Evaluation of Boosted Regression Tree for the Prediction of the Maximum 24-Hour Concentration of Particulate Matter

Wan Nur Shaziayani, Ahmad Zia Ul-Saufie, Syarifah Adilah Mohamed Yusoff, Hasfazilah Ahmat, and Zuraira Libasin

***Abstract*—Air pollution is a considerable health danger to the environment. The objective of this study was to assess the characteristics of air quality and predict PM_{10} concentrations using boosted regression trees (BRTs). The maximum daily PM_{10} concentration data from 2002 to 2016 were obtained from the air quality monitoring station in Kuching, Sarawak. Eighty percent of the monitoring records were used for the training and twenty percent for the validation of the models. The best iteration of the BRT model was performed by optimizing the prediction performance, while the BRT algorithm model was constructed from multiple regression models. The two main parameters that were used were the learning rate (lr) and tree complexity (tc), which were fixed at 0.01 and 5, respectively. Meanwhile, the number of trees (nt) was determined by using an independent test set (test), a 5-fold cross validation (CV) and out-of-bag (OOB) estimation. The algorithm model for the BRT produced by using the CV was the best guide to be used compared with the OOB to test the predicted PM_{10} concentration. The performance indicators showed that the model was adequate for the next day's prediction (PA=0.638, R^2=0.427, IA=0.749, NAE=0.267, and RMSE=28.455).**

***Index Terms*—Accuracy measures, air Pollution, boosted regression trees, PM_{10}, regression.**

I. Introduction

In Malaysia, air quality is monitored continuously throughout the country by the Department of Environment (DOE) at 65 stations. Afroz *et al*. [1] discussed air pollution caused by open burning and forest fires in Malaysia, which has become harmful to the public health and the environment. According to the [2], PM_{10} and O_3 are the major causes of unhealthy days recorded in Malaysia. PM_{10} is particulate matter with an aerodynamic diameter of less than 10 μm [3]. It is one of the main causes of pneumoconiosis, when it enters the bronchus, alveoli, and so on. The smaller the size of the dust particles, the deeper into the respiratory tract they enter [4].

Previously, many studies were conducted to predict future PM_{10} concentrations using a variety of methods. The multiple linear regression (MLR) method is the most common method used to predict PM_{10} concentrations. Juneng *et al*. [5] used the MLR method in their study to analyse the predictive relationship between the dependent variable (PM_{10}) and the independent variables. It was shown that local meteorological factors, particularly local surface air temperature, local humidity and local wind speed, dominate the fluctuations of PM_{10} over the Klang Valley during the summer monsoon. Moreover, Ul-Saufie *et al*. [6] used a quantile regression model to predict future (next day, next 2 days and next 3 days) PM_{10} concentration levels in Seberang Perai, Malaysia, and compared the results with the MLR. Despite the success of the MLR, according to [7], it presents problems in identifying the most important contributors when there is a high correlation or multicollinearity between the independent variables in the regression equation. Typically, one of the favoured techniques for predicting a complex system involves the use of artificial neural networks (ANN), such as the ANN model that was used by [8] to predict PM_{10} concentrations from the hourly data of a subway platform. According to [9], the predictive aspect of validation in the ANN model is not sufficient enough to fully assess the ability of the developed model to completely capture the underlying dynamics between independent and dependent variables.

BRTs are very reliable and flexible for dealing with complex responses, including interactions and nonlinearities [10]. The BRT algorithm is a single algorithm that is a combination of regression trees. The regression tree stops growing with repeated binary splits when certain criteria are met. In recent years, BRTs have been successfully implemented in air quality forecasting applications [11]-[14].Table I lists recent studies that have been conducted on air pollution in Malaysia. It shows that limited study have been conducted to predict PM_{10} concentrations using a BRT in Malaysia. A BRT works very well with large datasets and is robust with regard to missing values or outliers. Therefore, this study was conducted to predict PM_{10} concentrations using the BRT approach which had been developed by [15]. In contrast, this study used maximum daily data compared to hourly and averaged daily data that had been used by other researcher. Furthermore, this study used BRT to predict for the next day and it is different from BRT prediction that had been produced by [16].

