Air Quality, Atmosphere & Health (2021) 14:1647–1663
https://doi.org/10.1007/s11869-021-01045-3

Coupling of quantile regression into boosted regression trees (BRT) technique in forecasting emission model of PM_{10} concentration

Wan Nur Shaziayani[1] · Ahmad Zia Ul-Saufie[2] · Hasfazilah Ahmat[2] · Dhiya Al-Jumeily[3]

Received: 17 February 2021 / Accepted: 10 May 2021 / Published online: 24 May 2021

Abstract

Air pollution is currently becoming a significant global environmental issue. The sources of air pollution in Malaysia are mobile or stationary. Motor vehicles are one of the mobile sources. Stationary sources originated from emissions caused by urban development, quarrying and power plants and petrochemical. The most noticeable contaminant in the Peninsular of Malaysia is the particulate matter (PM_{10}), the highest contributor of Air Pollution Index (API) compared to other pollution parameters. The aim of this study is to determine the best loss function between quantile regression (QR) and ordinary least squares (OLS) using boosted regression tree (BRT) for the prediction of PM_{10} concentration in Alor Setar, Klang and Kota Bharu, Malaysia. Model comparison statistics using coefficient of determination (R^2), prediction accuracy (PA), index of agreement (IA), normalized absolute error (NAE) and root mean square error (RMSE) show that QR is slightly better than OLS with the performance of R^2 (0.60–0.73), PA (0.78–0.85), IA (0.86–0.92), NAE (0.15–0.17) and RMSE (9.52–22.15) for next-day predictions in BRT model.

Keywords Particulate matter (PM_{10}) · Quantile regression · Ordinary least squares (OLS) · Boosted regression tree

Introduction

The Air Pollution Index (API) describes the current state of air quality in a given region. The Department of Environment (DOE), Ministry of Environment and Water is one of the government agencies responsible for monitoring air quality at 68 stations in Malaysia. The API was then introduced to measure the cleanliness and efficiency of the air (Leong et al. 2020). The Malaysia Ambient Air Quality Guidelines (MAAQG) is used to determine the level of air quality in Malaysia and is used to measure the concentration levels of particles less than 2.5 μm ($PM_{2.5}$), particles less than 10 μm (PM_{10}), carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and ozone (O_3). When the concentration level is above the level specified in the MAAQG for a long period of time, it will cause negative effects on health and the environment. The API in Malaysia is listed in Table 1 with its categorization as good, moderate, unhealthy and hazardous. Generally, PM_{10} is identified as a major pollutant that causes unhealthy conditions (DOE 2018). Therefore, PM_{10} is the main focus of this study.

According to Azmi et al. (2010), the main causes of air pollution in Malaysia are either mobile sources from cars, buses and planes or stationary sources from power plants, open burning and wildfires, industrial facilities and others. The occurrence of haze in Malaysia is as a result of biomass burning since 1982 interrupting everyday life in Malaysia (Latif et al. 2018). Several haze episodes have been reported since then. These extreme episodes occurred in 1997, 2005 and 2015. Severe haze episodes were recorded in 1997 due to forest fires and large-scale plantations, especially in southern Sumatra and central Kalimantan, both in a neighbouring

✉ Ahmad Zia Ul-Saufie
ahmadzia101@uitm.edu.my

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 13500 Bukit Mertajam, Pulau Pinang, Malaysia

[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[3] Faculty of Engineering and Technology, Liverpool John Moores University, Liverpool, UK

Table 1 Malaysia Air Pollution Index

API	Air quality status
0–50	Good
51–100	Moderate
101–200	Unhealthy
201–300	Very unhealthy
> 300	Hazardous

country, Indonesia. The city of Kuching, Sarawak located in East Malaysia was one of the areas affected by air pollution and haze in Sarawak East Malaysia in 1997. The Kuching API was recorded above 850 during the haze, the most alarming haze in Malaysia (Zakri et al. 2018). A further episode of extreme haze was reported in 2005 (Sahani et al. 2014) which was mainly on the Peninsula's west coast of Malaysia. At that time, the smoke haze heavily affected the Klang Valley and its surrounding area. It reached its height at the haze emergencies on 11 August 2005, as the Air Pollution Index (API) reading in Port Klang and Kuala Selangor was recorded to be above 500. The latest extreme and long haze episode in Malaysia was reported in September 2015 (Huijnen et al. 2016). PM_{10} concentration is the most significant major pollutant released by human activity (Sapini et al. 2015). Specifically the PM_{10} concentration, in most cities of Southeast Asia (Reddington et al. 2014) and in Malaysia (Juneng et al. 2011), is justified as the main atmospheric pollutant. PM_{10} contributed most to Malaysia's API until 2017. In mid-2017, $PM_{2.5}$ had a greater impact on APIs in Malaysia until 2018 (DOE 2018).

There has been a growing interest in using many statistical models in the prediction of air pollution in recent years. One of these is regression techniques which have been used for a long time as predictive tools in many fields especially in the prediction of air pollution. The benefits of regression models are for its ease of use and efficient execution. However, these models are not very good in the prediction of complex situations, as the linear relationship between the selected parameters is determined (Abdullah et al. 2016). The statistical method is limited in clarifying the factors influencing PM_{10}, due to statistical assumptions and the homogeneity of the data. Recent studies have attempted to develop powerful computing intelligence models using machine learning algorithms such as the neural network to predict the complex PM_{10} concentration system, which indicate that such models can easily predict the desired value (Abdullah et al. 2017). However, machine learning, more specifically the neural network, is usually used as a black box where there is no specific understanding of the physical characteristic of the technique (Viotti et al. 2002).

The boosted regression tree (BRT) model, another type of machine learning, which combines the advantages of regression trees with the boosted adaptive method, has recently been used in air pollution prediction studies. The boosting method was first developed by Friedman in (2001), and later added a stochastic aspect to the boosting algorithm through a random sample of the training data sets (Friedman 2002). In addition, it can also be used as a general method that is useful to improve the model accuracy of each learning algorithm. The BRT produces an ensemble model by boosting the loss function (such as root mean square error) of the user-defined number of additional trees by minimizing it. In contrast to the black box technique, the BRT method would evaluate the response of variables based on the individual model variable. It is therefore possible to determine, rank and describe the relationship between variables (Yahaya et al. 2019). The BRT is also capable of handling various types of inputs (i.e. categorical and continuous data) and accepts missing values (Motevalli et al. 2019) and able to deal with multiple forms of loss functions (Ridgeway 2012), such as Gaussian, Laplace, quantile regression (QR), Bernoulli and Poisson.

The loss function is one of the BRT model factor considerations. Ordinary least squares (OLS) loss function has been used by many studies, for the purpose of minimizing the squared error for continuous predictors, which resulted in a better correlation between the observed value and the estimation of the generalized boosting model (GBM) (Gu et al. 2019). However, datasets that have outliers such as air pollution data are not suitable to be used in OLS function. According to Kudryavtsev (2009), QR has become an important robust alternative tool, as it is more resistant to outliers and it is free function and does not have any properties.

The QR has the ability to be more useful and precise, since the non-central location of a distribution can be represented in all quantiles (Lingxin and Naiman 2007). The QR has the capability of including models for all quantiles, evaluating the entire function and calculating the central tendency (such as mean, median and mode) in the entire function of the variable of interest. The advantage of QR is for its robustness against non-OLS distribution which was found by Schlink et al. (2010). It can also be adapted to unbalanced observational frequencies. Due to this property, QR was considered and selected as a loss function strategy for this study.

The aim of this study is to derive air pollution modelling based on the loss function of QR using the BRT method. It is clear from the literature that no study has been conducted using such a method to predict PM_{10} concentrations. The finding from the proposed methodology is compared with the prediction obtained from the OLS loss function using the BRT method.

Methodology

The process of data preparation has been conducted in detail to reach for developing the model evaluation as illustrated in Fig. 1. The flow diagram is adapted to the author's research and is reconstructed.

