The used of the Boosted Regression Tree Optimization Technique to Analyse an Air Pollution data.

Noor Zaitun Yahaya, Zul Fadhli Ibrahim, Jamaiah Yahaya

Abstract*: The stochastic boosted regression trees (BRT) technique has the capability to quantify and explain the relationships between explanatory variables. We applied this machine learning modelling technique to derive the relationships between the gases air pollutants, meteorological conditions and time system variables of particulate matter (PM_{10}) concentrations. In order to get lowest prediction error and to avoid over-fitting, the parameters of the BRT model need to be tuned. In this experiment, 25 BRT models were generated from 14 years' worth of hourly data (122,736 a one hour averaged data from January 2000 to December 2013 gathered from four Continuous Automated Air Quality Monitoring Stations in peninsular Malaysia (located in Klang, Selangor (CA0011), Perai, Penang (CA0003), Kota Bharu, Kelantan (CA0022) and Kemaman, Terengganu (CA0002)). Seventy percent of the data were used for training and 30 percent for validation of the models. An experiment was conducted to determine the best iteration that could model hourly PM_{10} concentrations by optimizing the BRT parameter which are learning rate (lr), tree complexity (tc) and number of trees (nt). Five different lr (0.001, 0.005, 0.01, 0.05 and 0.1) were tested with different tree complexities (1 to 20) in the BRT model development process. From the experiment, the combination of lr = 0.05 and tc = 5 for the training set for the BRT model achieved the lowest root mean squared error (RMSE) compared to the other tested combinations. It was also found that the number of trees increased with the increment in the number of samples. A high coefficient of determinant (R^2) value (0.90) for the linear relationship between the number of samples and nt was found for all the four stations. The optimum number of trees for the model was estimated by using 10-fold cross-validation. It was found that the best number of iterations for Klang, Perai, Kota Bahru and Kemaman were 12,327, 32,987, 16,370 and 57,634, respectively. The prediction accuracy of the model was tested by using the fraction of prediction namely a factor of two (FAC2), mean bias, mean gross error, RMSE, correlation coefficient (R), and index of agreement (IOA). The prediction performance of the final BRT model based on the R value was 0.81, 0.78, 0.85 and 0.81 for for Perai, Kemaman, Klang and Kota Bahru, respectively, which indicates that the BRT model developed and applicability of this can be used in other atmospheric environment data.

***Index Terms*: Boosted Regression Tree, Tuning parameters, Hourly PM_{10} model, particulate matter**

I. INTRODUCTION

Air pollution issues is a major concern that needs to be given immediate and serious attention by all relevant authorities around the globe because clean air is a basic requirement of human health and well-being [1]. Particulate matter contains microscopic liquid droplets or solids that can travel deep into human lungs and cause serious health problems. Numerous scientific studies link exposure to particulate pollution to premature death in people with heart or lung disease, non-fatal heart attacks, irregular heartbeat, aggravated asthma, decreased lung function, and increased respiratory symptoms such as irritation of the airway, coughing and breathing difficulties [2].

In Malaysia, the most prominent source of PM_{10} pollution in the country is motor vehicle emissions followed by industrial emission source [3,4], mobile emission sources contributing to at least 70%–75% of the total air pollution, while stationary sources generally contribute around 20%–25% and open burning and forest fires contribute approximately 3%–5% [5]. Several unhealthy days have been reported by the Department of Environment (DOE) Malaysia [6] at various locations in Klang Valley and other big cities in other states as a resulted of motor vehicle emissions.

Various statistical tools have been used in PM_{10} prediction and forecasting studies in Malaysia and elsewhere. One of the modelling techniques that has been used to predict PM_{10} concentration is the robust regression method, which is seen as a good method because it is able to minimize the influence of outliers in datasets compared to the traditional least squared method [7]. The most popular methodology is multiple regression analysis (MRA) to express response of dependent variable of multiples independent variables. Yet, MRA suffers from an inherent drawback in that it is difficult to identify which independent variable has the most influence on the dependent variable when there is multi-collinearity between the independent variable [8].

In contrast, multivariate data analysis techniques such as multiple linear regression (MLR) and principal component analysis (PCA) have been proven to be effective tools in studying the relationship between voluminous data such as air pollution and meteorological records[9], 10].Principal component analysis is the simplest of the true eigenfactor-based multivariate analyses; it is used to reduce the number of predictive variables and transform them into new variables that are mutually orthogonal, or uncorrelated, as well as to determine the dominant multivariate relationships[8] ,9]. On the other hand, MLR is frequently used to explain meteorological components because it allows the formation of explicit equations that are less complex.

Revised Manuscript Received on November 15, 2019

Noor Zaitun Yahaya, Senior Lecturer, School of Ocean Engineering, University Malaysia Terengganu, Terengganu, Malaysia

Zul Fadhli Ibrahim, Researcher, School of Ocean Engineering, University Malaysia Terengganu, Terengganu, Malaysia

Jamaiah Yahaya, Assoc. Professor, School of Informatic Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

While many statistical models have been developed to investigate and quantify the sources that contribute to PM_{10} emissions; the statistical approach is limited in terms of the ability to explain the factors that influence PM_{10} concentration due to the statistical assumptions that need to be made and the homogeneity of the data. To address this issue, recent studies have attempted to develop powerful computational intelligence models by using machine learning algorithms such as the neural network that can be trained to predict the complex PM_{10} concentration system and have demonstrated that such models can quickly predict the desired value [11], [7]. However, the neural network is a black box technique where there is no deep understanding of physical characteristics [12].

In order to address the above issues, Friedman [13] developed a computational intelligence modelling technique based on boosted regression trees (BRT) to model complex multivariable interactions and non-linear effects, which are common in an air pollution data. Initially, the gradient boosting algorithm developed, which called for all of the training data observations to be included in the function estimation process at each iteration. No matter how dimensionally large the predictor variable space is, or how many variables are used for the prediction, the model subcomponents can be represented by a two-dimensional graphical representation which can be easily plotted and interpreted [13].

Then, in 2002, Friedman added a stochastic element to the initial boosting algorithm that involved taking a random sample of observations for each iteration of *t.* Thus, the performance of the initial technique was improved by adding an element of randomness to the algorithm and creating the stochastic gradient boosting machine (GBM) or stochastic BRT [14]. This improvement involved taking subsamples of training data – between typically 40%–60% for each iteration – by indicating the percentage of the training data in the algorithm. Since then, the term 'stochastic gradient boosting' (SGB) has been simplified to 'gradient boosting' or more simply 'boosted trees'. Hereinafter, the term BRT is used to denote stochastic gradient boosting using least squares regression trees.

The BRT modelling technique has now become an important technique for modelling single response variables using several predictors [15]-[20]. Unlike a black box technique, the BRT approach is able to examine how the dependent variables respond to individual model variables. Thus, the interactions between the variables can be determined, ranked and visualized. There are several techniques that aim to improve the performance of a single model by fitting many models and combining them for prediction. 'Decision tree learning' or 'decision trees' is a tool in machine learning and/or data mining or big data analysis which maps observations about a certain item to conclusions about a certain item's target value and explained that the advantages of the decision-tree-based technique are its flexibility in handling a wide range of response types, ability to handle missing values in both response and explanatory variables and to rank statistics [17], [21].