Manuscript received November 12, 2020; revised January 22, 2021. This work has been carried out as part of the statutory activity of the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia. This research was funded by Malaysia Government under Fundamental Research Grant, grant number 600-IRMI/FRGS 5/3 (289/2019).

The authors are with the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 13500 Permatang Pauh, Malaysia (e-mail: shaziayani@uitm.edu.my, ahmadzia101@uitm.edu.my, syarifah.adilah@uitm.edu.my, zuraira946@uitm.edu.my, hasfazilah@uitm.edu.my).

TABLE I: RECENT STUDIES ON PM_{10} FORECASTING IN MALAYSIA

Author	Site	Data	Method	Accuracy
[16]	Peninsular, Malaysia (Klang, Kota Bharu, Kemaman and Perai)	2000-2013 (Hourly)	Boosted Regression Tree (BRT)	R^2=0.0075 - 0.9709
[17]	Nilai, Negeri Sembilan	2003-2010 (Average daily)	Multiple Linear Regression (MLR)	RMSE=14.23-23.03 R^2=0.35-0.62
			Artificial Neural Network (ANN)	RMSE=13.99-22.42 R^2=0.44-0.64
			Principal Component Analysis and Artificial Neural Network (PCA and ANN)	RMSE=11.10-15.64 R^2=0.46-0.78
[18]	Kuala Terengganu, Terengganu	2005-2011 (Average daily)	Multiple Linear Regression (MLR)	R^2=0.518
[19]	Pasir Gudang, Johor	2008-2010 (Daily)	Elman network (Elm)	R=0.70 MSE=0.0446 MAE= 0.1421
			Feed Forward Network (FNN)	R=0.70 MSE=0.0461 MAE= 0.1532
[20]	Kuching, Sarawak	2009-2015 (Hourly)	Artificial Neural Network (ANN)	R^2=0.9989941 RMSE=0.924999

II. MATERIALS AND METHODS

The main research site for this study was Kuching, Sarawak (Latitude: 1°36'27" N; Longitude: 110°22'42" E). Kuching is the capital city of Sarawak, and it has been classified by the DOE as an industrial area in the state. Sarawak is the largest state among the 13 states in Malaysia. It is located in northwest Borneo Island, and is bordered by the Malaysian state of Sabah to the northeast, Kalimantan (the Indonesian portion of Borneo) to the south, and Brunei to the north. The sampling station was named as the Kuching Air Monitoring Station (Latitude: 1°33' 44" N; Longitude: 110°23' 19" E). This area was selected to provide an overall representation and inference of the level of air quality in Kuching, Sarawak.

The parameter PM_{10}, CO, SO_2, NO_2, relative humidity (RH), temperature (T), and wind speed (WS) were used to predict for maximum 24-hour concentration of PM_{10}. The BRT model was fitted in the R version 3.4.2 software using the GBM (Generalized Boosted Regression Model) package version 1.6-3.1. The GBM offers three methods for estimating the optimal number of estimations, namely, the five-fold CV, independent test set (test), and out-of-bag estimation (OOB).

OOB assesses the decline of deviations from observations not used in selecting the next regression tree. Ridgeway *et al*. [21] states that the OOB use conservative methods to get the best iteration, as it underestimate the reduction of deviance. The advantage of this method is that the reduction of information available to study the structure of the model does not happened since it not eliminates a large set of independent data. According to Kohavi *et al*. [22], CV estimations of predictive performance may be erratic and repeated since it will do the cross validation according to the number folds in CV and then fit the final GBM model with number of iteration using all data. This study used five-fold CV. Lastly; the independent test set method uses a single holdout test set to select the optimal number of iterations. The disadvantage of this method is that the prediction uses a large number of observations, leaving a reduced data set to estimate the overall structure of the model. The steps for the BRT algorithm are summarized as follows [23]:

Input:

Data $\{(x_i, y_i)\}_{i=1}^{n}$ $x_i \in \Re^2$, $y_i \in \Re$ Loss Function $L(y_i, F(x))$

Output: Regression Tree $F_M(x)$

Step 1: Initialize model with a constant value:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

Step 2: For $m = 1,2,...,M$:

(A): For $i = 1,2,...,N$, Compute residuals:

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}$$

(B): Fit a regression tree to the residual $r_{i,m}$ values and create the leaf node area R_{jm} for $j = 1,2,...,J$.