Data preparation

Three urban sites were selected for this study. Table 2 shows the characterization of each station. All stations are located in the peninsular Malaysia. Alor Setar station (CA0040) is located in the northern region, Klang (CA0011) is located in the west coast region and Kota Bharu (CA022) is located in the east coast region as shown in Fig. 2. Data are operated by the

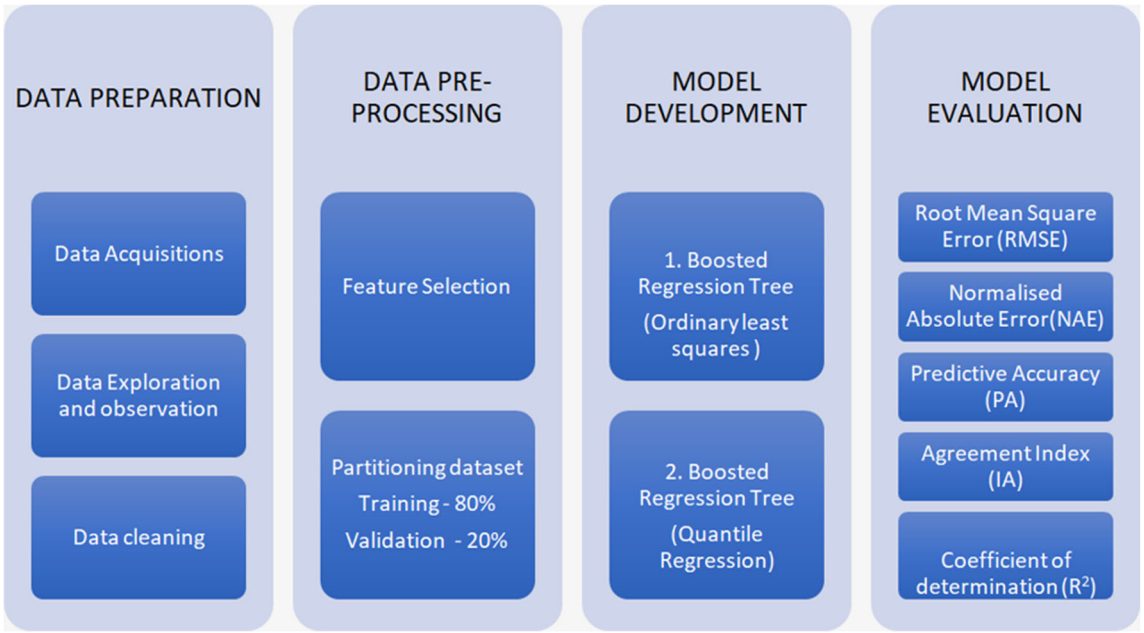

Fig. 1 Building the prediction model workflow

Department of Environment's continuous air quality monitoring (CAQM) stations in Malaysia. CAQM is an integrated ambient air quality monitoring device, is outfitted with a variety of ambient air analyses and sensors to identify particular pollutants. The analyses and sensors operate in a continuous mode, with data collected being captured on a microcomputer-based data acquisition system (DAS) that also controls the performance of the analyses and sensors. On an hourly basis, data is collected and transferred to a central computer for review and reporting. The United States Environmental Protection Agency (USEPA) has authorized the monitoring instruments and operational protocols of CAQM stations (Kamarul Zaman et al. 2017).

For data exploration, a descriptive analysis is carried out to determine the existence of extreme values or missing values. Missing data is a problem commonly faced by researchers in environmental studies. Data discontinuities are a major obstacle to the prediction models that require continuous information for the majority of the parts to be used. The absence of any data prevents the ability to accurately conclude or interpret the observation (Noor et al. 2014). The missing data must be processed, because complete data are required to perform statistical analysis. This study used linear interpolation for missing data imputation. According to Noor et al. (2015), this linear interpolation method estimates the missing data better than that of the other methods.

Data pre-processing

Maximum daily data used in this study were furnished by the Department of Environment (DOE), Ministry of Environment and Water of Malaysia for the period of 2002 to 2017. The data for this project are confidential, but may be obtained with Data Use Agreements with the Department of Environment (DOE), Ministry of Environment and Water of Malaysia. The data was 80% randomly selected for training and another 20% for the validation of the model (80% for model development and 20% to evaluate the performance of the model). The variables used in this study consist of gaseous nitrogen dioxide (NO_2; ppb), carbon monoxide (CO; ppb), sulphur dioxide (SO_2; ppb), ozone concentration (O_3; ppb), particulate matter concentration (PM_{10}; μgm^{-3}) and meteorological parameters such as wind speed (WS; km/h), relative humidity (RH; %) and temperature (T; °C), as the predictors used to predict

Table 2 Characteristics of monitoring station sites

Station ID	Location	Latitude	Longitude	Category
CA0040	Islamic Religious Secondary School, Mergong, Alor Setar, Kedah	N06° 08.218	E100° 20.880	Urban
CA0011	Raja Zarina Secondary School, Klang, Selangor	N03° 00.620	E101° 24.484	Urban
CA0022	Sekolah Menengah Kebangsaan Tanjong Chat, Kota Bharu, Kelantan	N06° 00.040	E102° 15.321	Urban

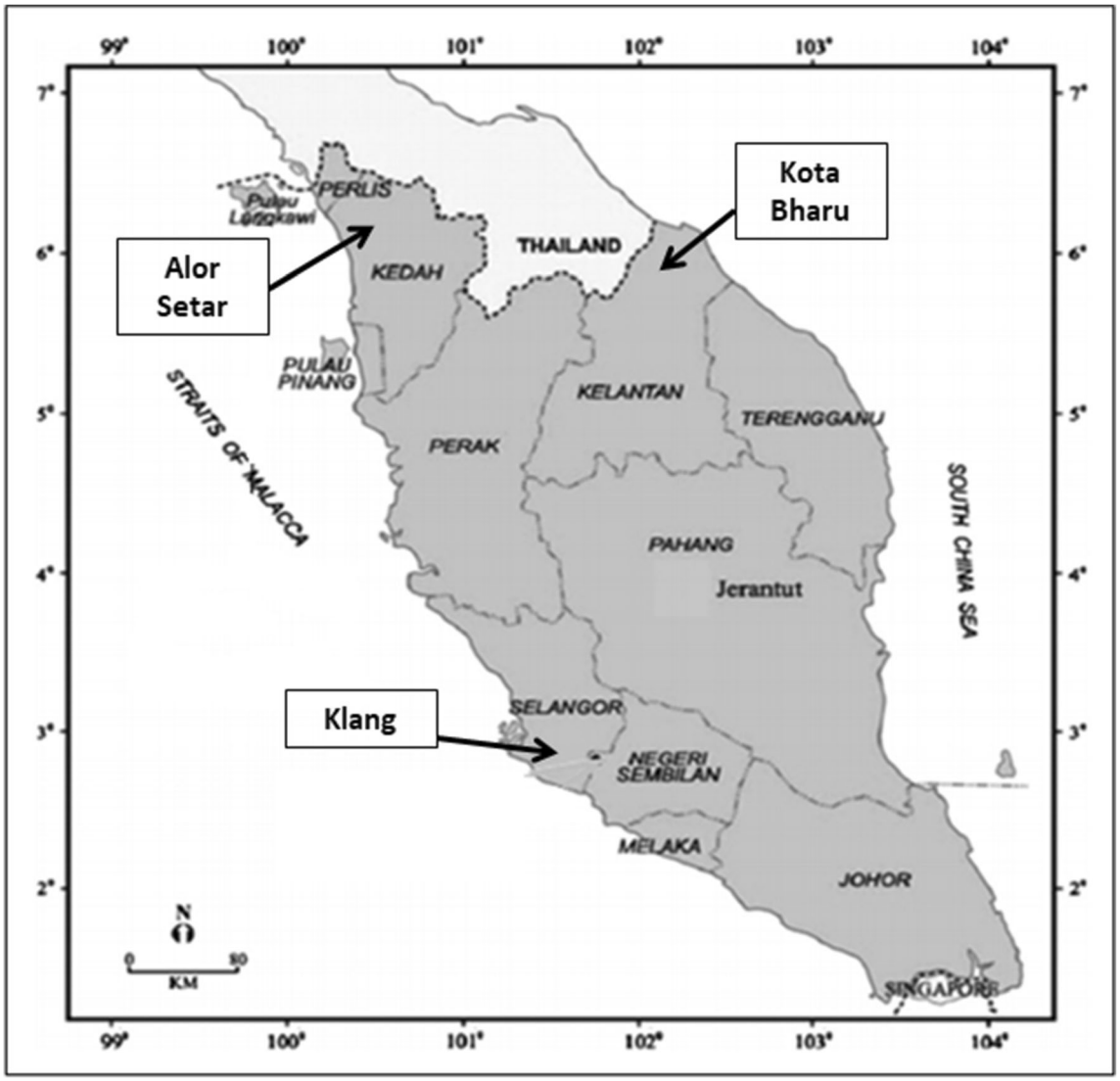

Fig. 2 Location of the monitoring sites (Ul-Saufie et al. 2012a, b)

PM_{10} concentrations 3 days ahead. All the selected parameters in this study have an influence on forecasts of PM_{10} concentrations for 3 days ahead, and had been used by previous researchers, as summarized in Table 3. The general models for this study are shown in Table 4. where

$PM_{10,D+1}$	Next-day prediction of PM_{10} concentration
$PM_{10,D+2}$	Next 2 days prediction of PM_{10} concentration
$PM_{10,D+3}$	Next 3 days prediction of PM_{10} concentration
$PM_{10,D}$	Particulate matter ($\mu g/m^3$)
CO_D	Carbon monoxides (ppb)
$NO_{2,D}$	Nitrogen dioxide (ppb)
$SO_{2,D}$	Sulphur dioxide (ppb)
$O_{3,D}$	Ozone (ppb)
RH_D	Relative humidity (%)
T_D	Temperature (°C)
WS_D	Wind speed (km/h)

Model development

BRT is a method used to increase the accuracy of a single model by fitting a number of models and combining them for prediction purposes. BRT uses regression trees from the classification and regression tree (CART) and constructs boosts to combine model sets (Grunwald et al. 2020). In the BRT, there are several tuning parameters that need to be controlled such as the number of trees (nt), the learning rate (lr) which is the shrinkage parameter used in each iteration to reduce the contribution of the tree, the complexity of the tree (tc) or the interaction depth which is the maximum tree depth of variable interactions. This study fitted BRT models with varying values for nt (10,000), lr (0.01) and tc (5). In version 3.4.2 of the *R* software, the BRT model was fitted with version 1.6-3.1 of the GBM. The GBM offers three methods for

