IOP Conference Series:
Earth and
Environmental Science

PAPER • **OPEN ACCESS**

Modified linear regression for predicting ambient particulate pollutants (PM_{10}) during High Particulate Event

To cite this article: I A Mohd Jafri *et al* 2023 *IOP Conf. Ser.: Earth Environ. Sci.* **1216** 012002

Modified linear regression for predicting ambient particulate pollutants (PM_{10}) during High Particulate Event

I A Mohd Jafri[1,2], N Mohamed Noor[1,2]*, N A A A Rahim[1,2], S E Baidrulhisham[1], N Ramli[1,2], A Z Ul-Saufie[2,3] and G Deak[2,4]

[1]Faculty of Civil Engineering & Technology, Universiti Malaysia Perlis, Jejawi 02600, Perlis, Malaysia
[2]Sustainable Environment Research Group (SERG), Centre of Excellence Geopolymer and Green Technology (CEGeoGTech), Universiti Malaysia Perlis, Jejawi 02600, Perlis, Malaysia
[3]Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), Shah Alam 40450, Selangor, Malaysia
[4]National Institute for Research and Development in Environmental Protection INCDPM, Splaiul Independentei 294, 060031 Bucharest, Romania

*Corresponding author: norazian@unimap.edu.my

Abstract. Particulate Matter (PM_{10}) is one of the most significant contributors towards haze or high particulate event (HPE) that occurs in Malaysia. HPE can severely affect human health, environment and economic so it is important to create a reliable prediction model in predicting future PM_{10} concentration especially during HPE. Therefore, the aim of this study is to investigate the performance of modified linear regression models in predicting the next-day Particulate Matter (PM_{10+24}) concentration at two areas in the peninsular Malaysia namely, Bukit Rambai and Nilai. Hourly air quality dataset during historic HPE in 1997, 2005, 2013 and 2015 were used for analysis. Pearson correlation was used to select the input of the PM_{10} prediction model where only parameters with moderate ($0.6 > r > 0.3$) and strong ($r > 0.6$) correlation with PM_{10} concentration were selected as independent variables input in creating the multiple linear regression (MLR) model. The performance of modified linear regression model was evaluated by using several performance indicator which is Prediction Accuracy (PA), Index of Agreement (d_2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The results show that the modified MLR (parameter with $r > 0.6$ included as input) gave the best prediction model for the next-day PM_{10} concentration in both Bukit Rambai and Nilai.

1. Introduction

Air pollution in the form of haze repeatedly occurred in Southeast Asia and Malaysia has been listed as one of the most affected countries. In Malaysia, haze or high particulate event (HPE) often occurs either originated from local anthropogenic sources or the wildfire pollution from neighbouring country. The local anthropogenic sources originating from various sources such as industrial activities, increases uses of vehicles, biomass burning, and peat land fire had been identified as some of the reason for

regional haze. [1] also stated that peat land fire was one of the significant contributor to high particulate event in Southeast Asian. Aside from that, the strong influence of wildfire pollution from the neighbouring countries especially during the seasonal monsoon also becoming the main contributor to the HPE. Particularly, on the southwest part of peninsular Malaysia, the HPE was influenced by the southwest monsoon that is usually enhanced the biomass burning with the dry and hot weather during these season [2].

During HPE, large scale of extreme air pollution lead to higher mortality rates and cardiovascular diseases. A study conducted by [3] observed that the number of diagnostics of lung cancer patients have been dramatically increased during haze period from 2010 to 2015. Aside from the impact of these event towards health, economic sector in Malaysia also suffered a severe loss because of the reductions of output in both manufacturing and construction, and decreases in tourism earnings particularly due to flight cancellation [4].