(C): For $j = 1,2,...,J$ compute

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma)$$

(D): Update current model:

$$F_m(x) = F_{m-1}(x) + \upsilon \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$$

Step 3: Get the final output $F_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} \upsilon \gamma_{jm} I(x \in R_{jm})$

The steps of BRT algorithm involve of fit decision tree to the data and the loss function is used to appraise how well the prediction of a study. Step 2 is called weak classifier additive which also includes four steps. $r_{i,m}$ is the negative gradient of the i-th sample in the m-th as a number of tree. R_{jm} is a leaf node with the j-th is the number of leaf in the tree. It is a looping process that fit a regression tree to the residuals. This means that once the first tree is fitted to the model, it will take into account the prediction error of the tree to match the next tree, and to improve its accuracy. The learning rate (υ) for this study was set at 0.01. The three methods for estimating the optimal number of iterations test, OOB, and CV monitor the test data to stops improving beyond a certain number of

iterations.

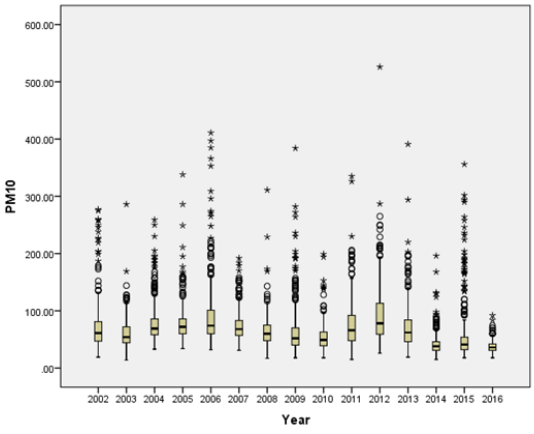
Performance indicators were used to determine the best BRT model from the three methods (OOB, test and CV) for future PM_{10} concentration predictions in Kuching, Sarawak. The models were validated by the root mean square error (RMSE), mean absolute error (MAE), index of agreement (IA), prediction accuracy (PA) and coefficient of determination (R^2). The equations used were reported by [24].

III. Results and Discussion

The model was developed using 80% of 3998 sets of data (3198) to forecast for the next day, while another 20% (800) were used to compare the performance for future predictions of PM_{10} concentrations in Kuching, Sarawak. Table II shows the descriptive statistics of the gaseous and meteorological parameters in Kuching, Sarawak. According to Uyanik *et al*. [25], if the skewness coefficient is a variable within the acceptable range of 1, the variable may not be said to be skewed. The PM_{10}, CO, NO_2, and SO_2 and relative humidity were highly skewed because their skewness coefficients were less than -1 or greater than +1, while the wind speed and temperature were moderately skewed since their skewness coefficients were between -½ and -1 or between +½ and +1. This showed that these variables were not normally distributed since the kurtosis coefficients differed greatly from the normal.

TABLE II: Descriptive Statistics for Maximum Daily Gaseous and Meteorological Parameters in Kuching, Sarawak

	WS	T	RH	SO_2	NO_2	CO	PM_{10}
Mean	11.69	33.18	94.35	0.36	1.31	90.36	67.67
Standard Deviation	2.67	2.60	3.01	0.26	0.55	50.07	39.97
Minimum	3.50	22.50	57.00	0.10	0.01	9.00	14.00
Maximum	28.40	43.00	100.00	2.40	12.30	508.00	526.00
Skewness	0.77	-0.69	-1.01	1.97	2.48	1.61	2.93
Kurtosis	1.57	0.77	6.74	6.10	36.75	4.79	14.66



