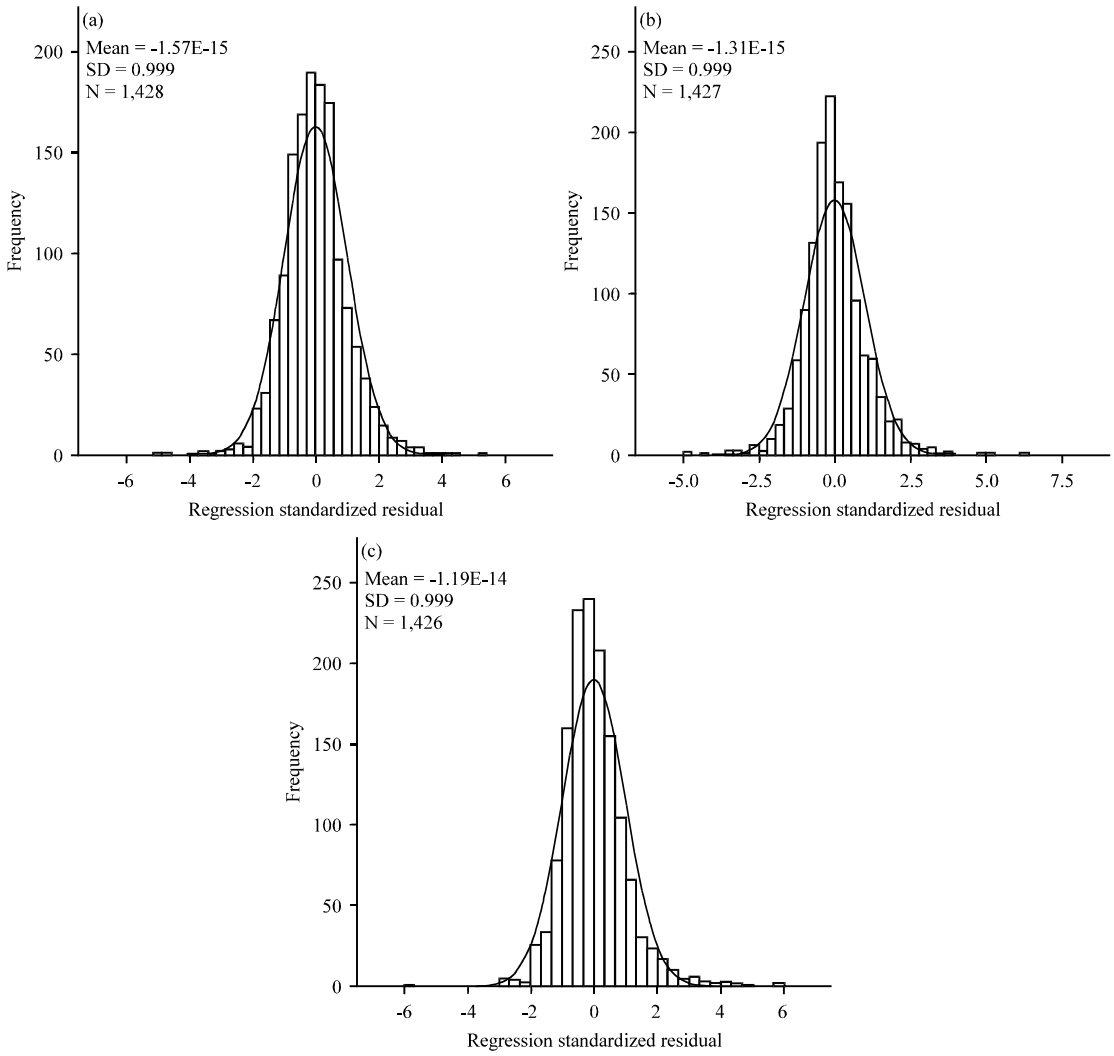

Fig. 1(a-c): Meteorological parameters based standardized residual analysis of PM_{10} for, (a) Next day, (b) Next 2 days and (c) Next 3 days

depicted that residuals were uncorrelated with constant variances as the residuals were contained in a horizontal band and hence obviously there were no defects in the models.