High particulate event has caused an adverse effects towards the environment, human health and economic sector in Malaysia. Accompanied by the severe atmospheric condition, it will worsen the extend of the disaster. Therefore, it was important to create a valid model to predict air pollution levels in order to determine future concentrations of ambient particulate level especially during HPE. In order to achieve an efficient predictive model, this study used Pearson correlation to select the high and moderate correlated air pollutant parameters to be used as independent variables for Multiple Linear Regression (MLR) model. It was aimed to reduce the number of input variables and to improve the performance of the model. The performances of the prediction models were evaluated using several performance measures.

2. Methodology

2.1. Area of study

The area of study was focused at the southwest region of peninsular Malaysia where these areas were the most affected during HPE. Two monitoring stations located at the southwest of peninsular Malaysia namely Bukit Rambai and Nilai were chosen. Table 1 describe the location and background of the monitoring stations. Bukit Rambai and Nilai was located at the southern region of peninsular Malaysia which is strategically located in the rapid growth industrial areas. These stations were prone to the transboundary smoke from the Sumatera regions as these stations resides at the west coast of peninsular Malaysia.

Table 1. Specific location of the monitoring stations and background

Monitoring Station	Latitude (N)	Longitude (E)	Study Area
Bukit Rambai, Melaka	02°12.789’	102°14.364’	Industrial
Nilai, Negeri Sembilan	02°49.246’	101°48.877’	Industrial

2.2. Data Collection

Hourly data of air pollutants and meteorological parameters in the year that Malaysia experienced historic high particulate event (1997, 2005, 2013 and 2015) were chosen in this study. Table 2 shows the air pollutants and weather parameters that were obtained from Department of Environment (DOE), Malaysia.

Table 2. Air pollutants and meteorological parameters

Air Quality & Weather Parameters	Formula/ Abbreviation	Unit
Particulate Matter	PM_{10}	$\mu g/m^3$
Nitrogen oxides	NO_x	ppm
Sulphur dioxides	SO_2	ppm
Surface ozone	O_3	ppm
Nitrogen dioxides	NO_2	ppm
Carbon monoxide	CO	ppm
Temperature	T	^oC
Wind speed	WS	km/h
Relative humidity	RH	%

2.3. Data Pre-treatment

The missing observation of PM_{10}, other gases and weather parameters were first fill-in before the analysis were done. These missing data will be treated by using Linear Interpolation (LI) method using IBM SPSS Software Version 26. As stated by [5][6], it is important to fill in the missing data before any analysis because the success of the modelling depends on the quality of the dataset.

2.4. Pearson Correlation

The Pearson correlation coefficient was used to measure the correlation between PM_{10} concentration and other air pollutants and meteorological parameters. The degree of the correlation can be identify by the calculated Pearson correlation (r) value. If the correlation coefficient (r) is higher than 0.6, it was declared as high correlation; moderately correlation with the r-value between 0.3 and 0.6 and; weak correlated with the r-value smaller than 0.3 [7]. The parameters that has high to moderate values of r were used as independent variables input for developing prediction model using Multiple Linear Regression (MLR) model. Statistical Package for the Social Sciences (SPSS) version 26 was used to perform the correlation analysis and the formula for the correlation coefficient was as follows [7]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where x_i and y_i are the i-th sample values of PM_{10} measurements and other parameters respectively. $\bar{x}$, $\bar{y}$ are the mean value of PM_{10} measurement and other parameters correspondingly.

2.5. Multiple Linear Regression

Multiple linear regression (MLR) is one of the most used methods for forecasting. Firstly, the air pollution data was segregated into two sets of data which is 80% and 20% of data. 80% of the original data was used in developing the MLR model to predict the next-day PM_{10} concentration using SPSS version 26. The remaining 20% of the dataset was used as a reference data to evaluate the model's accuracy. MLR attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observe the data [8][9]**.** The general equation of MLR was shown in equation below[10]:

$$y = b_0 + \sum_i 1 b_i X_i + \varepsilon \quad (2)$$

Where, y is the dependent variable, b_0 is the value of y with all parameters set to 0, b_i is the regression coefficients, X_i is the independent variables and ε is the stochastic error.