Table 3 Variable selection by the previous studies in the prediction of PM_{10} concentration level

Authors	NO_2	SO_2	CO	O_3	PM_{10}	T	RH	WS	Others
Chelani et al. (2002)						√	√	√	TV_s, WD
McKendry (2002)	√		√		√				TV_s, MV_s, NO, $PM_{2.5}$
Lu et al. (2004)	√	√	√		√	√			NO, NO_x, WD, SR
Corani (2005)		√			√	√			P
Brunelli et al. (2007)						√		√	WD, P
Fernando et al. (2012)					√				MV_s
Perez (2012)					√				MV_s
Nejadkoorki and Baroutian (2012)			√		√				TV_s, MV_s, NO
Popescu et al. (2013)					√			√	WD
Liu et al. (2015)	√				√	√	√	√	MV_s
Navares and Aznarte (2020)	√	√	√	√	√				Pollen
This study	√	√	√	√	√	√	√	√	

Abbreviations of the parameters: *TV_s* temporal variables, *WD* wind direction, *MV_s* meteorological variables, *NO* nitrogen monoxide, *NO_x* nitrogen oxide, *SR* solar radiation, *P* atmospheric pressure

estimating the optimum number of trees, i.e. the cross validation (CV), the independent test set (test) and the out-of-bag estimation (OOB).

This research used 10-fold cross validation as suggested by Ridgeway (2010) to get the optimum number of trees, and then, ten separate testing sets were averaged. Rather than worrying about the block being suitable for testing, CV employs them all, one at a time, and summarizes the results at the end. The independent test set (test) approach uses a single holdout base dataset to determine the optimum number of tree (Ridgeway 2007). This research used a 50% held out test set to find the optimum number of trees as suggested by Ridgeway (2017). Out-of-bag estimation (OOB) is used to evaluate the classifier. According to Martinez-Munoz and Suarez (2010), individual classifiers are trained in standard bagging on independent bootstrap samples extracted with replacement from the set of original data. In general, the size of these samples is chosen to align with the number of the original training dataset. This prescription is arbitrary and does not have to be optimal in terms of the ensemble's generalization accuracy. The accuracy of the voting classifier is equal to the average of classifier. Bag.fraction 0.5 was used in this research, as suggested by Ridgeway (2020), to improve predictive performance while using the OOB method.

BRT constructs a model as a weighted sum of functions similar to other boosting algorithms. The BRT algorithm steps are summarized accordingly:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{1}$$

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2}$$

Start the model with a constant value $F_0(x)$.The BRT algorithm steps consist of a suitable decision tree and a loss function to determine how well a study is predicted. At each stage, the decision tree $h_m(x)$ is chosen to minimize the loss given the current model F_{m-1} and its fit $F_{m-1}(x_i)$. The residuals $r_{i,m}$ are computed:

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{3}$$

$r_{i,m}$ is the negative gradient of the *i*th sample in the *m*th as the number of trees. $h_m(x)$ is set to use the $r_{i,m}$ as the target variable. Fit a regression tree to the residual $r_{i,m}$ values and create the leaf node area $R_{j,m}$ for $j = 1, 2, \ldots, J$. The weights are obtained by solving the problem of minimization:

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma) \tag{4}$$

Table 4 General model of BRT

3 days ahead prediction	Model
Next-day prediction (D+1)	$PM_{10,D+1}$~gbm ($PM_{10,D}$, CO_D, $NO_{2,D}$, $SO_{2,D}$, RH_D, T_D, WS_D, $O_{3,D}$)
Next 2-day prediction (D+2)	$PM_{10,D+2}$~gbm ($PM_{10,D}$, CO_D, $NO_{2,D}$, $SO_{2,D}$, RH_D, T_D, WS_D, $O_{3,D}$)
Next 3-day prediction (D+3)	$PM_{10,D+3}$~gbm ($PM_{10,D}$, CO_D, $NO_{2,D}$, $SO_{2,D}$, RH_D, T_D, WS_D, $O_{3,D}$)

The square error is the loss function for the deterministic prediction:

$$L(y_i, F(x_i)) = \frac{1}{2}(y_i - F(x_i))^2 \quad (5)$$

For quantile regression, the expression below is used when the α (quantile) value is in range 0 to 1.

$$L(y_i, F(x_i), \alpha) = \begin{cases} \alpha(y_i - F(x_i)), & \text{if } y_i \geq F(x_i) \\ (\alpha - 1)(y_i - F(x_i)), & \text{otherwise} \end{cases} \quad (6)$$

$R_{j,\,m}$ is a leaf node, the jth being the number of leaf in the tree and υ is a learning rate. Update the current model:

$$F_m(x) = F_{m-1}(x) + \upsilon \sum_{j=1}^{J} \gamma_{jm} I\left(x \in R_{jm}\right) \quad (7)$$

It is a method of looping that fits the regression tree. Then, once the first tree is added to the model, tree error prediction will be taken into account to balance and boost the accuracy of the next tree.

$$F_m(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} \upsilon \gamma_{jm} I\left(x \in R_{jm}\right) \quad (8)$$

Model evaluation

Performance indicators in this research work are used to determine the accuracy and errors of BRT with different loss function (OLS and QR). The indicators used to identify the best method for the prediction of PM_{10} concentration were the root mean square error (RMSE), normalized absolute error (NAE), predictive accuracy (PA), agreement index (IA) and coefficient of determination (R^2). The RMSE and NAE were used to find a model error where a value closer to 0 demonstrated a better model. Meanwhile, the other three performance indicators, i.e. IA, PA and R^2, were used to verify the accuracy of the model outcome, where a higher accuracy is given by a value closer to 1. The equations displayed in Table 5 have been indicated by Ul-Saufie et al. (2015).

N = Number of sample hourly measurement of a selected sites.

P_i = Predicted values of hourly data.

O_i = Observed values of hourly.

$\overline{O}$ = Mean of the observed values of hourly data.

$\overline{P}$ = Mean of the predicted values of hourly data.

Results and discussion

The descriptive statistics and box plots for maximum daily PM_{10} concentrations in Alor Setar, Klang and Kota Bharu from 2002 to 2017 are shown in Fig. 3. Concentrations of PM_{10} were very high in Klang, Selangor with maximum concentrations 643 μg/m^3 over the threshold limit of 150 μg/m^3, followed by Alor Setar (385 μg/m^3) and Kota Bharu (198 μg/m^3). This relates to the fact that Klang is the 13th busiest shipping port and the 16th busiest port in the world. Klang is one of the densely populated and developed areas in Malaysia as there are many industries and business activities in Port Klang. Alor Setar, Klang and Kota Bharu witnessed high particulate events as well as extreme events that promote the increase in PM_{10} concentrations since the skewness value for Alor Setar (4.03), Klang (4.89) and Kota Bharu (1.72). The distribution is highly skewed, as described in Shaziayani et al. (2018), if the skewness is less than − 1 or greater than + 1. Box plot shows that Alor Setar experienced the highest PM_{10} concentration in 2016. According to the DOE, this condition is affected by land and forest fires in Sumatra Central, Indonesia, carried by the Southwest Monsoon winds. Klang reached the highest PM_{10} level during the haze emergency declared on 11 August 2005 as the Air Pollution Index (API) exceeded 500. Due to massive land and forest fires in Sumatra and Kalimantan, Indonesia, Kota Bharu had suffered degradation in air quality during Southwest Monsoon from August to September 2015.

The MAAQG control values for CO, NO_2, O_3, PM_{10} and SO_2 are 8750 ppb (8-h mean reading), 40 ppb (24-h mean reading), 60 ppb (8-h mean reading), 50 μg/m^3 (24-h mean reading) and 40 ppb (24-h mean reading). The analysed data for Alor Setar, such as mean, median, standard deviation, skewness, kurtosis and maximum data, are listed in Table 6. The mean values for all five air pollutants which are PM_{10}

Table 5 Performance indicator

Performance indicator	Equation
Root mean square error (RMSE)	$\frac{1}{n-1}\sum_{i=1}^{n}(P_i - O_i)^2$
Normalized absolute error (NAE)	$\frac{\sum_{i=1}^{n} Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i}$
Index of agreement (IA)	$1 - \left[\frac{\sum_{i=1}^{n}(P - O_i)^2}{\sum_{i=1}^{n}\left(\lvert P_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert\right)^2}\right]$
Prediction accuracy (PA)	$\frac{\sum_{i=1}^{n}(P_i - \overline{O})^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$
Coefficient of determination (R^2)	$R^2 = \left(\frac{\sum_{i=1}^{n}(P_i - \overline{P})(O_i - \overline{O})}{n.S_{pred}.S_{obs}}\right)^2$

	Alor Setar,Kedah	Klang, Selangor	Kota Bharu, Kelantan
Mean	41.99	75.05	48.73
Median	38	68	45
Standard Deviation	20.84	37.78	18.11
Skewness	4.03	4.89	1.72
Kurtosis	40.05	44.82	6.22
Maximum	385	643	198