Comparison of performance: Performance indicators were used to compare performance for future prediction of PM_{10} concentration in Seberang Perai, Pulau Pinang. Table 6 showed the values for performance indicators. Accuracies measured were prediction accuracy, coefficient of determination and index of agreement, while the errors measured were normalized absolute error and root mean square error. The performance indicators reflected greater accuracy in next day PM_{10} concentration prediction compared to the next 2-day and next 3-day predictions. However, the result showed that MLR could predict future PM_{10} concentration until the next 3 days. Index of agreement with values greater than 0.9 indicated that the predicted values were highly accurate until the next 3 days. Table 6 also showed the comparisons between different parameters as inputs. The result

Table 6: Performance indicator for future PM_{10} concentration prediction.

PI	Next day[1]	Next day[2]	Next 2 days[1]	Next 2 days[2]	Next 3 days[1]	Next 3 days[2]
NAE	0.126	0.124	0.161	0.155	0.181	0.167
PA	0.923	0.927	0.865	0.881	0.823	0.849
R^2	0.851	0.858	0.748	0.775	0.676	0.720
RMSE	11.374	11.211	14.815	14.652	16.799	15.611
IA	0.959	0.960	0.923	0.925	0.895	0.912

1: Based on meteorological parameters (WS, T, RH) and PM_{10}, 2: Based on meteorological parameters (WS, T, RH), gaseous (CO, NO_2, SO_2) and PM_{10}

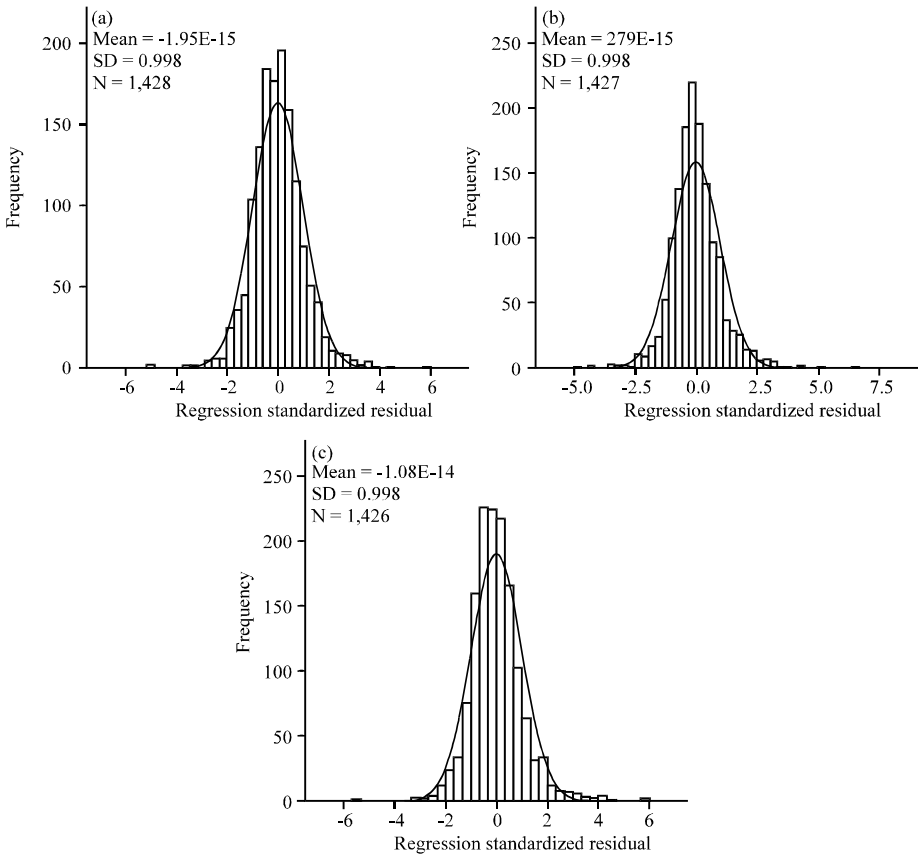

Fig. 2(a-c): Gaseous meteorological parameters based standardized residual analysis of PM_{10} for, (a) Next day, (b) Next 2 days and (c) next 3 days