	Mean	Standard Deviation	Minimum	Maximum	Skewness	Kurtosis
2002	71.989	43.752	19	277	2.503	7.3570
2003	60.581	25.651	14	286	2.617	16.566
2004	78.347	33.491	33	259	2.29	6.671
2005	32.389	1049.015	338	78.253	0.128	0.255
2006	91.304	56.063	32	411	2.870	10.478
2007	74.213	25.220	31	192	1.704	3.842
2008	64.445	26.762	17	311	3.343	23.554
2009	65.216	43.335	18	384	3.062	13.040
2010	53.907	22.750	18	199	2.500	10.347
2011	76.168	40.946	15	335	2.226	8.529
2012	91.407	50.488	26	526	2.699	15.528
2013	72.236	40.442	19	391	2.714	13.328
2014	41.923	18.963	15	196	3.537	19.742
2015	57.526	51.433	18	356	3.022	9.809
2016	37.591	10.074	18	92	1.446	3.622

Fig. 1. Descriptive statistics and box plots for maximum daily PM_{10} concentrations in Kuching, Sarawak.

Fig. 1 shows the box plots and descriptive statistics for the maximum daily PM_{10} concentrations in Kuching, Sarawak from 2002 to 2016. A box plot is a simple graphical display that is suitable for some important data features such as the central tendency, dispersion, skewness and potential outliers through three quartiles and the minimum and maximum observations [26]. The skewness values showed that the PM_{10} had an extreme event. The maximum value for the PM_{10} concentration was 526 mg m^{-3} (07 October 2012) while the minimum value was 14 mg m^{-3}. In 2012, the country experienced several short spells of haze due to transboundary air pollution as a result of forest fires in Central and Northern Sumatra, Indonesia. These had contributed to a slight deterioration in the overall air quality. For many years, the recorded PM_{10} concentrations in Kuching, Sarawak had been up to 200 mg m^{-3}, which is very unhealthy.

Fig. 2 shows the long-term monthly record of air quality data in Kuching, Sarawak. The emission of PM_{10} was found to be higher from August to October. According to [27], the inter-annual phenomenon known as the El-Nino Southern Oscillation (ENSO) is substantially related to an increase in sea surface temperatures. Thus, the rainfall rate over Southeast Asia is decreasing and will affect the Malaysia-Indonesia region, and anomalous easterly winds during August to October may enhance pyrogenic emissions across international borders from Kalimantan to Kuching, Sarawak. Incidents of open burning and forest fires increased due to ENSO, such as the 1997 forest fires in Borneo and Sumatra. The concentrations of CO, SO_2 and NO_2 were due to emissions from motor vehicles used by locals as well as tourists. The NO_2, CO and SO_2 profiles possessed a similar pattern throughout the years. Since an increase in relative humidity will lower the temperature in Kuching, Sarawak from November to December, it will also decrease the PM_{10} concentration.

Performance indicators were used to compare the performances for future predictions of PM_{10} concentration in

Kuching, Sarawak. Table III shows the values of the performance indicators. The accuracy measures used were the prediction accuracy, coefficient of determination, and index agreement, while the error measures used were the normalized absolute error and root mean square error. The results showed that the CV was the best method for estimating the optimal number of estimations in the BRT.

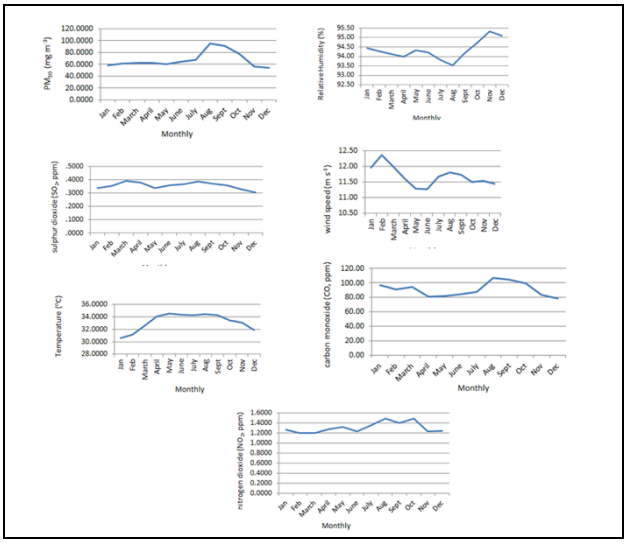


Fig. 2. Monthly trends of air quality in Kuching, Sarawak.