In a recent study, the characteristics of the BRT technique were found to offer advantages over other methods such as linear regression and multiple regression analyses in the context of air pollution modelling [21]. Also, Yahaya et al. (2013) applied the BRT technique to explore air pollution data (particle number count (PNC) concentrations and gases), meteorological and traffic data in the City of Leeds in the UK in which promise a good results to elaborate the model fitting, variables influence most and also interaction between variables [22]. Three methods can be used to estimate the optimal number of iterations through the fitted GBM*:* the independent test set (test), out-of-bag estimation (OOB), and cross-validation [21],[18]. In a recent study, [23] extended the model developed by [22] to model nitrogen oxides (NO_x), meteorological variables and traffic variables from the City of Leeds UK data. Results show an agreement that BRT technique showing a good result to visualize BRT output graphically in term of the partial dependence plots [23].

The ability of the BRT model to give in-depth information about the relationship between the response variable and predictor variables give a great insight into the PM in ambient air was demonstrated by a study undertaken in the City of Birmingham, UK [24]. Also, in more recent research [19], [20] applied the BRT model to determine the relationship between ground level ozone and [PNC] data in several continuous air quality stations located in a coastal environment in Malaysia, in which the Pearson's correlation coefficient, the R value and the coefficient of determination (R^2) between the estimator variance and modelled data indicated that the performance of the model was a good fit. A correlation coefficient value approaching 1 showed a better model performance between the two variables (ozone and [PNC]). The correlation of determination (R^2) values for fine and coarse particles were 0.90 (0.81) and 0.94 (0.87), respectively, which indicated that both the observed and modelled number counts for both the fine and coarse particles were in good correlation with each other [20]. These results were similar to those obtained in a previous study by [19], thereby demonstrating the reliability of the BRT technique for air quality studies. In the BRT, there are five tuning parameters that need to be controlled (in addition to the distribution): the training sample size relative to the training population (*bag.fraction*), the number of iterations (*nr*), the learning rate (*lr*), the maximum tree depth (*interaction depth*), and the number of observations in each terminal nod. For each iteration, only the random subset of the residuals is used to build the tree. The bag-fraction signifies the fraction of observations in the training data to sample for each iteration. The fraction of the training set observations is randomly selected to propose the next tree in the expansion. In this paper, *lr*s of 0.1 to 0.001 are tested with different *tc*s and the optimum number of trees is estimated by using the 10-fold cross-validation method in order to determine the optimum tuning parameters of the BRT for different air quality datasets from stations located in peninsular Malaysia.

II. METHODOLOGY

A. Study Sites and Data Used

Four air quality monitoring station sites in peninsular Malaysia were selected for this study, each with different background and meteorology characteristics.

All four stations are operated by the DOE Malaysia. The CA0011 (Klang, Selangor) and CA0022 (Kota Bharu, Kelantan) stations are in the urban background category, whereas the CA002 (Kemaman, Terengganu) and CA003 (Perai, Penang) stations are categorized as industrial background stations. Table 1 provides a summary of the characteristics of each station.

Table 1: Characteristics of monitoring station sites

Stn. ID	Latitude	Longitude	Cat.
CA011	3°0'41.77"N	101°24'3.23"E	Urban
CA022	6°6'28.42"N	102°15'5.01"E	Urban
CA003	5°23'21.78"N	100°23'11.32"E	Ind.
CA002	4°15'55.32"N	103°25'50.15"E	Ind.

The two urban stations, the CA0011 or Klang station and the CA0022 or Kota Bharu station are located in cities on the west coast of peninsular Malaysia and the east coast of Malaysia, respectively. Based on data from 2010 census, Klang city has a city area of 626.8 km^2 and a population of 240,016, while Kota Bharu has a population of 44,757 and a size is 394 km^2. It should be noted that the weekend of these two cities falls on different days: Saturday and Sunday for Klang on the west coast and Friday and Saturday for Kota Bharu on the east coast which that also need to take into considerations.

The two industrial stations, the CA002 or Kemaman station and the CA0003 (Perai station) are located in the states of Penang and Terengganu, respectively, The CA0002 station is located at SK Bukit Kuang, Kemaman in the southwest of the Teluk Kalong industrial area. The second station, CA0003 station is located at Sek. Keb. Cenderawasih, Tmn Inderawasih, Bandar Seberang Jaya, Perai.

This study used the hourly air pollution and meteorological data gathered by the four air quality monitoring stations over a period of 14 years from 2000 to 2013 (a total of 122,736 hourly observations). Data were obtained on the following pollutants and meteorological parameters: PM_{10}, nitric oxide (NO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), temperature, relative humidity, wind speed and wind direction. The monitoring of PM_{10} was performed by using the Beta Attenuation Method Model 1020 from Met-One Instruments, Inc, USA. Instruments from Teledyne Technology Inc., USA model 100A/100E and the Teledyne API Model 300/300E, were used to monitor SO_2 and CO, respectively. The Teledyne API Model 200/200E was also used to monitor the NO and NO_2 concentrations.

B. Data Summary

The PM_{10}, SO_2, NO, NO_2 and CO concentrations were observed for the period January 2000 to December 2013 for each study site as well as the metrological parameters. All the data measurements were averaged by using a similar averaging time of 1 hour for comparison purposes. The data measurements were processes by referring to quality assurance and standard ratification techniques for gases [6]. Also a data screening process was conducted that involved the deletion of several spur data. Every outlier and extreme value were analysed by time series plot and the spur value was detected and deleted before further analyses were conducted. Figure 1 illustrates the time series plots for the 1-hour averaged data for PM_{10} for all four stations.

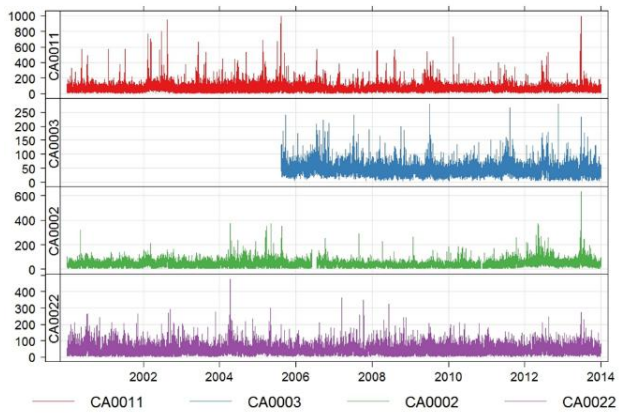


Fig 1: Hourly PM_{10} concentrations for CA0011 (Klang), CA0003 (Perai), CA0002 (Kemaman) and CA0022 (Kota Bharu) from January 2000 to end of December 2013

C. BRT Model Development

A BRT algorithm was fitted to the hourly PM_{10} concentrations for each dataset in order to investigate how ambient PM_{10} concentrations are influenced by ambient gases (NO, NO_2, CO, SO_2), meteorological conditions (wind speed, wind direction, temperature, humidity) and time system (hour of the day, weekday and Julian day (day of the year)). The change in the dependent variable, PM_{10}, is related to the above-mentioned independent variables, as well as traffic data (flow and speed). Wind speed and direction, humidity and temperature have been shown to be the most dominant meteorological parameters so they were included in the simulations. The time system variables, namely hour of the day (t_{hour}), day of the week ($t_{weekday}$), day of the year (t_{jd}) have also been found to be important for model development in air quality studies, so these parameters were also included in the simulations. The use of the hour of the day (0–23) was intended to capture the diurnal emission trend of possible sources at the monitoring site, which are mostly related to motor vehicle emissions. The weekday was included because it influences the level of mobility of people as they go to work or school and undertake other activities, which in turn has an effect on the level of emissions such as motor vehicle emissions.