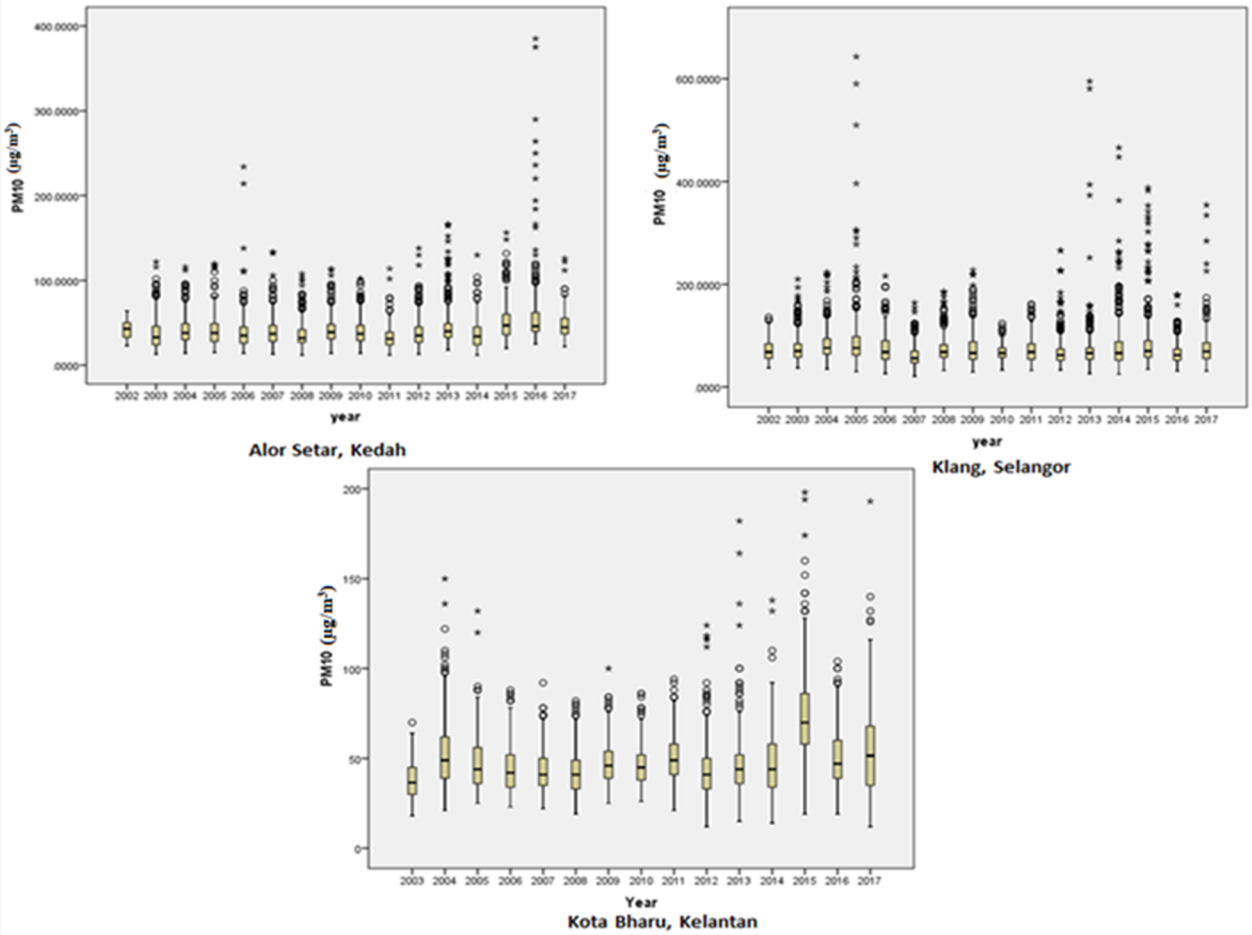

Fig. 3 Descriptive statistics and box plots for maximum daily PM_{10} concentration

Table 6 Descriptive statistics for Alor Setar, Kedah

Parameters	Mean	Median	Standard deviation	Skewness	Kurtosis	Maximum
PM_{10} (µg/m^3)	41.99	38.00	20.84	4.03	40.05	385
O_3 (ppb)	34.27	32.00	14.86	0.82	1.05	118
CO (ppb)	560.30	540.00	246.71	1.71	7.36	3060
NO_2 (ppb)	15.20	14.00	5.85	1.10	2.97	58
SO_2 (ppb)	1.05	1.00	0.93	0.99	2.32	8
RH (%)	89.35	91.00	8.07	−1.77	3.81	100
T (°C)	32.42	32.70	2.77	−1.23	3.21	39.5
WS (km/h)	10.53	10.70	3.74	0.30	1.78	33.5

(41.99 μg/m^3), O_3 (34.27 ppb), CO (560.30 ppb), NO_2 (15.20 ppb) and SO_2 (1.05 ppb) indicate that the average concentration in Alor Setar for 16 years was below the Malaysia Ambient Air Quality Guidelines (MAAQG) for the period from 2002 to 2017. Furthermore, the mean values for meteorological parameters are represented by RH (89.35%), T (32.42 °C) and WS (10.53 km/h). Skewness shows positive values for all air pollutant values. The highest positive skewness value for CO, NO_2, O_3, PM_{10} and SO_2 is 1.71, 1.10, 0.82, 4.03 and 0.99 indicating the existence of extreme events.

Table 7 gives the summary of the descriptive statistics for all parameters' maximum daily data of Klang for 2002 to 2017. The mean values for the area in 16 years are higher than their respective median which indicates that the pollutant distributions are positively skewed (also called right-skewed). The maximum value for air pollutants was PM_{10} 643 μg/m^3, O_3 127ppb, CO 10,500 ppb, NO_2 128 ppb and SO_2 150 ppb. Klang has the highest mean and median values compared to other locations. This may be due to the fact that extensive industry operates in Port Klang, the most densely populated and developed region in Malaysia (AL-Dhurafi et al. 2017, 2018). It has the smallest standard deviation, despite the highest central tendency value, indicating that this area has continuously encountered very high concentrations.

Table 8 demonstrates the result of the descriptive analysis of air pollutant concentration and meteorological parameter for Kota Bharu, Kelantan. The mean values for PM_{10} (48.73 μg/m^3), O_3 (29.21 ppb), CO (926.26 ppb) and NO_2 (15.15 ppb) were higher than the median value. Therefore, the distributions of these measurements were skewed to the right, indicating that there were several observations of high concentration of air pollutant occurred in the years 2002–2017. Meanwhile, the mean value for RH (91.86%) and T (31.36 °C) was lower than the median value which indicates the distribution of data was skewed to the left. These results show that the weather in Kota Bharu is mainly hot and dry, which means that the observation of humidity this year seems to be less humid.

The relative influence (RI) was computed to identify the strength of each predictor-response variable relationship. According to Sayegh et al. (2016), the BRT modelling technique can be used to identify the influence of different predictors on response variable. The most important predictor identified for the maximum daily PM_{10} concentration for the next day (D + 1) was PM_{10} concentration for the previous day, where Alor Setar has 90.17%, Kota Bharu 59.72% and Klang 54.68%. PM_{10} concentration for the previous day played a remarkable role in explaining more than 50% of the variance in the BRT model. The least important predictor was found to be SO_2, where Alor Setar has 0.30%, Kota Bharu 2.77% and Klang 3.02% (Fig. 4).

The BRT models using OLS loss function and compared test, 10-fold CV and OOB methods are shown in Table 9. Performance indicator has been used to assess the accuracy of the fit to the BRT model in order to determine which method better predicts PM_{10} concentration in Alor Setar, Klang and Kota Bharu for the 3 days ahead. This study predicts up to 3 days ahead because, according to Perimula (2012), the government will be able to announce warning status if the API exceeds 101 for more than 72 h.

The best OLS loss function in BRT models with the lowest total ranking is shown in Table 10. For error measurements, the values are ranked from the smallest (rank = 1) to the largest (rank = 3), and for accuracy measurements, the values are ranked from the largest (rank = 1) to the smallest (rank = 3). The total ranking has been determined. This procedure was repeated until the next 3-day (D + 3) prediction to decide the best BRT models for the three stations in this study.

The results show that for the next-day prediction independent test set is better than OOB and CV for all sites. The coefficient of determination (R^2) for Alor Setar, Klang and Kota Bharu was 0.70, 0.60 and 0.65, respectively, while the

Table 7 Descriptive statistics for Klang, Selangor

Parameters	Mean	Median	Standard deviation	Skewness	Kurtosis	Maximum
PM_{10} (μg/m^3)	75.05	68	37.78	4.89	44.82	643
O_3 (ppb)	44.74	42	19.33	0.66	0.48	127
CO (ppb)	1611.43	1440	774.87	2.65	16.04	10,500
NO_2 (ppb)	38.34	37	12.67	0.36	0.89	128
SO_2 (ppb)	6.60	5	6.52	8.67	119.11	150
RH (%)	83.71	84	6.93	−0.71	1.37	100
T (°C)	33.34	33.6	2.22	−0.74	0.74	38.5
WS (km/h)	9.15	9.60	5.02	25.33	1326.95	271

Table 8 Descriptive statistics for Kota Bharu, Kelantan

Parameters	Mean	Median	Standard deviation	Skewness	Kurtosis	Maximum
PM_{10} (μg/m^3)	48.73	45	18.11	1.72	6.22	198
O_3 (ppb)	29.21	29	10.99	0.23	−0.05	69
CO (ppb)	926.26	850	475.32	16.44	689.59	21,712
NO_2 (ppb)	15.15	14	6.23	1.22	4.15	63
SO_2 (ppb)	0.903112	1	1.54	19.54	836.17	71.4
RH (%)	91.86	92	6.80	−5.91	59.70	100.2
T (°C)	31.36	31.7	2.33	−0.63	0.21	37.5
WS (km/h)	9.50	9.8	8.04	31.07	1346.61	360

RMSE value was 10.35, 22.13 and 10.27, respectively. The R^2 values between the fitted model data and the data set were found to be more than 0.5, suggesting that the model is appropriate and good for the next day's prediction by using an independent test set. The R^2 between the observations and the fitted model obtained from this study indicates how well the BRT model fits.