TABLE III: PERFORMANCE INDICATOR FOR FUTURE PM_{10} CONCENTRATION PREDICTIONS

MODEL	NAE	RMSE	IA	PA	R^2
OOB	0.2724	28.5867	0.7242	0.634	0.3708
CV	0.2673	28.4554	0.7492	0.6375	0.4269
test	0.2726	28.6046	0.7277	0.6329	0.3806

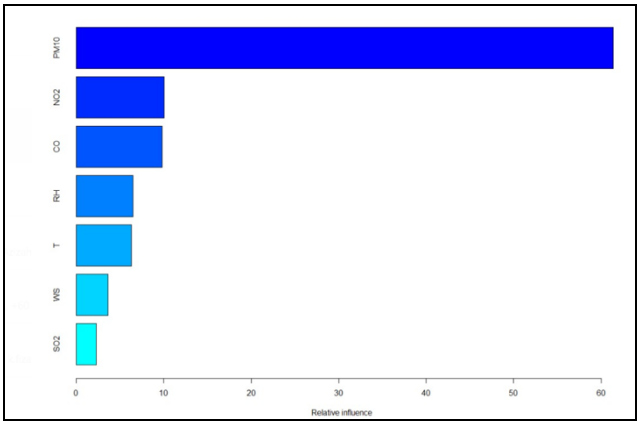


Fig. 3. Relative influence of significant variables.

As illustrated in Fig. 3, the most influential variable for predicting the maximum PM_{10} concentration for the next day was the PM_{10} concentration for the previous day, with 61.41%. Meanwhile, the second most influential variable would be NO_2 or CO, with 10%. The least significant influence for predicting the maximum PM_{10} was SO_2 with a maximum of only 2.32%. Therefore, the outcome would be more precise if the previous PM_{10} concentrations were used as one of the parameters to predict future PM_{10} concentrations. According to Chaloulakou *et al.* [28], the determination of the regression coefficient, R^2 was improved to 0.65 by using the previous day's PM_{10} concentrations as the extra input. In addition, Caselli *et al.* [29] found that the use of the PM_{10} concentrations for the previous day as independent variables for the prediction of PM_{10} concentrations give better results than a model without the previous day's PM_{10} concentrations for prediction models of PM_{10}. [30] also obtained a similar result as [28] and [29].

IV. CONCLUSION

This study has proved that BRT method can be used as alternative method to predict the maximum daily PM_{10} for the next day in Kuching, Sarawak. OOB, CV and test were used to determine number of tree. An assessment of the performance of the model verified that the CV gives a higher quality of prediction with a lower error NAE 0.2673, RMSE 28.4554 and with greater accuracy 0.7492 (IA), 0.6375 PA, and 0.4269 (R^2) compared with the OOB and test. Since this study used maximum daily data so it will be more relevant to help the government or authorities to provide early warning to the people about severe haze in the future.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization: Wan Nur Shaziayani, Ahmad Zia Ul-Saufie and Hasfazilah Ahmat; Formal analysis: Wan Nur Shaziayani, Syarifah Adilah Mohamed Yusoff and Zuraira Libasin; Methodology: Wan Nur Shaziayani, Ahmad zia Ul-Saufie and Syarifah Adilah Mohamed Yusoff; Supervision, Ahmad Zia Ul-Saufie and Hasfazilah Ahmat; Validation, Ahmad Zia Ul-Saufie and Syarifah Adilah Mohamed Yusoff; Writing – original draft: Wan Nur Shaziayani; Writing – review & editing: Ahmad Zia Ul-Saufie, Hasfazilah Ahmat and Zuraira Libasin. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENT

The authors are grateful to the Department of Environmental Malaysia (DoE) for providing necessary data for this research.