A weekday is defined as any of the days on which most people go to work or school, which is Monday to Friday for Klang and Perai, and Sunday to Thursday for Kemaman and Kota Bharu. The day of year was used because it reflects seasonal meteorological effects which are not accounted for by the meteorological variables. There are two types of decision tree: (1) those for numerical data, which are known as 'regression trees' and (2) those for categorical data, which are known as 'classification trees'. Since the response data in this study are continuous numerical data, therefore the analysis is based on the BRTs regression trees type. Table 2 provides the details of the variables that were used to model 1-hourly PM_{10} concentrations.

Table 2 Variables used to model 1-hourly PM10 concentrations (µg/m3)

Variable	Response and predictors	Description and units	Variable type
Response	PM_{10}	Diameter (Dp) ranged below 10 µg/m^3	Continuous
Predictors	Time system		
1.	T hour	Time of the day (t_{hr})	Discrete (1–24)
2.	T Julian day	Julian day (t_{jd}))	Discrete (1–365)
3.	T weekday	Weekend or weekday ($t_{weekend/weekday}$)	Discrete (1–7)
	Meteorological factors		
4.	T	Ambient temperature (˚C)	Continuous
5.	rh	Relative humidity (% rh)	Continuous
6.	ws	Wind speed (m/s)	Continuous
7.	wd	Wind direction (degree)	Discrete (1–360)
	Gaseous		
8.	NO_2	Nitrogen dioxide (ppb)	Continuous
9.	NO	Nitrogen oxide (ppb)	Continuous
10.	CO	Carbon monoxide (ppb)	Continuous
11.	SO_2	Sulphur dioxide (ppb)	Continuous

D. Experimental steps

The **first step** of the experiment was to find a good set of settings for the algorithm parameters *lr, tc and nt, i*.e. the learning rate, interaction depth or tree complexity and number of trees, respectively. To conduct this, four models were developed by using four different datasets and in each of the simulations the parameters (*lr*, *tc* and *nt*) were set at different values. The learning rate is a shrinkage parameter that is applied in each iteration to shrink the contributions of a tree. The smaller the *lr* value the bigger the number of trees in the model [19]. The interaction depth is the maximum depth of variable interactions. For example, a *tc* of 1 means that there are no other predictor variables in a tree, while a *tc* of 2 means there are two predictors in a tree that interact, and so on. These two parameters (*lr* and *tc)* have a significant impact on the number of trees expansion and need to be determined to develop an optimum iteration or the best algorithm to suit a dataset. As a rule of thumb, the lowest *lr* value will give the highest *tc* value and produce the best and most accurate model [16], [18] and the number of trees can be estimated by using the k-fold cross-validation method.

The aim of the **second step** of the experiment was to evaluate the predictive performance of the different models by varying the interaction depth parameter by using different k-fold cross-validations, where the value of k was set as 10 during model fitting. Basically, a spatial version of the k-fold cross-validation method used by [15] was used and implemented by using an R package and the *gbm* library [25]. Briefly, the rationale for adopting this approach is that the cross-validation method has the advantage of using all the data for both training and validation by repeating the process k-times on different combinations of subsamples and calculating the mean performance of the v-models. The k-fold cross-validation method partitions the data into k-subsets, where k-models are built based on the 'k minus 1' subsets and the model performance is tested based on the last remaining subset [25]; this process is repeated several times with different number of trees for model parameter until the optimum number of trees with the minimal error is determined.

All the models were fitted in the R package R 3.3.2 using the '*gbm*' or 'generalized boosted machine' package version 2.1 [27], [25]. Each model was trained within 1 hour by using a modern computer with an Intel Core i7 (3632QM) CPU. Hence all 25 BRT models were completely trained within 1 to 2 days. The full dataset that was used for each site was first divided into two parts, where 70% was used as the training set to develop the model and 30% was used as the independent testing set. The latter was mainly used to determine the optimum interaction depth as this value is best identified by using data not involved in the model development process [18]. Also, a random component was injected in order to improve the prediction performance. This was achieved using a bag fraction of 0.5 or 50% of the training set observations randomly selected to fit each consequent tree [14].

Finally, a multiple *gbm* models were fitted using different combinations of learning rates of 0.001 and *tc* of 5 are the best fit for PNC via the 10-fold cross-validation method was determined. In line with [21], [22], a Gaussian error distribution was assumed for all models.

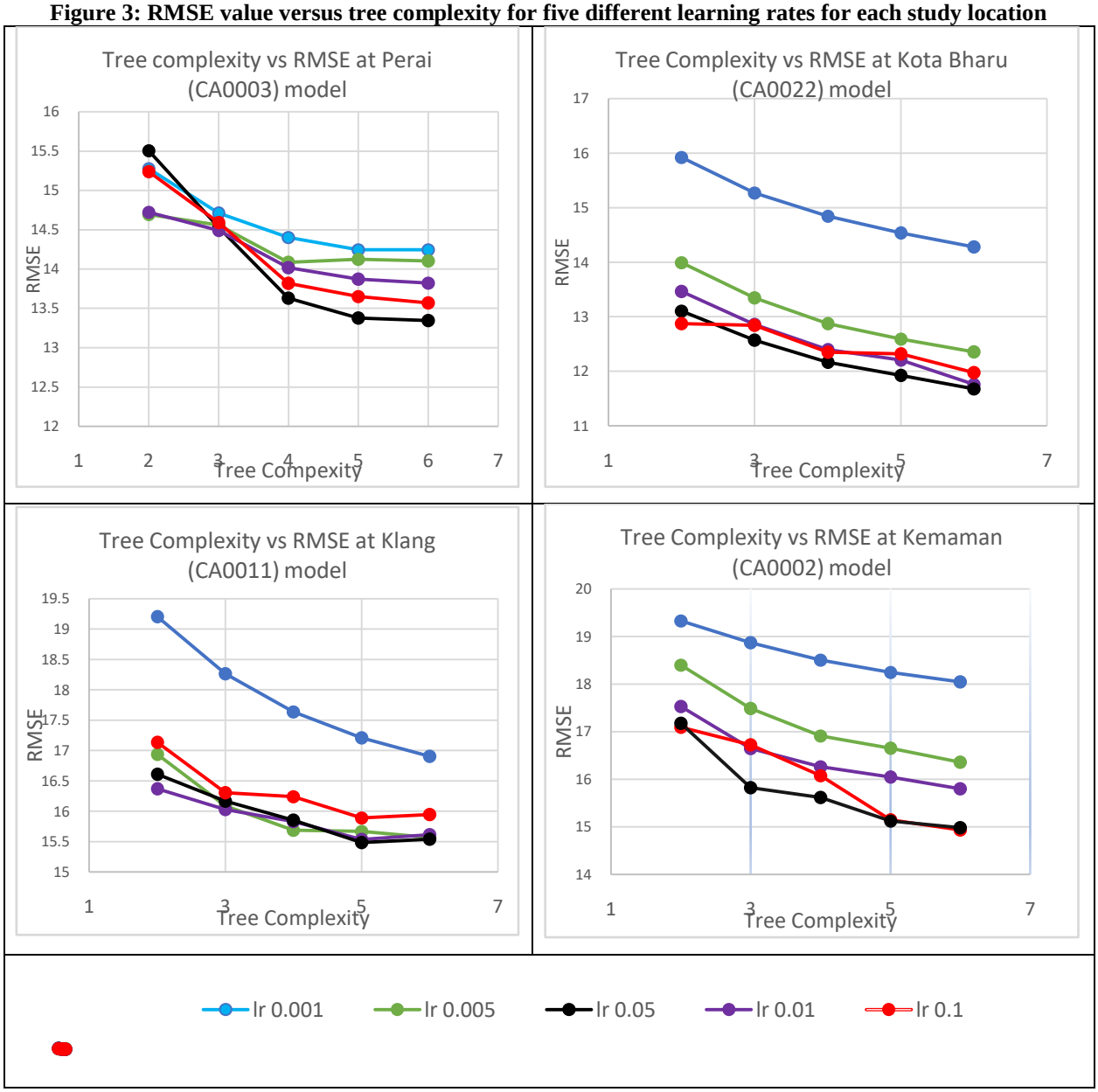
The final models were tested to the holdout test fraction with the optimum number of trees and were then evaluated by using error bias analyses to compare the observed PM_{10} concentration values with the predicted values according to a range of evaluation statistics namely the factor of two (FAC2), Root Mean Square Error (RMSE), correlation coefficient (R), coefficient of efficiency (COE) and index of agreement (IOA) [25],[29].