A comparison among the performances of the lowest error (NAE and RMSE) value and comparable IA, PA and R^2 values as for Alor Setar (independent test set), Klang (CV)

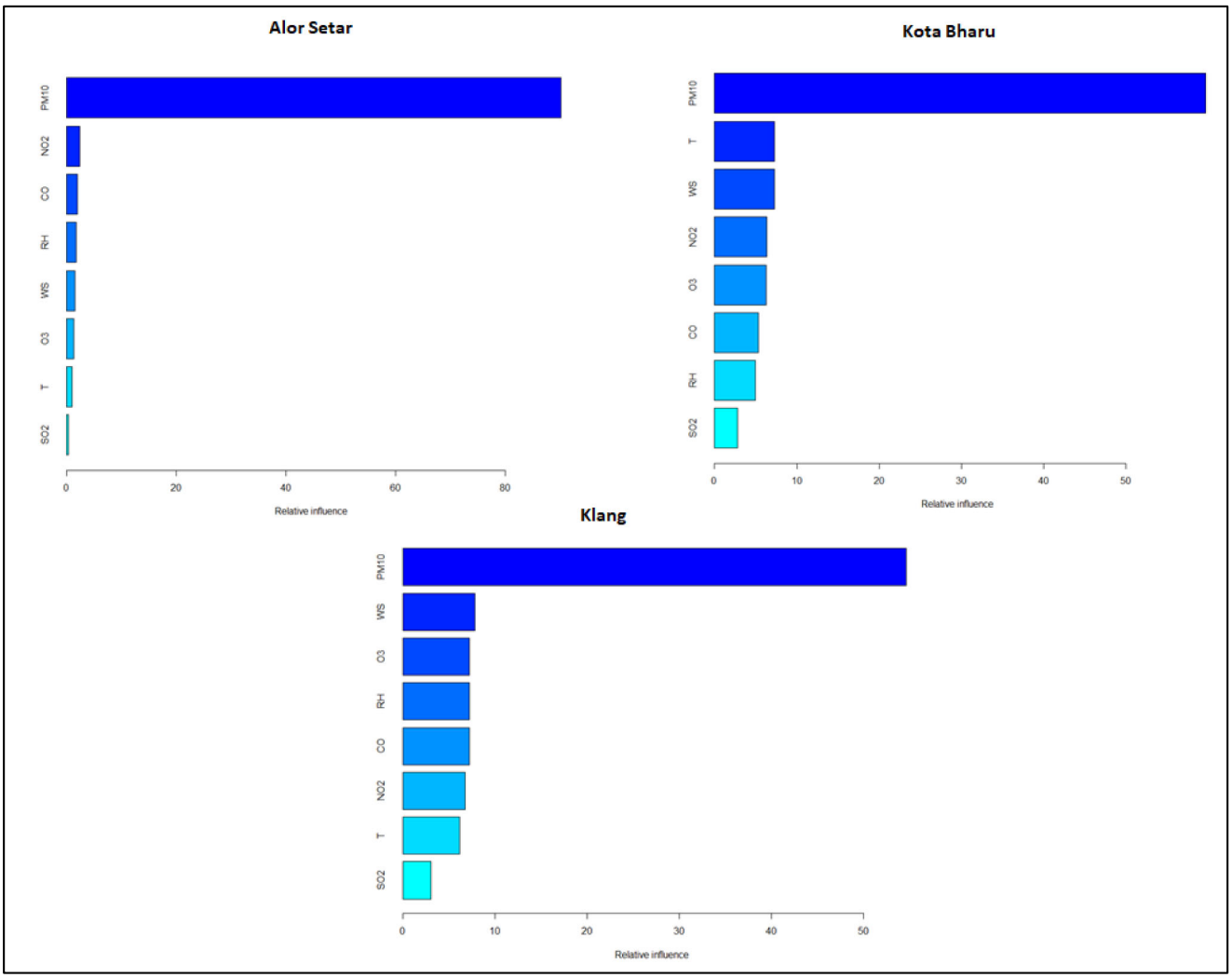

Fig. 4 Relative influence of the selected predictors

Table 9 Comparison method of best iteration (OLS)

Station	Predicted day	Method	Best iteration	RMSE	NAE	IA	PA	R^2
Alor Setar	Next day	OOB	256	10.0661	0.1529	0.9115	0.8370	0.6992
		CV	663	10.4801	0.1541	0.9122	0.8376	0.7003
		Test	**440**	**10.3452**	**0.1528**	**0.9127**	**0.8381**	**0.7011**
	Next 2-day	OOB	236	13.1370	0.2224	0.7816	0.6504	0.4223
		CV	256	13.2581	0.2222	0.7877	0.6507	0.4226
		Test	**350**	**13.2971**	**0.2224**	**0.7903**	**0.6516**	**0.4238**
	Next 3-day	OOB	230	14.7003	0.255	0.6627	0.5345	0.2852
		CV	**465**	**14.7194**	**0.2549**	**0.6876**	**0.5415**	**0.2927**
		Test	322	14.7446	0.2561	0.6769	0.5361	0.2869
Klang	Next day	OOB	255	22.4456	0.1753	0.8441	0.7729	0.5963
		CV	998	22.1405	0.1735	0.8621	0.7764	0.6018
		Test	**991**	**22.1348**	**0.1734**	**0.8623**	**0.7766**	**0.6020**
	Next 2-day	OOB	231	27.0512	0.2312	0.7043	0.6261	0.3913
		CV	**391**	**26.8149**	**0.2289**	**0.7303**	**0.6304**	**0.3967**
		Test	912	27.2296	0.2309	0.7345	0.6164	0.3793
	Next 3-day	OOB	233	30.3456	0.2519	0.6024	0.5378	0.2887
		CV	**378**	**30.1907**	**0.2508**	**0.6327**	**0.5400**	**0.2911**
		Test	2406	30.4640	0.2508	0.6533	0.5310	0.2815
Kota Bharu	Next day	OOB	341	10.2873	0.1527	0.8845	0.8086	0.6527
		CV	708	10.2792	0.1511	0.8898	0.8091	0.6534
		Test	**412**	**10.2702**	**0.1518**	**0.8868**	**0.8092**	**0.6535**
	Next 2-day	**OOB**	**253**	**13.7274**	**0.2091**	**0.7390**	**0.6435**	**0.4133**
		CV	842	13.7969	0.2068	0.7605	0.6373	0.4054
		Test	558	13.7114	0.2062	0.7594	0.6422	0.4116
	Next 3-day	OOB	247	15.4608	0.2294	0.6409	0.5508	0.3028
		CV	595	15.3865	0.2270	0.6720	0.5531	0.3053
		Test	**565**	**15.3831**	**0.2271**	**0.6709**	**0.5533**	**0.3056**

and Kota Bharu (OOB) indicates that the best method for each site is different for the second-day prediction.

Table 10 Ranking of performance indicators for the BRT model to predict D + 1 PM_{10} concentration

Station	Method	Best iteration	RMSE	NAE	IA	PA	R^2	Sum
Alor Setar	OOB	256	1	2	3	3	3	12
	CV	663	3	3	2	2	2	12
	Test	**440**	**2**	**1**	**1**	**1**	**1**	**6**
Klang	OOB	255	3	3	3	3	3	15
	CV	998	2	2	2	2	2	10
	Test	**991**	**1**	**1**	**1**	**1**	**1**	**5**
Kota Bharu	OOB	341	3	3	3	3	3	15
	CV	708	2	1	1	2	2	8
	Test	**412**	**1**	**2**	**2**	**1**	**1**	**7**

However, the next 3-day prediction suggests that the CV is the best method for Alor Setar and Klang, but for Kota Bharu independent test set is the best method which predicts PM_{10} concentration. Overall, the model's performance verified that the next-day prediction is better than the next 2-day and next 3-day prediction.