2.6. Performance Indicator

In order to evaluate the performance of the regression model, several performance indicators such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Index of Agreement (d_2), and Prediction Accuracy (PA) was used to describe the goodness of fit for the predicting models of PM_{10} concentration. The performance indicator formulae is shown as in table 3 [11]:

Table 3. Performance Indicators and Formulae

Performance Indicator	Formula		Description
Mean Absolute Error (MAE)	$NAE = \sum_{i=1}^{n} \frac{Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i}$	(3)	MAE value closer to zero indicates better method
Root Mean Squared Error (RMSE)	$RMSE = \frac{1}{N} \sum_{i=1}^{N} \lvert P_i - O_i \rvert$	(4)	RMSE value closer to zero indicates better method
Index of Agreement ($\boldsymbol{d_2}$)	$d_2 = 1 - [\frac{\sum_{i=N}^{N} (P_i - O_i)^2}{\sum_{i=N}^{N} (\lvert P_i - \overline{O} \rvert + \lvert O_i - \overline{O} \rvert)^2}$	(5)	d_2 value closer to 1 indicates better method
Prediction Accuracy (PA)	$PA = \sum_{i=1}^{N} \frac{[(P_i - \overline{P})(O_i - \overline{O})]}{(N-1)\sigma_P \sigma_O}$	(6)	PA value closer to 1 indicates better method

3. Result and Discussion

3.1. Data Summary

Table 4 shows the data summary of hourly PM_{10} concentration for Bukit Rambai and Nilai during 1997, 2005, 2013 and 2015. In general, all respective year of study shows high value of maximum hourly PM_{10} concentration which exceeded the hourly Recommended Malaysia Ambient Air Quality Guideline (RMAAQG) which was 150 µg/m^3. The highest peak of PM_{10} concentration was recorded during the historic haze event in 1997 which was 515 µg/m^3 in Nilai and 510 µg/m^3 in Bukit Rambai. PM_{10} concentration for Bukit Rambai and Nilai also recorded having the highest standard deviation during 1997 with 61.59 µg/m^3 and 64.24µg/m^3 respectively due to high variability in dataset distribution.

Table 4. Descriptive Statistics Summary of PM_{10} concentration for Bukit Rambai and Nilai during 1997, 2005, 2013 and 2015

		Bukit Rambai				Nilai			
		1997	2005	2013	2015	1997	2005	2013	2015
N	Valid	8337	7427	8163	8759	8436	8757	8644	8620
	Missing	423	1333	597	1	324	3	116	140
Mean		71.70	57.13	57.24	69.69	64.16	63.18	58.17	69.90
Median		46.00	52.00	50.00	58.00	40.00	56.00	52.00	58.00
Std. Dev		61.59	20.05	43.75	41.47	64.24	33.57	29.74	44.02
Variance		3793.23	401.92	1913.70	1720.00	4127.38	1127.24	884.33	1938.00

Minimum	13.00	18.00	18.00	24.00	13.00	19.00	17.00	13.00
Maximum	415.00	178.00	515.00	338.00	510.00	330.00	304.00	353.00

3.2. Pearson Correlation

The relationship between PM_{10} concentration with the meteorological parameters and gaseous pollutants is given in Figure 1. Overall, Bukit Rambai and Nilai show high correlation of PM_{10} concentration with CO with r-value of 0.67 and 0.64 respectively. Both of the locations in this study are classified as industrial area. CO which mainly released by motor vehicle and machinery that used diesel fuel seem to have the strongest correlation with PM_{10} concentration. Besides, the seasonal fires from Indonesia can also be the main contribution to this correlation. It was reported by [12] that the seasonal fires was greatly inflated during El Niño and drought which causes large amounts of terrestrially-stored carbon into the atmosphere. Several moderate positive correlation was detected in Bukit Rambai which included PM_{10} - SO_2 concentration (r = 0.33), PM_{10} - Temperature (r =0.29) and PM_{10} - Humidity (-0.29). Bukit Rambai is situated next to Ayer Keroh industrial area and this area is also surrounded by a few coal-fired power plant.