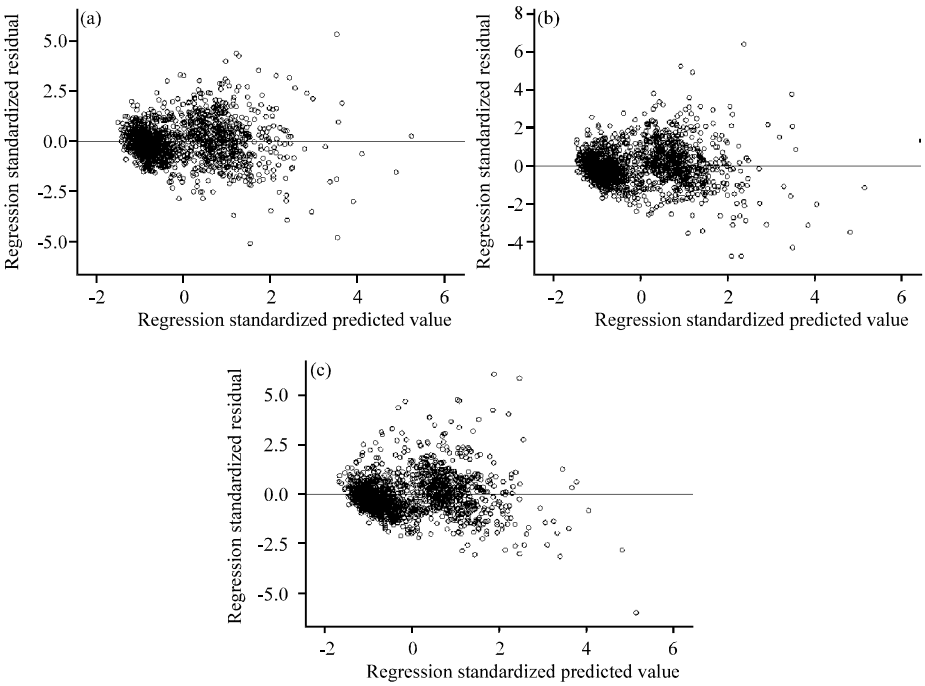

Fig. 3(a-c): Correlation of fitted values with residuals of PM_{10} for, (a) Next day, (b) Next 2 days and (c) Next 3 days