REFERENCES

[1] R. Afroz, M. N. Hassan, and N. A. Ibrahim, "Review of air pollution and health impacts in Malaysia," *Environ. Res.*, vol. 92, no. 2, pp. 71-77, 2003.

[2] Department of Environment, Ministry of Sciences, Technology and the Environment, Kuala Lumpur, Malaysia, "Malaysia environmental quality report," 2015.

[3] F. W. Han and L. Jian, "Flow field characteristics and coal dust removal performance of an arc fan nozzle used for water spray," *PLoS ONE*, vol. 13, no. 9, 2018.

[4] Q. Liu, W. Nie, Y. Hua, H. Peng, and Z. Liu, "The effects of the installation position of a multi-radial swirling air-curtain generator on dust diffusion and pollution rules in a fully-mechanized excavation face, a case study," *Powder Technol.*, vol. 329, pp. 371-385, 2018.

[5] L. Juneng, M. T. Latif, and F. Tangang, "Factors influencing the variations of PM10 Aerosol Dust in Klang Valley, Malaysia during the Summe," *Atmospheric Environment*, vol. 45, no. 26, pp. 4370-4378, 2011.

[6] A. Z. Ul-Saufie, A. S. Yahaya, N. A. Ramli, N. R. Awang, and H. A. Hamid, "Future PM_{10} concentration prediction using quantile regression models," in *Proc. 2nd International Conference on Environmental and Agriculture Engineering*, Singapore, vol. 37, 2012, pp. 15-19.
[7] S. A. Abdul-Wahab, C. S. Bakheit, and S. M. Al-Alawi, "Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations," *Environmental Modelling & Software*, vol. 20, no. 10, pp. 1263-1271, 2005.
[8] S. Park, M. Kim, H. G. Namgung, K. T. Kim, K. H. Cho, and S. B. Kwon, "Predicting PM_{10} concentration in Seoul metropolitan subway stations using artificial neural network (ANN)," *Journal of Hazardous Materials*, vol. 341, pp. 75-82, 2018.
[9] S. M. Cabaneros, J. K. Calautit, and B .R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," *Environmental Modelling & Software*, vol. 119, 2019, pp. 285-304.
[10] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, 2008, pp. 802–813.
[11] S. M. Cabaneros, J. K. Calautit, and B. R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," *Environmental Modelling and Software*, vol. 119, pp. 285–304, 2018.
[12] N. Z. Yahaya, Z. F. Ibrahim, and J. Yahaya, "The used of the boosted regression tree optimization technique to analyse an air pollution data," *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, pp. 1565–1575, 2019.
[13] P. D. Ivatt and M. J. Evans, "Improving the prediction of an atmospheric chemistry transport model using gradient boosted regression trees," *Atmos. Chem. Phys. Discuss*, vol. 20, issue 13, 2019.
[14] N. Z. Yahaya, S. M. Phang, A. A. Samah, I. N. Azman, and Z. F. Ibrahim, "Analysis of fine and coarse particle number count concentrations using boosted regression tree technique in coastal environment," *Environment Asia*, vol. 11, no. 3, 2018, pp. 221–234.
[15] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, 2001, pp. 1189–1232.
[16] N. Z. Yahaya, Z. F. Ibrahim, and J. Yahaya, "The used of the boosted regression tree optimization technique to analyse an air pollution data," *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, pp. 1565–1575, 2019.
[17] A. Z. Ul-Saufie, A. S. Yahaya, N. A. Ramli, N. Rosaida, and H. A. Hamid, "Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA)," *Atmospheric Environment*, vol. 77, pp. 621–630, 2013.
[18] S. Abdullah, M. Ismail, and S. Y. Fong, "Multiple linear regression (MLR) models for long term PM10 concentration forecasting during different monsoon seasons," *Journal of Sustainability Science and Management*, vol. 12, no. 1, pp. 60–69, 2017.
[19] A. Afzali, M. Rashid, B. Sabariah, and M. Ramli, "PM10 Pollution: Its predictionand meteorological influence in Pasir Gudang, Johor," presented at 8th International Symposium of the Digital Earth (ISDE8), 2014.