In this study, parameters with moderate (0.6 > r > 0.3) and strong (r > 0.6) correlation with PM_{10} concentration were selected as independent variables input in creating the MLR model. Hence, from Figure 1, CO were selected as the strong correlation parameter whereas SO_2, RH and T were selected as moderate correlation parameters for Bukit Rambai. Meanwhile, only CO was selected as strong correlation parameter in Nilai.

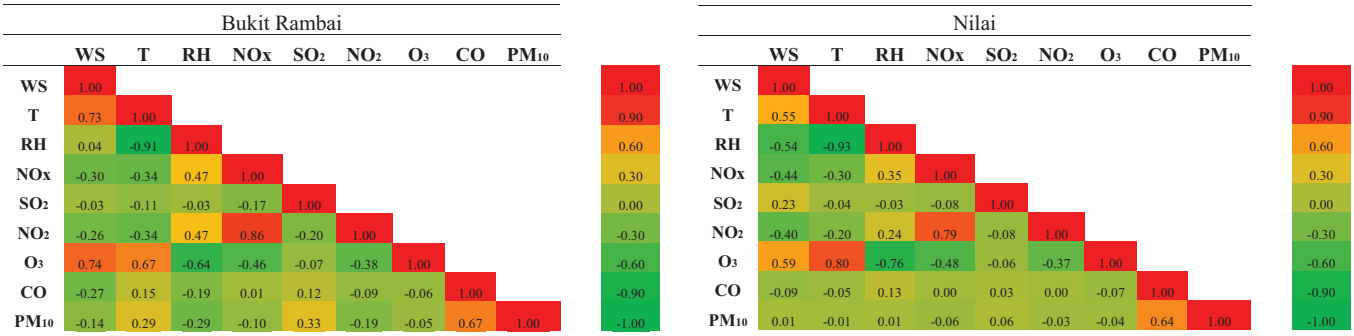

Figure 1. Pearson correlation coefficient matrix between PM_{10}, meteorological parameter and gaseous pollutants for Bukit Rambai and Nilai monitoring stations.

3.3. Prediction Model

Table 5 shows the linear equations for predicting the next-day PM_{10} (PM_{10+24}) concentration during HPE. Two classification of models were developed i.e. the modified MLR (only significant parameters from Pearson analysis were used as input) and MLR models (all parameters were included as input).

Table 5. MLR Model Summary of the next-day PM_{10} Concentration

Monitoring Station	Parameters	Equation
Bukit Rambai	MLR	$PM_{10+24} = 28.692 + 0.771PM_{10} - 0.275WS + 0.004T - 0.195H + 100.516NO_X - 29.012SO_2 + 53.149NO_2 - 17.09O_3 + 0.483CO$
	PC-MLR (H)	$PM_{10+24} = 12.243 + 0.742PM_{10} + 7.164CO$
	PC-MLR (H+M)	$PM_{10+24} = 35.212 + 0.767PM_{10} + 2.091CO + 6.596SO_2 - 0.188H - 0.274T$
Nilai	MLR	$PM_{10+24} = 46.019 + 0.627PM_{10} - 0.589WS - 0.363T - 0.268H + 19.742NO_X - 282.743SO_2 + 160.685NO_2 + 105.237O_3 + 12.472CO$
	PC-MLR (H)	$PM_{10+24} = 13.549 + 0.73PM_{10} + 5.198CO$

Where MLR – Multiple Linear Regression,
PC-MLR (H) – Parameter with high correlation with PM_{10} from Pearson Analysis
PC-MLR (H+M) – Parameter with high and moderate correlation with PM_{10} from Pearson Analysis

Table 6 shows the performance evaluation for the predicted values of next-day PM_{10} concentration (PM_{10+24}). From the four indicator applied, it was reflected the modified MLR models have a less error and greater accuracy compared to general MLR models for the next-day prediction. For Bukit Rambai, PC-MLR (H) model is selected as the best prediction model compared to others as it shows less error (RMSE and MAE value of 12.11 and 13.89 respectively) and higher value of d_2 and PA (0.98). Compared to MLR, PC-MLR (H+M) model gave better prediction in Bukit Rambai. The rank of performances of the prediction models in Bukit Rambai is given as PC-MLR (H) > PC-MLR (H+M) > MLR.