III. RESULTS

An experiment was conducted to identify the best tree complexity that could be produced by different learning rates and to test the relationships between the number of trees and the sample size used in the iterations. In the first step of the experiment, sample of the cross-validation *(cv)*, performance were plotted using the *gbm* command to illustrate four different iterations or *nt*, one at each station, as shown in Figure 3 (a, b, c, d). The RMSE values were plotted between the *tree complexity* and the error or predictive deviation for the model fitted to the dataset as an example. Five lines were plotted for each dataset for the tree complexity versus the number of iterations to determine the minimum error and best fit for each station. Figure 3 highlights the RMSE value of the relationship between tree complexity and the five different learning rates which ranged from 0.001 to 0.1 and which were applied to all four datasets.

It is clear that from the figure that the *lr* value of 0.05 consistently gave the lowest RMSE for the four different models. In contrast, the smallest *lr* value of 0.001 gave the highest RMSE value. It was also found that a *lr* of 0.001 required more iterations to produce a better model. Thus, the smallest *lr* as a setting to model these datasets has two major drawbacks: a high RMSE and high time consumption. In this experiment, the maximum of number of trees was set at 10,000 in order to try to lower the computation cost and reduce time consumption. While a higher value for the tree complexity would improve the RMSE value for BRT models, a higher *tc* value leads to more time consumption and is therefore not practical given that much more time consumption is needed for a relatively minor improvement [18]. The result of the experiment indicated that a *tc* of 5 gave the minimum RMSE for almost all the models. Thus, a *tc* of 5 was determined as the best value to fit these data. Similar procedures were performed on all the iterations to determine the optimal *nt* for the stations.

Figure 3: RMSE value versus tree complexity for five different learning rates for each study location



It is often challenging to develop a good and representative environmental model because it usually involves the need to manage big data. In the real world, data are captured and become available second by second on an hourly, daily or monthly basis over numerous years, so it is not uncommon to have to deal with thousands of datasets. Therefore, the influence of the number of datasets used in setting the parameters for the BRT and their influence on the BRT fitted model were also examined in this study.

For this purpose, each dataset was divided into five groups consisting of 200, 500, 1000, 3000 and 8000 samples and then these groups of samples were fitted into the BRT models by using the best algorithm parameter settings from the best iteration, namely a *tc* of 5 and a *lr* of 0.05. This resulted in, for example, for the Perai station, 355 trees for 200 samples, 548 trees for 500 samples, 850 for 1000 samples, 1951 for 3000 and 3338 for 8000 samples, which were plotted in order to identify the relationship between the two variables. Figure 4(a, b, c, d) illustrates the relationship between these two variables for all stations.

It is clear from the figure that there is a highly positive correlation between the sample size and the number of trees. There is a linear relationship between the number of trees and number of samples with an R^2 value of more than 0.9, which indicates that the technique is reliable to be used in prediction. It also shows the consistency of the R^2 values for all datasets, which indicates that the BRT model is good and consistent and can be used to analyse air pollution and meteorological data.The predicted number of trees gives an idea to limit the maximum number of trees used in the initial setup of the *gbm* algorithm. At the early part of the experiment number of trees of 10000 trees were set with the *lr* of 0.05 and a *tc* ranging from 1 to 5 by 10-fold cross-validation were performed. Results shows that the number of trees was equal to the maximum number of trees, which indicates that more trees are required to achieve the optimum. Therefore, smaller *lr* value implies that a greater number of trees is needed to achieve convergence because the *lr* value is inversely proportional to the number of trees [26]. The selection of a few different sample sizes of a dataset to get the number of trees based on the linear relationship between the two variables, as shown in Figure 3 (a, b, c, d), was more efficient and took much less time compared to setting the limit of the number of trees value to be the same as the number of samples in the dataset.