[20] N. L. Zakri, S. M. Saudi, A. Juahir, H. E. Toriman, M. F. Abu, I. M. M. Muaz, and K. M. Feroz, "Identification source of variation on regional impact of air quality pattern using chemometric techniques in Kuching, Sarawak," *International Journal of Engineering & Technology*, vol. 7, no. 3, 14, p. 49, 2018.
[21] G. Ridgeway, "Generalized boosted models: A guide to the gbm package," *Update*, vol. 1, no. 1, 2019.
[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–114.
[23] M. Gong, Y. Ai, J. Qin, J. Wang, P. Yang, and S. Wang, "Gradient boosting machine for predicting return temperature of district heating system: A case study for residential buildings in Tianjin," *Journal of Building Engineering*, vol. 27, p. 100950, 2020.
[24] H. C. Lu, "Estimating the emission source reduction of PM10 in central Taiwan," *Journal of Chemosphere*, vol. 54, no. 7, pp. 805-814, 2004.
[25] G. K. Uyanik and N. Guler, "A study on multiple linear regression analysis," *Procedia-Social and Behavioral Sciences*, vol. 106, pp. 234–240, 2013.
[26] C. J. Changa, D. C. Li, Y. H. Huang, and C. C. Chen, "A novel gray forecasting model based on the box plot for small manufacturing data sets," *Applied Mathematics and Computation*, vol. 265, pp. 400–408, 2015.
[27] K. Yusof, A. Azid, M. S. Samsudin, and J. M. Asrul, "An overview of transboundary haze studies: The underlying causes and regional disputes on Southeast Asia region," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 13, no. 4, 2017, pp. 747-753.
[28] A. Chaloulakou, P. Kassomenos, N. Spyrellis, P. Demokritou, and P. Koutrakis, "Measurements of PM10 and PM2.5 particle concentrations in Athens, Greece," *Atmospheric Environment*, vol. 37, pp. 649-660, 2003.
[29] M. Caselli, L. Trizio, and G. D. Gennaro, "A simple feedforward neural network for the PM10 forecasting: Comparison with a radial basis function network and a multivariate linear regression model," *Water Air and Soil Pollution*, vol. 201, pp. 365–377, 2009.
[30] A. Z. Ul-Saufie, A. S. Yahaya, N. A. Ramli, N. Rosaida, and H. A. Hamid, "Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA)," *Atmospheric Environment*, vol. 77, pp. 621–630, 2013.

Wan Nur Shaziayani is a Ph.D student in Universiti Teknoligi MARA, Malaysia in the field of statistics. She holds a master's degree in mathematics from the Universiti Tenkologi Malaysia. She also holds a bachelor of science degree in statistics from the Universiti Teknologi MARA, MALAYSIA. Her research interest areas are applied statistics, environmental modelling and education research.

Ahmad Zia Ul-Saufie Mohamad Japeri is an associate professor in Universiti Teknologi Mara. His research interest is on applied statistics in environment such as air pollution modelling, environmental modelling, machine learning and survey research.

Syarifah Adilah Mohamed Yusoff is a senior lecturer in Universiti Teknologi MARA. She holds a PhD degree in computer sciences from Universiti Sains Malaysia. Her research interest is on computer sciences such as optimization, machine learning and education research.

Hasfazilah Hamat is a senior lecturer in Universiti Teknologi MARA. She holds a PhD degree in air quallty from Universiti Sains Malaysia. Her research interest is on applied statistics in environment such as air pollution modelling, environmental modelling and survey research.

Zuraira Libasin is a senior lecturer in Universiti Teknologi MARA, Malaysia in the field of statistics. She holds a master's degree in mathematics from the Universiti Tenkologi MARA. Her research interest areas are applied statistics, environmental modelling and education research.