Descriptive analysis shows that the data for this study is non-central condition because it contains outlier; therefore, this study uses quantile regression as explained by Kudryavtsev (2009). Performance indicators have been used to identify the best quantile to predict the next-day (D + 1) PM_{10} concentration at Alor Setar as summarized in Table 11. Of the five performance indicators used, NAE and IA indicate that 0.5 quantile gave better fit than other quantiles, but the valley differed by just 0.01 with 0.55 quantile. However, RMSE, PA and R^2 have shown that 0.55 quantile is the best quantile in PM_{10} concentration models. 0.55 quantile was therefore used to predict the PM_{10} concentration models for the OOB method. For CV and OOB

Table 11 Performance indicators for PM_{10} concentration prediction (D + 1)

Method	Quantile	Best iteration	NAE	RMSE	IA	PA	R^2
OOB	0.1	537	0.267904	15.991054	0.685428	0.793510	0.628524
	0.2	546	0.201553	12.512929	0.818682	0.831978	0.690941
	0.3	650	0.165560	10.207522	0.894825	0.853560	0.727252
	0.4	588	0.150423	9.475352	0.914884	0.856961	0.733060
	0.5	407	0.145459	9.345616	0.918590	0.854599	0.729024
	0.55	**318**	**0.146444**	**9.326012**	**0.917657**	**0.854638**	**0.729091**
	0.6	312	0.148931	9.461313	0.917807	0.853149	0.726552
	0.65	307	0.154895	9.727199	0.915689	0.851896	0.724420
	0.7	301	0.164896	10.222952	0.910567	0.848863	0.719271
	0.8	286	0.208543	12.432113	0.882167	0.834148	0.694549
	0.9	277	0.324641	18.010877	0.802655	0.803960	0.645188
CV	0.1	3466	0.230134	13.12464	0.821847	0.844543	0.711969
	0.2	3228	0.186435	10.99957	0.881592	0.851891	0.724412
	0.3	2047	0.163486	9.97347	0.907356	0.853797	0.727657
	0.4	944	0.150205	9.49027	0.917772	0.856138	0.731653
	0.5	**706**	**0.145297**	**9.484484**	**0.920923**	**0.853164**	**0.726579**
	0.55	921	0.146527	9.655575	0.920727	0.851436	0.723638
	0.6	716	0.14933	9.922706	0.918218	0.848511	0.718673
	0.65	791	0.155147	10.32006	0.914679	0.846717	0.715639
	0.7	793	0.16241	10.73059	0.90988	0.844388	0.711707
	0.8	521	0.198085	12.46628	0.887964	0.837048	0.699388
	0.9	746	0.284587	17.64743	0.825464	0.816684	0.665772
Test	0.1	2379	0.236263	13.54239	0.804801	0.841709	0.707198
	0.2	2161	0.188082	11.07842	0.878106	0.852244	0.725012
	0.3	2060	0.163522	9.972937	0.907353	0.853876	0.727791
	0.4	907	0.15027	9.484897	0.917661	0.856346	0.732008
	0.5	**668**	**0.145298**	**9.477619**	**0.920683**	**0.853023**	**0.726338**
	0.55	600	0.146047	9.519797	0.921000	0.852851	0.726046
	0.6	806	0.149524	9.954591	0.918100	0.84827	0.718266
	0.65	634	0.154392	10.15938	0.915851	0.848536	0.718717
	0.7	843	0.162507	10.75863	0.909660	0.844025	0.711095
	0.8	1047	0.200318	13.27227	0.880712	0.829483	0.686802
	0.9	541	0.288694	17.47310	0.825032	0.816543	0.665542

methods, the presented results demonstrate that 0.5 gave better fit than other quantiles.

After choosing the right quantile to present the best PM_{10} concentration prediction models for the next day, repeat the same process for finding the best quantile for the next 2-day and next 3-day prediction for all selected locations.

The best quantile to predict the next 2-day (D + 2) PM_{10} concentration at Alor Setar is reported in Table 12. Results show that 0.5 is the best quantile for CV and Test method for the next 2-day prediction, while for OOB is 0.4. The chosen quantile for the next 3-day at Alor Setar is described in Table 13. The findings revealed that all methods (OOB, CV and Test) have the same result, which is 0.55 as the best quantile.

After selecting the best weighting for OOB, CV and Test, the next step is to determine the best method for the next-day, the next 2-day and the next 3-day prediction. The best weighting function was identified for the next day, the next 2 days and the next 3 days in Table 14 for all three monitoring stations by repeating the same procedure for the proposed PM_{10} concentration prediction method.

The best prediction model for next-day PM_{10} concentration in Alor Setar is OOB (quantile = 0.55) with an error of 0.1464 (NAE) and 9.3260 (RMSE), with an accuracy of 0.9177 (IA), 0.8546 (PA) and 0.7291 (R^2). For Klang and Kota Bharu, CV (quantile 0.5) is the best method. The CV and Test models were selected to predict the PM_{10} concentration for the next 2-day while for the next 3-day only Alor Setar shows that OOB

Table 12 Performance indicators for PM_{10} concentration prediction (D + 2)

Method	Quantile	Best iteration	NAE	RMSE	IA	PA	R^2
OOB	0.1	299	0.363715	20.51956	0.526868	0.616668	0.379594
	0.2	298	0.288589	17.02283	0.590397	0.656111	0.429707
	0.3	289	0.250096	15.08631	0.656008	0.667207	0.444363
	0.4	**281**	**0.226822**	**13.6431**	**0.720543**	**0.676696**	**0.457092**
	0.5	281	0.216947	12.79989	0.767547	0.675379	0.455315
	0.55	281	0.217607	12.68753	0.781193	0.672139	0.450957
	0.6	278	0.221589	12.72408	0.788015	0.672194	0.451032
	0.65	272	0.230308	13.03989	0.786863	0.666954	0.444027
	0.7	268	0.242643	13.4476	0.785895	0.669762	0.447773
	0.8	266	0.289563	15.60705	0.758638	0.665199	0.441693
	0.9	253	0.424113	22.04045	0.665915	0.626922	0.392324
CV	0.1	873	0.337707	18.91487	0.572998	0.661126	0.436301
	0.2	769	0.274776	16.11904	0.65214	0.671201	0.449699
	0.3	913	0.24255	14.42645	0.715937	0.670518	0.448784
	0.4	574	0.224878	13.40232	0.755227	0.670488	0.448744
	0.5	**648**	**0.216564**	**12.84958**	**0.790077**	**0.670562**	**0.448844**
	0.55	743	0.217634	12.86845	0.798897	0.669231	0.447063
	0.6	717	0.220777	12.94091	0.800999	0.669613	0.447574
	0.65	701	0.228319	13.21856	0.800638	0.669104	0.446893
	0.7	598	0.241782	13.77886	0.796029	0.669002	0.446757
	0.8	430	0.28651	15.96568	0.765365	0.663513	0.439456
	0.9	728	0.398319	22.43355	0.68394	0.633522	0.400628
Test	0.1	1872	0.326885	18.36329	0.597421	0.66398	0.440075
	0.2	1445	0.270644	15.86809	0.671778	0.667623	0.444918
	0.3	1120	0.242326	14.39568	0.719462	0.669347	0.447218
	0.4	1861	0.224681	13.38723	0.769915	0.664307	0.440509
	0.5	**1244**	**0.217868**	**12.94139**	**0.79478**	**0.668038**	**0.445471**
	0.55	3147	0.219849	12.98603	0.799145	0.665574	0.44219
	0.6	2819	0.222878	13.13031	0.803858	0.66691	0.443968
	0.65	1625	0.230164	13.45881	0.801663	0.66566	0.442305
	0.7	1877	0.241219	14.02964	0.79772	0.66578	0.442465
	0.8	598	0.286518	16.1793	0.766589	0.662632	0.43829
	0.9	716	0.398211	22.38851	0.684235	0.633931	0.401145

(quantile = 0.55) is the best method with performance indicators 0.2463 (NAE), 14.4598 (RMSE), 0.6496 (IA), 0.5553 (PA) and 0.3078 (R^2). Overall, the results showed that quantile values of 0.5, 0.55 and 0.6 obtained the best quantile results when combined with the BRT method.

The best loss function representing each monitoring station can be identified according to the results of the performance indicator in Table 15. Of the five performance indicators applied, all sites indicate that QR was slightly better than OLS. This is supported by Khan et al. (2019), which states that QR can be utilized for the prediction of extreme events.

Norazrin et al. (2018) investigated the Bayesian regression model using conjugate prior distribution and get the results for RMSE (4.66 to 9.88), IA (0.900 to 0.929), PA(0.830 to 0.866) and R^2 (0.614 to 0.665). While, Park et al. (2018) predicted PM_{10} concentration in Seoul metropolitan subway stations using artificial neural network (ANN) model and presented R^2 of 0.39 to 0.81. On the other hand, Abdullah et al. (2020) showed the results from performance error RMSE (126.73–164.98) and NAE (0.33–0.43) by using multiple linear regression for PM_{10} forecasting during episodic trans-boundary haze event in Malaysia. In addition, Shaziayani et al. (2018) reported that feed forward back propagation performs better than general regression neural network in Seberang Jaya, Pulau Pinang with an IA of as much as 0.7796 for the next day, 0.6033 for the next 2-day and 0.8024 for the next 3-day predictions.