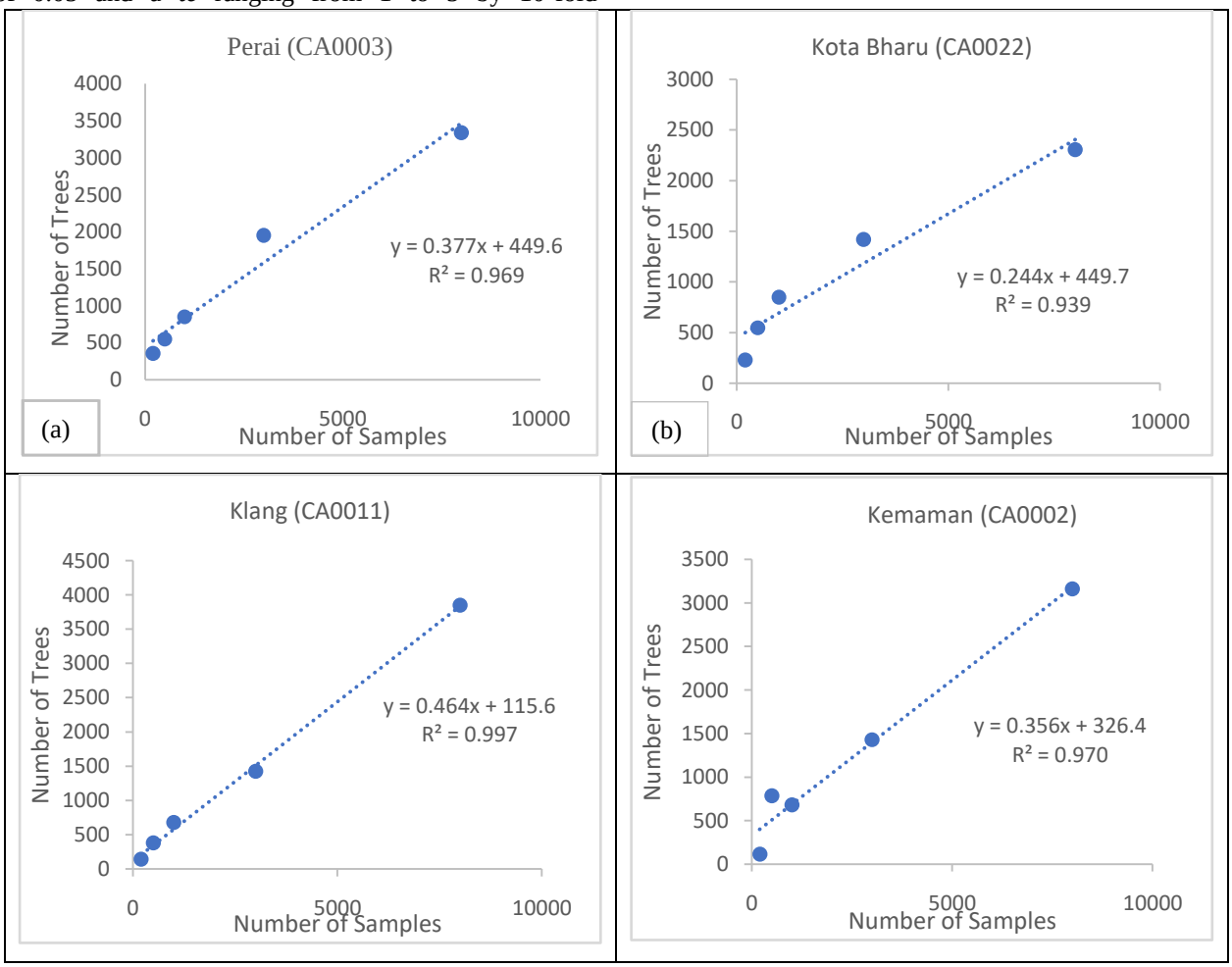


Figure 3: Relationship between dataset sample size and number of trees for each study location

The second stage of the experiment involved examining the influence of the number of samples on the number of estimated trees in all stations. This was done by applying the BRT settings (lr = 0.05 and tc = 5) to all the datasets in order to determine the number of samples by using the 10-fold cross-validation technique. This technique was used to estimate the best number of trees for each dataset. The number of trees obtained from the *best-iter* for Perai station (from 72,133 data) was 28,490 trees. When the estimated number of trees was calculated from the linear equations derived from Figure 4a, the number of trees for Perai was estimated as 27,680 trees. Similar steps were applied to Kota Bharu, Klang and Kemaman stations for data samples of 29,219, 57,175 and 42,105, respectively, the estimated number of trees for Kota Bharu, Klang and Kemaman was 28,465, 58,229 and 54,584 by 10-fold cross-validation, respectively. Perai and Kota Bharu had nearly the same number of trees because of missing data for some of the independent variables in the Kota Bharu dataset such relative humidity which were not available from the year 2000 to early 2003. Thus it is clear that missing values in the predictors affects the number of trees in the model, but *gbm* has the capability to handle missing data, as previously reported by [17]. Table 4 summarizes the model parameters and performance obtained from the 10-fold cross-validation for the four stations.

Table 3: Model parameters and performance obtained from the 10-fold cross-validation procedure.

Model		Number of samples (1)	Number of trees from CV (2)	Linear equation (4)	R^2 (5)	Number of trees from (4)
Perai (CA0003)	Bharu	72133	28490	Y = 0.3775x + 449.66	0.9691	27680
Kota (CA0022)		117095	28465	Y = 0.2444x + 449.78	0.939	29219
Klang (CA0011)		122735	58299	Y = 0.4649x + 115.65	0.0075	57175
Kemaman (CA0002)		117715	54584	Y = 0.3568 + 326.41	0.9709	42105

The number of trees that was estimated by using the 10-fold cross-validation method were compared with the number of trees estimated by using linear regression to determine whether the linear equations in Table 3 were reliable for estimating the number of trees at the initial setup of the *gbm*. The comparisons for all four stations are shown in Figure 4, which clearly indicates the two methods of estimation are strongly correlated at 0.87. Although high correlated, the number of trees estimated by using the linear equation is not the final number of trees that can be used for the prediction of PM_{10} concentrations. More trees must be added to the number of trees estimated by using the linear equation to give some head room for error.

A. Performance Evaluation of BRT Model

The following steps were undertaken prior to analysing the performance of the predictive model, which in this study was a predictive model for PM_{10}. First, the PM_{10} prediction data were pooled from the 10-fold cross-validation and saved in .csv files.

Next, the data were exported into excel spreadsheets and combined with the monitored data. Then the data were analysed using the R package version 3.3.2 with the openair package [30] and the results were reported statistically and graphically. Figure 5 (a, b, c, d) depicts a comparison between the hourly observed and modelled PM_{10} concentrations from 2006 to 2013 for Perai station and from 2000 to 2013 for the other stations. It can be seen from the figure that the fluctuation in the hourly PM_{10} concentration was predicted well by BRT except for some extreme values of PM_{10}.

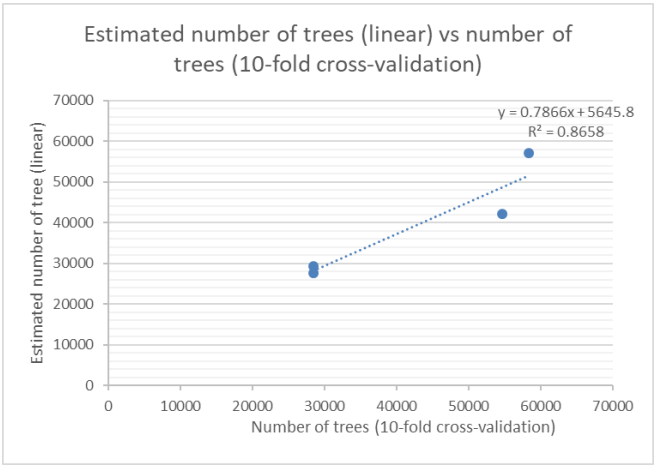


Figure 4. The number of tree by using the 10-fold cross-validation method verses estimated by using linear regression

It is essential to tune the *gbm* parameters to optimize the precision and accuracy of the prediction. The performance of all four models was assessed based on the following: factor of two (FAC2), mean bias (MB), mean gross error (MGE), root mean square error (RMSE), coefficient of correlation (R), coefficient of efficiency (COE) and index of agreement (IOA). The results of all the optimized BRT models are summarized in Table 5. From the table it can be seen that the FAC2 values of the BRT model for each station was acceptable as it is within suggested value range of 0.5 to 2. As for the MB results, the Perai and Klang prediction means were 2.11 and 2.65, respectively, which shows that there was overestimation of the PM_{10} concentration, while the Kemaman and Kota Bharu prediction means were 0.51 and 1.48, respectively, which indicates that there was an underestimation of the PM_{10} concentration. From the table, the MGE for Perai, Kemaman, Klang and Kota Bharu was 7.67, 9.69, 16.62 and 10.13, respectively, while the RMSE value was 11.37, 14.61, 24.18 and 13.83, respectively. The correlation of coefficient (R) and R^2 between the observations and the fitted model obtained from this analysis show how well the BRT model fits. It was found that the R^2 values between the fitted model data and the dataset were more than 0.5, which indicates that the model is acceptable and good. The Pearson's correlation coefficients of the observed and predicted PM_{10} concentrations were 0.808, 0.784, 0.849, and 0.796 for Perai, Kemaman, Klang and Kota Bharu station, respectively, which means that about 80% of the variance explained by the model is statistically valid. This means that more than 50% of the variation in the response variable (i.e., the PNC) is explained by the variation in the explanatory variables. Finally, *the p-values* of this model were less than 0.001 ($p < 0.001$), which shows that the model is statistically significant for all four stations.