In Nilai, PC-MLR (H) model also outperformed compared to general MLR model with the value of d_2 and PA of 0.97. It can be concluded that the modified MLR (with high correlation parameters from Pearson analysis included as input) improved the prediction model for the next-day PM_{10} concentration in both Bukit Rambai and Nilai.

Table 6. The Performance Measure of the Prediction Model for Bukit Rambai and Nilai.

Location	Model	Performance Measure			
		RMSE	MAE	d_2	PA
Bukit Rambai	MLR	14.05	9.05	0.97	0.97
	PC-MLR (H)	**12.11**	**7.10**	**0.98**	**0.98**
	PC-MLR (H+M)	13.89	8.94	0.97	0.97
Nilai	MLR	18.67	12.55	0.94	0.86
	PC-MLR (H)	**13.89**	**7.95**	**0.97**	**0.97**

Where MLR – Multiple Linear Regression,
PC-MLR (H) – Parameter with high correlation with PM_{10} from Pearson Analysis
PC-MLR (H+M) – Parameter with high and moderate correlation with PM_{10} from Pearson Analysis
RMSE – Root Mean Squared Error, MAE – Mean Absolute Error, d_2 – Index of Agreement,
PA – Prediction Accuracy

Figure 2 shows the scatter plot of the observed and predicted value using the best selected prediction model (PC-MLR(H)) for the two locations. Scatter plot are purposely made to evaluate the strength of the relationship between predicted values and observed values. For Bukit Rambai, the modified MLR model that used high correlation parameter of (CO) has high R^2 which is 0.964. Furthermore, Nilai which also used high correlation parameter (CO) also show high value of R^2 which is 0.948. These high value of R^2 indicates that there is strong agreement between the predicted data and the observed data [13].

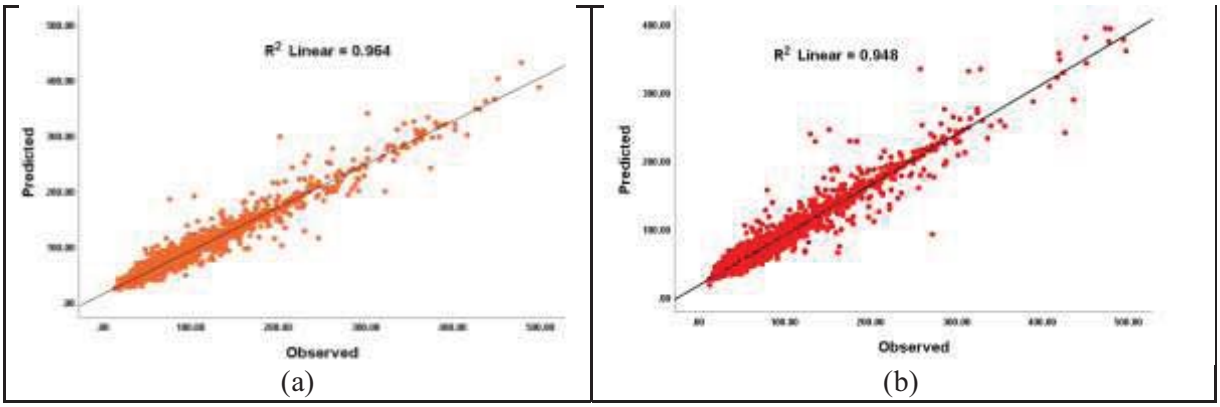

Figure 2. The scatter plot of the observed and predicted value by using PC-MLR (H) model for the prediction of the next-day PM_{10} concentration at (a) Bukit Rambai and (b) Nilai