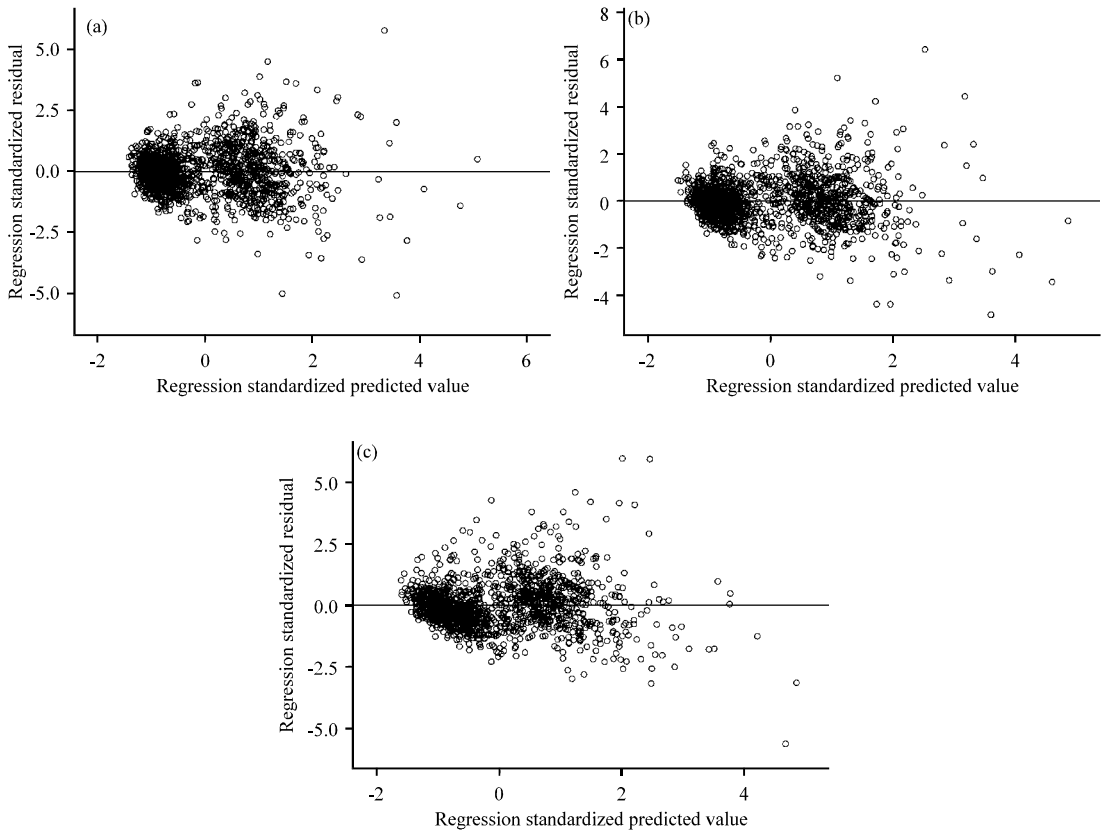

Fig. 4(a-c): Correlation of fitted values with residuals of PM_{10} for, (a) Next day, (b) Next 2 days and (c) Next 3 days

Table 7: Comparison results with other researcher using multiple linear regression

Area	Type of data	R^2	AI	References
Athens	Hourly	0.53-0.59	0.64-0.72	Grivas and Cholokou (2003)
Volos, Greece	Daily	0.55	0.86	Papanastasiou *et al.* (2007)
Athens and Helsinki	Daily	0.67-0.91	n/a	Vlachogianni *et al.* (2011)
Athens	Daily	0.5982	0.8656	Sfetsos and Vlachogiannis (2010)
Perai, Malaysia	Daily	0.720-0.858	0.912-0.960	Ul-Saufie *et al.* (2012)

showed meteorological parameters with gases as inputs performed better than meteorological parameter without gases. However, all the models could be utilized for PM_{10} concentration prediction as the values for prediction accuracy were greater than 0.8.

Various researcher have obtained multiple linear regression for predicting PM_{10} concentration. The result show that Coefficient of Determinations (R^2) were between 0.53-0.91 and Index of Agreement (IA) is from 0.64-0.86. Our result show that is close agreement between these obtained by previous researchers. Table 7 show comparison results with other researchers.

CONCLUSION

The result of fitting the best multiple linear regression models for PM_{10} concentration prediction using predictors such as air pollutants (NO_2, SO_2, CO and PM_{10}) and meteorological parameters (T, RH and wind speed). The result showed that using meteorological parameters with gases as inputs worked better than meteorological parameters without gases. The values of R^2, PA and IA would increase as more variables were added to the model. Similar conclusions were found by Mendenhall and Sincich (1995). Tree model predicting PM_{10} concentration had been successfully developed for next day, next 2 days and next 3 days.

The quality and reliability of the developed models were evaluated via performance indicators (NAE, RMSE, PA, IA and R^2). Assessment of model performance indicated that multiple linear regression method could be used for long term PM_{10} concentration predictions. The models could be easily implemented for public health protection by providing early warnings to the respective population. Besides, the models were useful in helping

authorities to reduce air pollution impact preventative measures in Seberang Perai, Malaysia.