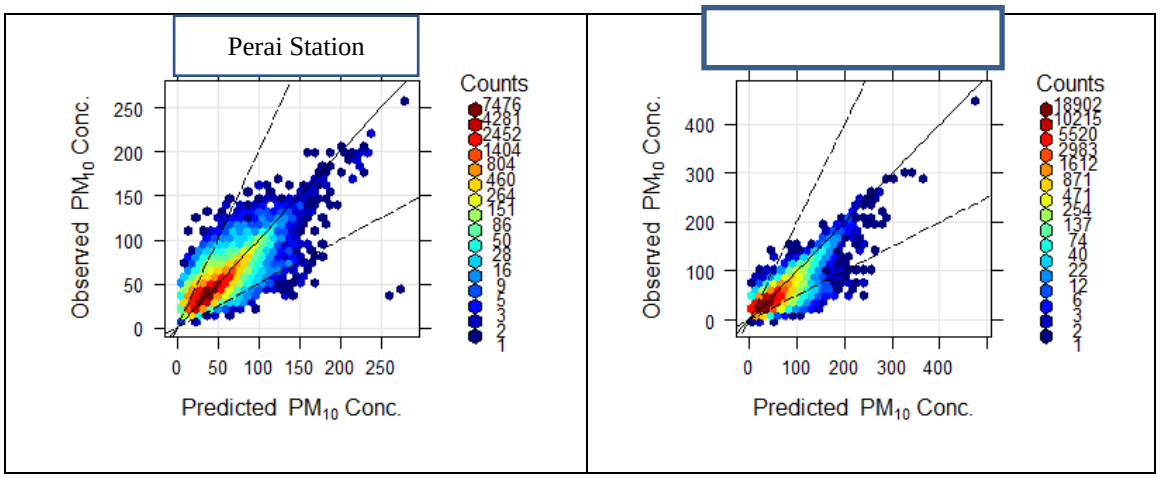
Table 4: Statistical measures of the BRT model to estimate the PM_{10} concentrations at all stations.

BRT Model	Perai	Kemaman	Klang	Kota Bharu
Number of samples	72133	117095	120154	117715
Factor of two	0.96	0.937	0.938	0.923
Mean bias	2.109	-0.508	2.651	-1.479
Mean gross error	7.677	9.689	16.62	10.125
Root mean square error	11.366	14.609	24.184	13.828
Coefficient of correlation	0.808	0.784	0.849	0.796
R-squares (fitted model)	0.652	0.614	0.720	0.633
Coefficient of efficiency	0.44	0.423	0.432	0.391
Index of agreement	0.72	0.711	0.716	0.696

Figure 5 shows the observed PM_{10} and predicted PM_{10} concentrations for Perai, Kota Bharu, Klang and Kemaman in scatter plots using hexagonal binning. The counts of the data in the bins are depicted on a colour scale. The solid line in the scatters plot shows the 1:1 relation between the observed and predicted PM_{10} concentration, and represents the perfect model. The dashed lines at the graphs show the 1:0.5 and 1:2 relations indicate FAC2 (a factor of 2) value. From the Figure 6 it can be seen that just few of the data were not within the FAC2 lines, while most of the data was close to the perfect model lines.

Figure 6 shows the conditional quantiles for the hourly predicted readings of PM_{10} concentration and hourly observed readings of PM_{10} concentration for each station. Conditional quantiles are a very useful way of comparing modelled against observed continuous measurements [31]. In the figure, the diagonal blue line shows the results for the perfect model, the red line shows the median of the prediction value while the other two dashed lines show the 25/75 percentile and 10/90 percentile, respectively. It was calculated by splitting the data into evenly spaced bins. The Perai, Kota Bharu, Klang and Kemaman median value of prediction lies near to the blue line until PM_{10} concentration values of 130, 270, 460 and 250 ug/m^3 are reached, respectively. This indicates there was a huge difference between the predicted PM_{10} concentration and the observed PM_{10} concentration when the median and lies far from the blue line. A perfect model should lie on the blue line and have a very narrow spread or shaded area. Note that there is still some spread even in a perfect model because a specific quantile interval will contain a range of values.

Finally, the histogram in Figure 6 shows the count of the predicted and observed in counts. The grey shaded bars of the histogram show the counts for the modelled PM_{10} concentrations, while the blue outlined bars show the observed PM_{10}. A comparison of the count between the observed and modelled PM_{10} shows that there is both underestimation and overestimation for each bin. The BRT models tended to overestimate at low concentrations; for example, the Perai BRT model showed a concentration below 30 ug/m^3 which was an underestimation. On the other hand, at higher concentrations, the Perai and Klang BRT models tended to overestimate, whereas the Kota Bharu and Kemaman BRT models tended to underestimate.



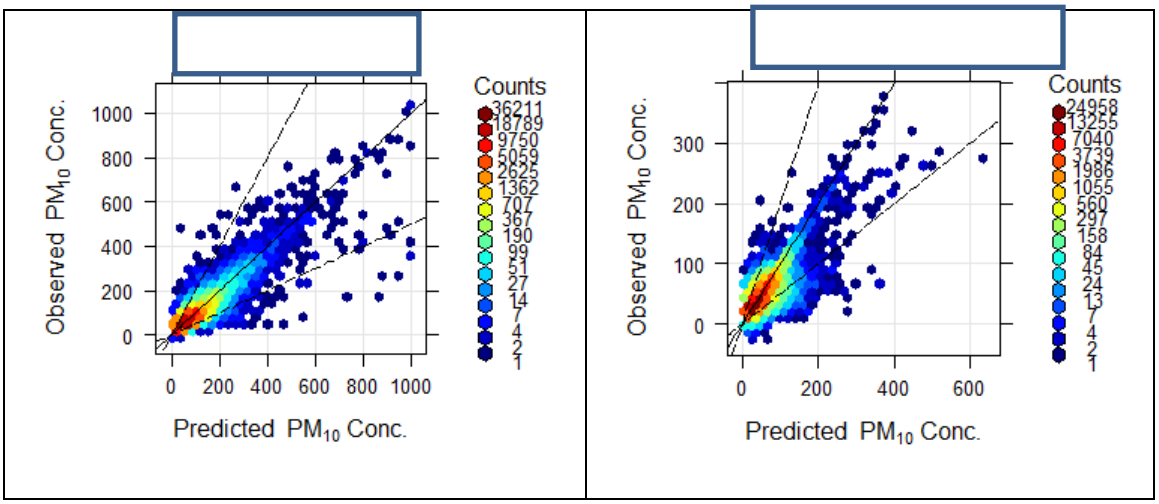


Figure 5: Scatter plots based on hexagonal binning of hourly observed PM_{10} and modelled PM_{10} concentrations for all four stations