Table 13 Performance indicators for PM_{10} concentration prediction (D + 3)

Method	Quantile	Best iteration	NAE	RMSE	IA	PA	R^2
OOB	0.1	300	0.380873	21.52871	0.489821	0.521352	0.271318
	0.2	299	0.309502	18.46747	0.515503	0.535856	0.286625
	0.3	286	0.27178	16.6726	0.548701	0.539934	0.291003
	0.4	283	0.252316	15.47293	0.589012	0.547938	0.299696
	0.5	272	0.244396	14.64987	0.633554	0.55345	0.305755
	0.55	**270**	**0.246321**	**14.45983**	**0.649632**	**0.555341**	**0.307848**
	0.6	270	0.253148	14.47623	0.662903	0.552334	0.304523
	0.65	266	0.26346	14.68193	0.66915	0.550629	0.302647
	0.7	260	0.281239	15.15586	0.670203	0.549539	0.301449
	0.8	253	0.33933	17.32438	0.651928	0.540822	0.291962
	0.9	247	0.487294	23.57309	0.593902	0.524174	0.274263
CV	0.1	992	0.35873	20.4491	0.512107	0.530793	0.281234
	0.2	693	0.297879	17.83777	0.550349	0.539899	0.290966
	0.3	1533	0.261724	16.09796	0.610221	0.542735	0.29403
	0.4	828	0.246604	15.16012	0.636867	0.550614	0.30263
	0.5	1262	0.243505	14.5868	0.681152	0.551573	0.303685
	0.55	**1594**	**0.245915**	**14.51953**	**0.692946**	**0.553111**	**0.30538**
	0.6	1068	0.253651	14.64835	0.694083	0.548511	0.300322
	0.65	808	0.263544	14.90997	0.694612	0.546833	0.298488
	0.7	810	0.278159	15.31503	0.695164	0.54996	0.301911
	0.8	828	0.333233	17.60118	0.675653	0.540595	0.291716
	0.9	697	0.460661	23.56437	0.612804	0.532199	0.282726
Test	0.1	926	0.35965	20.48675	0.511432	0.531874	0.28238
	0.2	1917	0.291029	17.51332	0.571968	0.534348	0.285014
	0.3	3558	0.259282	15.95166	0.6225	0.54401	0.295414
	0.4	1872	0.244724	15.05166	0.649943	0.553874	0.306224
	0.5	1878	0.243075	14.56068	0.684663	0.553187	0.305465
	0.55	**2952**	**0.245934**	**14.5097**	**0.695175**	**0.554395**	**0.3068**
	0.6	2899	0.252724	14.6662	0.698867	0.550385	0.302378
	0.65	1724	0.263638	14.94919	0.698801	0.54829	0.300081
	0.7	974	0.278018	15.33929	0.696575	0.549713	0.30164
	0.8	1704	0.33469	17.89451	0.678132	0.538551	0.289514
	0.9	1030	0.456324	23.71731	0.613868	0.528703	0.279023

Overall, this implies that the values of performance indicators of this study are almost the same as those of previous researchers. This paper shows that alpha 0.5, 0.55 and 0.60 are the best quantile as recommended by Ul-Saufie et al. (2012), which is appropriate for data on air pollution in Malaysia. Therefore, the proposed model can be used as an alternative method to predict the concentration of PM_{10} in Malaysia.

Figure 5 shows the comparison between the observed value and predicted value of Alor Setar, Kota Bharu and Klang for the validate data set. The optimum setting value from the training data set is tuned with the number of learning rate at 0.01 and iteration at 10,000. By using the optimum value found in the training process, the accuracy of this BRT prediction is found to be 60.33 to 91.77%.

Conclusion

Overall, these results indicate that the quantile regression has fulfilled the assumptions and the good model for BRT for predicting maximum daily PM_{10} concentration. The study findings show that the values of NAE (0.15–0.17), RMSE (9.33–22.25), R^2 (0.60–0.73), IA (0.85–0.92) and PA (0.78–0.85) were good for the next-day predictions. Most of the results used 0.5 as the best quantile which represents the median data, but 0.55 and 0.6 had also been chosen as the best

Table 14 Comparing the result between quantile regression

Station	Predicted day	Method	Alpha	Best iteration	NAE	RMSE	IA	PA	R^2
Alor Setar	Next day	**OOB**	**0.55**	**318**	**0.1464**	**9.3260**	**0.9177**	**0.8546**	**0.7291**
		CV	0.5	706	0.1453	9.4845	0.9209	0.8532	0.7266
		Test	0.5	668	0.1453	9.4776	0.9207	0.8530	0.7263
	Next 2-day	OOB	0.4	281	0.2268	13.6431	0.7205	0.6767	0.4571
		CV	**0.5**	**648**	**0.2166**	**12.8496**	**0.7901**	**0.6706**	**0.4488**
		Test	0.55	2952	0.2459	14.5097	0.6952	0.5544	0.3068
	Next 3-day	**OOB**	**0.55**	**270**	**0.2463**	**14.4598**	**0.6496**	**0.5553**	**0.3078**
		CV	0.55	1594	0.2459	14.5195	0.6929	0.5531	0.3053
		Test	0.55	2952	0.2459	14.5097	0.6952	0.5544	0.3068
Klang	Next day	OOB	0.65	295	0.1807	22.8858	0.8377	0.7673	0.5877
		CV	**0.5**	**1248**	**0.1653**	**22.2483**	**0.8509**	**0.7774**	**0.6033**
		Test	0.4	2360	0.1659	22.6276	0.8402	0.7835	0.6127
	Next 2-day	OOB	0.65	263	0.2420	27.3318	0.6974	0.6262	0.3914
		CV	0.6	825	0.2340	26.7720	0.7310	0.6337	0.4008
		Test	**0.6**	**1365**	**0.2338**	**26.6723**	**0.7409**	**0.6366**	**0.4046**
	Next 3-day	OOB	0.65	251	0.2579	30.3537	0.5950	0.5459	0.2974
		CV	0.6	520	0.2490	30.0872	0.6151	0.5491	0.3010
		Test	**0.6**	**917**	**0.2489**	**30.0279**	**0.6272**	**0.5487**	**0.3006**
Kota Bharu	Next day	OOB	0.6	303	0.1568	10.4668	0.8816	0.8069	0.6498
		CV	**0.5**	**1301**	**0.1483**	**10.2735**	**0.8917**	**0.8097**	**0.6544**
		Test	0.6	536	0.1542	10.4305	0.8891	0.8093	0.6538
	Next 2-day	OOB	0.6	279	0.2088	13.5492	0.7542	0.6608	0.4359
		CV	0.4	1049	0.2061	13.8734	0.7520	0.6623	0.4378
		Test	**0.5**	**677**	**0.2006**	**13.4385**	**0.7667**	**0.6625**	**0.4380**
	Next 3-day	OOB	0.5	288	0.2269	15.5590	0.6379	0.5548	0.3072
		CV	0.5	648	0.2244	15.4857	0.6639	0.5550	0.3075
		Test	**0.6**	**1332**	**0.2286**	**15.4000**	**0.6871**	**0.5557**	**0.3083**

Table 15 Comparing the best performance of statistical models for predicting PM_{10} concentration

Station	Predicted day	Method	Distribution	Best iteration	RMSE	NAE	IA	PA	R^2
Alor Setar	Next day	**OOB**	**QR(0.55)**	**318**	**9.3260**	**0.1464**	**0.9177**	**0.8546**	**0.7291**
		Test	OLS	440	10.3452	0.1528	0.9127	0.8381	0.7011
	Next 2-day	**CV**	**QR(0.5)**	**648**	**12.8496**	**0.2166**	**0.7901**	**0.6706**	**0.4488**
		Test	OLS	350	13.2971	0.2224	0.7903	0.6516	0.4238
	Next 3-day	**OOB**	**QR(0.55)**	**270**	**14.4598**	**0.2463**	**0.6496**	**0.5553**	**0.3078**
		CV	OLS	465	14.7194	0.2549	0.6876	0.5415	0.2927
Klang	Next day	**CV**	**QR(0.5)**	**1248**	**22.2483**	**0.1653**	**0.8509**	**0.7774**	**0.6033**
		Test	OLS	991	22.1348	0.1734	0.8623	0.7766	0.6020
	Next 2-day	**Test**	**QR(0.6)**	**1365**	**26.6723**	**0.2338**	**0.7409**	**0.6366**	**0.4046**
		CV	OLS	391	26.8149	0.2289	0.7303	0.6304	0.3967
	Next 3-day	**Test**	**QR(0.6)**	**917**	**30.0279**	**0.2489**	**0.6272**	**0.5487**	**0.3006**

Table 15 (continued)

Station	Predicted day	Method	Distribution	Best iteration	RMSE	NAE	IA	PA	R^2
		CV	OLS	378	30.1907	0.2508	0.6327	0.5400	0.2911
Kota Bharu	Next day	**CV**	**QR(0.5)**	**1301**	**10.2735**	**0.1483**	**0.8917**	**0.8097**	**0.6544**
		Test	OLS	412	10.2702	0.1518	0.8868	0.8092	0.6535
	Next 2-day	**Test**	**QR(0.5)**	**677**	**13.4385**	**0.2006**	**0.7667**	**0.6625**	**0.4380**
		OOB	OLS	253	13.7274	0.2091	0.7390	0.6435	0.4133
	Next 3-day	**Test**	**QR(0.6)**	**1332**	**15.4000**	**0.2286**	**0.6871**	**0.5557**	**0.3083**
		Test	OLS	565	15.3831	0.2271	0.6709	0.5533	0.3056

quantile because the model has more number of outliers compare to the other models. Overall, the results showed that the number of quantile is greater than the median value (0.5). In conclusion, QR is an alternative loss function for BRT to predict the 3 days ahead of PM_{10} concentration for all sites and suitable for data containing influence outlier. This model can help local authority to take action to reduce the effect of haze in Malaysia.