4. Conclusion

Hourly air quality parameters in southern region peninsular Malaysia during the year of historic haze events (1997, 2005, 2013, 2015) were used to model the next-day PM_{10} concentration during HPE. The performances of the two types of MLR models i.e. the general MLR and the modified MLR model were developed and compared. The modified MLR model used only the significant parameters that correlated to PM_{10} concentration based on calculation using Pearson Correlation with the r value ranging from 0.3 to 1.0. The modified models were observed to give better performance compared to the general MLR model. Pearson correlation seem to be a good alternative method to reduce the number of input variables when developing MLR model. Using high correlation parameters as independent variables give a higher accuracy and less error compared to using moderate correlation parameters. These models is hope to be a useful predictive tool in providing the relevant and up to date method to the government authorities to strategize a suitable mitigation plan to reduce the impact of air pollution especially during HPE.

Acknowledgement

Author would like to thank the Ministry of Higher Education Malaysia for the FRGS/1/2020/TK0/UNIMAP/02/53 and the Department of Environment, Malaysia (DOE) for providing the air quality dataset.

References

[1] Harrison M E, Page S E and Limin S H 2009 The global impact of Indonesian forest fires *Biologist* **56** 156-163
[2] Abas M, Oros D R and Simoneit B R T 2004 Biomass burning as the main source of organic aerosol particulate matter in Malaysia during haze episodes *Chemosphere* **55** 1089–95
[3] Hassan A, Latif M T, Soo C I, Faisal A H, Roslina A M, Andrea Y L B and Hassan T 2017 Short communication: Diagnosis of lung cancer increases during the annual southeast Asian haze periods *Lung Cancer* **113** 1–3
[4] Varma A 2003 The economics of slash and burn: a case study of the 1997–1998 Indonesian forest fires *Ecological Economics* **46** 159–71
[5] Kwak S K and Kim J H 2017 Statistical data preparation: management of missing values and outliers *Korean J Anesthesiol* **70** 407
[6] Subasi A 2020 Introduction *Practical Machine Learning for Data Analysis Using Python* 1–26

[7] Awang N R, Elbayoumi M, Ramli N A and Yahaya A S 2016 Diurnal variations of ground-level ozone in three port cities in Malaysia *Air Qual Atmos Health* **9** 25–39
[8] Awang N R, Ramli N A, Yahaya A S and Elbayoumi M 2015 Multivariate methods to predict ground level ozone during daytime, nighttime, and critical conversion time in urban areas *Atmos Pollut Res* **6** 726–34
[9] Ul-Saufie A Z, Yahaya A S, Ramli N A and Hamid H A 2012 Performance of multiple linear regression model for long-term PM 10 concentration prediction based on gaseous and meteorological parameters *Journal of Applied Sciences* **12** 1488–94
[10] Abdullah S, Napi N N L M, Ahmed A N, Mansor W N W, Mansor A A, Ismail M, Abdullah A M and Ramly Z T A 2020 Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia *Atmosphere (Basel)* **11**
[11] Noor N M, al Bakri Abdullah M M, Yahaya A S and Ramli N A 2015 Comparison of linear interpolation method and mean method to replace the missing values in environmental data set *Materials Science Forum* **803** 278–81
[12] Huijnen V, Wooster M J, Kaiser J W, Gaveau D L A, Flemming J, Parrington M, Inness A, Murdiyarso D, Main B and van Weele M 2016 Fire carbon emissions over maritime southeast Asia in 2015 largest since 1997 *Sci Rep* **6**
[13] Noor N M, Ahmad Shukri Y, Nor Azam R and Mohd Mustafa Al Bakri A 2008 Estimation of missing values in air pollution data using single imputation techniques *Science Asia* **34** 341–5