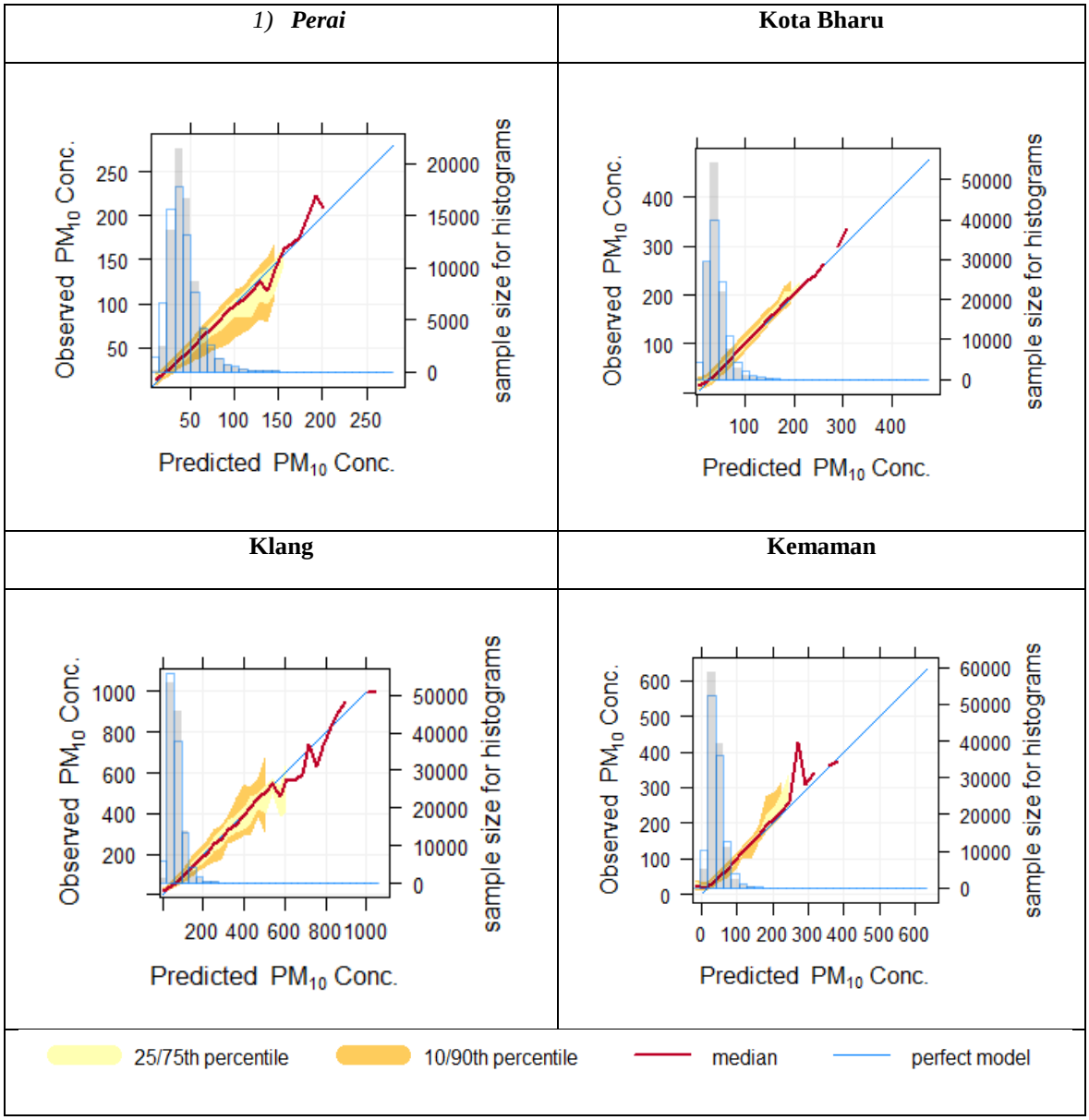


Figure 6: Conditional quantiles of hourly observed and modelled PM10 concentrations for Perai, Kota Bharu, Klang and Kemaman stations

IV. CONCLUSION

The datasets selected for this study were shown to be appropriate for BRT analysis. In this study, it was found that a *tc* of 5 was the best fit for the data. The rule of thumb for the *lr* and *tc* value was that the smaller *lr* and the bigger *tc* will always give the best optimum number. However, using the smallest possible *lr* and the biggest possible *tc* was deemed to be impractical because doing so would result in high time and computational costs when training a big dataset. Thus, the *lr* and *tc* values that were needed to optimize the results would compromise time consumption when training the models.

Therefore, this study tested five different *lr* values and five different *tc* values with a default number of trees that was limited to 10000 trees. This process required 25 models to be trained with 10000 trees and the 10-fold cross-validation estimated that the best number of trees was more than 10000 trees. Each model was trained within 1 hour by using a modern computer with an Intel Core i7 (3632QM) CPU. Hence all 25 BRT models were completely trained within 1 to 2 days. When using a higher number of trees, it takes a lot of time to completely train 25 models but the time needed also depends on the size of the data and the maximum number of trees at initial setup. For example, when the maximum number of trees was set at 60000 it took up to 1 day to train a model on one set of parameters, thus a lot of time was consumed to completely train all 25 models with this number of trees.

Finally, the models were set with best *lr* and *tc* combination to train several dataset samples sizes to determine the linear relation between the number of trees to set in the *gbm* algorithm and the sample size. The models were then trained by using the tuned *gbm* parameters and were ready for predictions. The number of trees increases in a linear manner as the number of samples grows, as demonstrated by the replication of the experiment four times by using four different datasets of air quality data from different locations. Therefore, the BRT model satisfied the statistical performance indicators.

ACKNOWLEDGMENT

The authors would like to convey their great appreciation to the Air Division, Department of Environment Malaysia and Universiti Malaysia Terengganu for supporting this research.Authors should consider the following points.