ACKNOWLEDGMENT

This study was funded by Universiti Sains Malaysia under Grant 304/PAWAM/60311017. Thank you to Universiti Sains Malaysia and Universiti Teknologi MARA for providing financial support to carry out this study and also thanks to the Department of Environment Malaysia for their support.

REFERENCES

Afroz, R., M.N. Hassan and N.A. Ibrahim, 2003. Review of air pollution and health in Malaysia. Environ. Res., 92: 71-77.

Alvim-Ferraz, M.C., M.C. Pereira, J.M. Ferraz, A.M.C. Almeida e Mello and F.G. Martins, 2005. European directives for air quality: Analysis of the new limits in comparison with asthmatic symptoms in children living in the Oporto metropolitan area, Portugal. Hum. Ecol. Risk Assess.: Int. J., 11: 607-616.

Brunekreef, B. and S.T. Holgate, 2002. Air pollution and health. Lancet, 360: 1233-1242.

Chaloulakou, A., G. Grivas and N. Spyrellis, 2003. Neural network and multiple regression models for PM10 prediction in Anthens: A comparative assessment. J. Air Waste Manage. Assoc., 53: 1183-1190.

Department of Environment, Malaysia, 2002. Malaysia environmental quality report 2004. Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia, Kuala Lumpur, Malaysia.

Ghazali, N.A., 2006. A study to assess the effect of weather parameters in influencing the air quality in Malaysia. M.Sc. Thesis, Universiti Sains Malaysia, Malaysia.

Grivas, G. and A. Chaloulakou, 2006. Artificial neural network models for prediction of PM_{10} hourly concentrations in the Greater Area of Athens, Greece. Atmos. Environ., 40: 1216-1229.

Hoek, G., B. Brunekreef, B. Goldbohm, P. Fischer and P.A. van der Brand, 2002. Association between mortality and indicators of traffic-related air pollution in the Netherlands: A cohort study. Lancet, 360: 1203-1209.

Kappos, A.D., P. Bruckmann, P. Eikmann, N. Englert and U. Heinrich *et al*., 2004. Health effects of particles in ambient air. Int. J. Hygiene Environ. Health, 207: 399-407.

Kovac-Andric, E., J. Brana and V. Gvozdic, 2009. Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. Ecol. Infom., 4: 117-122.

Mendenhall, W. and T.L. Sincich, 1995. Statistics for Engineering and the Sciences. 4th Edn., Prentice-Hall Inc., New Jersey, USA., ISBN-13: 978-0023805813, Pages: 1008.

Papanastasiou, D.K., D. Melas and I. Kioutsioukis, 2007. Development and assessment of neural network and multiple regression models in order to predict PM_{10} levels in a medium-sized Mediterranean city. Water Air Soil Pollut., 182: 325-334.

Sfetsos, A. and D. Vlachogiannis, 2010. A new methodology development for the regulatory forecasting of PM10. Application in the Greater Athens Area, Greece. Atmospheric Environ., 44: 3159-3172.

Ul-Saufie, A.Z., A.S. Yahaya, N.A. Ramli and H.A. Hamid, 2011. Comparison between multiple linear regression and feed forward back propagation neural network models for predicting PM_{10} concentration level based on gaseous and meteorological parameters. Int. J. Sci. Technol., 1: 42-49.

Vlachogianni, A, P. Kassomenos, A. Karppinen, S. Karakitsios and J. Kukkonen, 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM_{10} in Athens and Helsinki. Sci. Total Environ., 409: 1559-1571.

Yusof, N.F.F.M., N.A. Ghazali, N.A. Ramli, A.S. Yahaya, N. Sansuddin and W. Al-Madhoun, 2008. Correlation of Pm_{10} concentration and weather parameters in conjunction with haze event in Seberang Perai, Penang. Proceedings of the International Conference on Construction and Building Technology, June 16-20, 2008, Kuala Lumpur, Malaysia, pp: 211-220.