REFERENCES

1. WHO: WHO Air Quality Guidelines Global Updates: Particulate matter, ozone, nitrogen dioxide and sulfur dioxide, Copen- hagen, Denmark, 2006.
2. US EPA. (2015). Health and Environmental Effects of Particulate Matter (PM). Retrieved from https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm
3. R. Afroz, M.N. Hassan, N.A. Ibrahim. (2003) Review of air pollution and health impacts in Malaysia. Environ Res; 92, pp. 71-7.
4. Alam Sekitar Malaysia Sdn Bhd (ASMA) (2007). Standard Operating Procedure for Continuous Air Quality Monitoring. Shah Alam, Selangor Malaysia.
5. A. Azhari, M.T. Latif, & A.F. Mohamed. (2018). Road traffic as an air pollutant contributor within an industrial park environment. Atmospheric Pollution Research, 9(4), 680–687. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1309104217304361
6. Department of Environment Malaysia. Malaysian environmental quality report. Ministry of Natural Resources and Environment. 2013.
7. A.Z. Ul-Saufie, A.S. Yahaya, N.A. Ramli, N. Rosaida, & H.A. Hamid, (2013). Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmospheric Environment, 77, 621–630. https://doi.org/10.1016/J.ATMOSENV.2013.05.017
8. S.A. Abdul-Wahab, C.S. Bakheit, & S.M. Al-Alawi (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environmental Modelling & Software, 20(10), 1263–1271. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364815204002129
9. V. Gvozdić, E. Kovač-Andrić, & J. Brana. (2011). Influence of Meteorological Factors NO2, SO2, CO and PM10 on the Concentration of O3 in the Urban Atmosphere of Eastern Croatia. Environmental Modeling & Assessment, 16(5), 491–501. https://doi.org/10.1007/s10666-011-9256-4
10. E. Kovač-Andrić, J. Brana, & V. Gvozdić, (2009). Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. Ecological Informatics, 4(2), 117–122. https://doi.org/10.1016/J.ECOINF.2009.01.002
11. S. Abdullah, M. Ismail, A. Najah Ahmed, & S.Y. Fong (2017). Evaluation for Long Term PM_{10} Concentration Forecasting using Multi Linear Regression (MLR) and Principal Component Regression (PCR) Models. Environmental Asia, 9(2):101-110 · July 2016
12. P. Viotti, G. Liuti, & P. Di Genova. (2002). Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. Ecological Modelling, 148(1), 27–46. https://doi.org/10.1016/S0304-3800(01)00434-3
13. J.H. Friedman (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics 2001; 29(5): 1189-1232
14. J.H. Friedman. (2002). Stochastic gradient booleasting. Computational Statistics & Data Analysis, 38(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2
15. Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., Taylor, P., (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology-Progress Series 321, 267–281
16. G. Ridgeway. (2007). Generalized Boosted Models: A Guide to the gbm Package. R package vignette, URL. http://CRAN.R-project.org/package=gbm
17. G. De'ath. (2007). Boosted trees for ecological modeling and prediction. Ecology, 88(1), 243–251. https://doi.org/10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2
18. Elith, J., Leathwick, J.R., Hastie, T., (2008). A working guide to boosted regression trees. Journal of Animal Ecology 77 (4), 802–813.
19. N.Z.Yahaya, N.A. Ghazali, S. Ahmad, M.A. Asri, Z.F. Ibrahim, & N.A.Ramli. (2017). Analysis of daytime and nighttime ground level ozone concentrations using boosted regression tree technique. EnvironmentAsia, 10(1). https://doi.org/10.14456/ea.2017.14
20. N.Z. Yahaya, S.M. Phang, A.A. Samah, I.N. Azman and Z.F. Ibrahim (2018). Analysis of Fine and Course Particle Number Count Concentrations Using Boosted Regression Tree Technique in Coastal Environment. Journal of Environment Asia, Vol 11 (3) Sept 2018
21. D.C. Carslaw, & P.J. Taylor. (2009). Analysis of air pollution data at a mixed source location using boosted regression trees. Atmospheric Environment, 43(22–23), 3563–3570. Retrieved from https://www.sciencedirect.com/science/article/pii/S1352231009003069?via%3Dihub
22. N.Z. Yahaya (2013) Temporal and Spatyial Variations of Ultra-fine Particles in the Urban Environment. PhD Thesis, Institute for Transport Studies, University of Leeds, United Kingdom.
23. A. Sayegh, J.E. Tate, & K. Ropkins. (2016). Understanding how roadside concentrations of NOx are influenced by the background levels, traffic density, and meteorological conditions using Boosted Regression Trees. Atmospheric Environment, 127, 163–175. https://doi.org/10.1016/J.ATMOSENV.2015.12.024
24. D. Derwent, A. Fraser, J. Abbott, M. Jenkin, P. Willis. (2010). Evaluating the performance of air quality models. Report to the Department for Environment, Food and Rural Affairs, the Scottish Executive, Welsh Assembly Government and the Department of the Environment Northern Ireland.

25. A. Suleiman, M.R. Tight, & A.D.Quinn. (2016). Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter. Environmental Modeling & Assessment, 21(6), 731–750. https://doi.org/10.1007/s10666-016-9507-5
26. G. Ridgeway. (2012). gbm: Generalized Boosted Regression Models. R package version 1.6e3.2. http://CRAN.R-project.org/package=gbm
27. R Core Team, R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL. http://www.R-project.org/. 2014.
28. Derwent D, Fraser A, Abbott J, Jenkin M, Willis P, Murrellst (2010). Evaluting the performance of air quality models. Report to the Department for Environment, Food and Rural Affairs, the Scottish Executive, Welsh Assembly Government and the Department of the Environment Northern Ireland.
29. C.J. Willmott, S.M. Robeson, K.A. Matsuura. (2012). Refined index of model performance. International Journal of Climatology; 32(13): 2088-94
30. Carslaw, D. C., & Ropkins, K. (2012). openair — An R package for air quality data analysis. Environmental Modelling & Software, 27–28, 52–61. https://doi.org/10.1016/J.ENVSOFT.2011.09.008
31. WILKS, D. S. (2005). Statistical Methods in the Atmospheric Sciences, Volume 91, Second Edition (International Geophysics). 2nd ed. Academic Press (cit. on p. 243).

AUTHORS PROFILE

Dr Noor Zaitun Yahaya (AHEA UK).
She obtained her PhD from Institute for Transport Studies, The University of Leeds, United Kingdom and Master Degree in Civil Engineering from University Sains Malaysia. Currently she work as a university Malaysia Terengganu, Malaysia as a senior lecturer and she is the leader for An Air Quality and Environment Research Group for the University Malaysia Terengganu. Her research focusses on the nano-particles analysis and the used of artificial intelligent approach to analyse big data. She is currently member of the Board of Technologies of Malaysia, President of the Clean Air Society Forum of Malaysia, (a professional bodies in Malaysia) and board member of The *International Union of Air Pollution Prevention and* Environmental Protection Associations (*IUAPPA*) or The World Clean Air Congress. Her product called *MyAtmos* won Gold Medal in International The International Invention, Innovation & Technology Exhibition (ITEX) in year 2016. She is an active researcher with more than 50 publications in the international journals and proceedings for the last 5 years and more than 20 lecture series all around locally and internationally.

Zul Fadhli bin Ibrahim. He is currently enrolled as a Postgraduate and Researcher at School of Ocean Engineering. He obtained his Bachelor Degree in Environmental Technology from University Malaysia Terengganu in the year 2014. Her research focuses on analyzing particulate matter data from various station in Peninsular Malaysia and applies an artificial intelligent approach which name The Scholastic Boosted Regression Trees Technique as part of his study. Along the way he presented his research in renowned conference such as in World Clean Air Congress and the Better Air Quality conference in Busan , South Korea (2016), World Expo in Astana, Kazakstan in 2017, International Conference on Air Quality and Sustainable Environment 2017 in Malaysia. He is an active member of the Clean Air Society Forum of Malaysia since 2015.

Assoc. Prof. Dr Jamaiah Yahaya, Dr. Jamaiah Yahaya is the Associate Professor at Faculty of Information Science and Technology (FTSM), The National University of Malaysia (UKM) since July, 2011. Prior that she worked as a senior lecturer in School of Computing, Northern University of Malaysia (UUM) and a system analyst at University of Science Malaysia (USM). Her bachelor degree was Bachelor of Science in Computer Science and Mathematics from University of Wisconsin-La Crosse, USA (1986), Master of Science in Information System from University of Leeds, UK (1998), and PhD in Computer Science from The National University of Malaysia (UKM) (2007). Her PhD thesis was the development of software certification model and later, she continued her PhD research as a post-doctoral fellow in UKM (2008). She was appointed in a few management posts such as the head of PhD program and acting Deputy Dean of Academic in FTSM, UKM. Her research interests are software quality, software development and management, and software assessment and impact. She is an active researcher with more than 100 publications in the international journals and proceedings for the last 5 years.