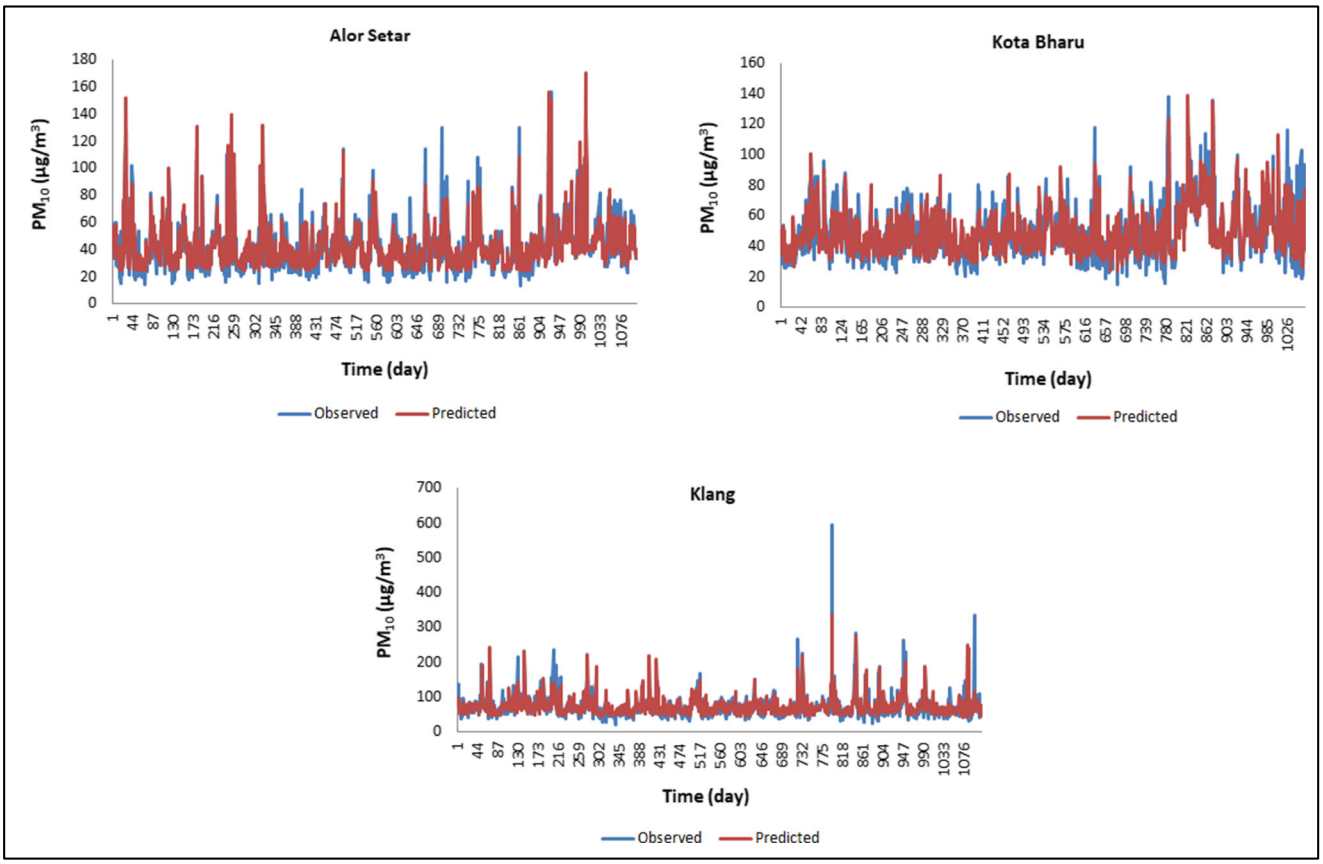

Fig. 5 The observed and predicted maximum daily PM_{10} concentration

Acknowledgements Thank you to Universiti Teknologi MARA for their support and also thanks to the Department of Environment Malaysia for providing air quality monitoring data.

Funding The research was funded by 600-IRMI/FRGS 5/3 (289/2019).

Data availability The data for this project are confidential, but may be obtained with Data Use Agreements with the Department of Environment (DOE), Ministry of Environment and Water of Malaysia.

References

Abdullah S, Ismail M, Fong SY, Ahmed AMAN (2016) Evaluation for long term PM10 concentration forecasting using multi linear regression (MLR) and principal component regression (PCR) models. EnvironmentAsia 9:101–110. https://doi.org/10.14456/ea.2016.13

Abdullah S, Ismail M, Fong SY, Ahmed AMAN (2017) Evaluation for long term PM10 concentration forecasting using multi linear regression (MLR) and principal component regression (PCR) models. Environ Asia 9:101–110

Abdullah S, Napi NNLM, Ahmed AN, Mansor WNW, Mansor AB, Ismail M, Abdullah AM, Ramly ZTA (2020) Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia. Atmosphere 11: 1–14. https://doi.org/10.3390/atmos11030289

AL-Dhurafi N, Masseran N, Zamzuri ZH, Razali AM (2017) Modeling unhealthy Air Pollution Index using a peaks-over- threshold method. Environ Eng Sci 35:101–110

AL-Dhurafi NA, Masseran N, Zamzuri ZH (2018) Compositional time series analysis for Air Pollution Index data. Stochastic Environ Res Risk Assess 32(10):2903–2911

Azmi SZ, Latif MT, Ismail AS, Juneng L, Jemain AA (2010) Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. Air Qual Atmos Health 3:53–64. https://doi.org/10.1007/s11869-009-0051-1

Brunelli U, Piazza V, Pignato L, Sorbello F, Vitabile S (2007). Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy. Atmos Environ, 41:2967–2995

Chelani AB, Gajghate DG, Hasan MZ (2002) Prediction of ambient PM_{10} and toxic metals using artificial neural networks. J Air Waste Manage Assoc 52:805–810

Corani G (2005) Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecol Model 185:513–529

DOE (2018) Department of Environment, Malaysia. Malaysia Environmental Quality Report 2018. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia

Fernando HJS, Mammarella MC, Grandoni C, Fedele P, Di Marco R, Dimitrova R, Hyde P (2012) Forecasting PM_{10} in metropolitan areas: efficacy of neural networks. Environ Pollut 163:62–67

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Friedman JH (2002) Stochastic gradient boosting. Computational Stat Data Anal 38:367–378

Grunwald L, Schneider AK, Schröder B, Weber S (2020) Predicting urban cold-air paths using boosted regression trees. Landscape Urban Planning:201. https://doi.org/10.1016/j.landurbplan.2020.103843

Gu H, Wang J, Ma L, Shang Z, Zhang Q (2019) Insights into the BRT (boosted regression trees) method in the study of the climate-growth relationship of Masson pine in subtropical China. Forests 10:1–20. https://doi.org/10.3390/f10030228

Huijnen V, Wooster MJ, Kaiser JW, Gaveau DLA, Flemming J, Parrington M, Inness A, Murdiyarso D, Main B, Van Weele M (2016) Fire carbon emissions over maritime Southeast Asia in 2015 largest since 1997. Sci Rep 6

Juneng L, Latif MT, Tangang F (2011) Factors influencing the variations of PM10 aerosol dust in Klang Valley, Malaysia during the summer. Atmos Environ 45:4370–4378. https://doi.org/10.1016/j.atmosenv.2011.05.045

Kamarul Zaman NAF, Kanniah KD, Kaskaoutis DG (2017) Estimating particulate matter using satellite based aerosol optical depth and meteorological variables in Malaysia. Atmos Res 193:142–162. https://doi.org/10.1016/j.atmosres.2017.04.019

Khan N, Shahid S, Juneng L, Ahmed K, Ismail T, Nawaz N (2019) Prediction of heat waves in Pakistan using quantile regression forests. Atmos Res 221:1–11. https://doi.org/10.1016/j.atmosres.2019.01.024

Kudryavtsev AA (2009) Using quantile regression for rate-making. Insurance, Math Econ 45:296–304

Latif MT, Othman M, Idris N, Juneng L, Abdullah AM, Hamzah WP, Khan MF, Sulaiman NMN, Jewaratnam J, Aghamohammadi N, Sahani M, Xiang CJ, Ahamad F, Amil N, Darus M, Varkkey H, Tangang F, Jaafar AB (2018) Impact of regional haze towards air quality in Malaysia. A review. Atmos Environ 177:28–44. https://doi.org/10.1016/j.atmosenv.2018.01.002

Leong WC, Kelani RO, Ahmad Z (2020) Prediction of Air Pollution Index (API) using support vector machine (SVM). Jf Enviro Chemical Eng 8:103208

Lingxin H, Naiman DQ (2007). Quantile regression, United Kingdom : Sage Publications

Liu W, Li X, Chen Z, Zeng G, León T, Liang J, Huang G, Gao Z, Jiao S, He X, Lai M (2015) Land use regression models coupled with meteorology to model spatial and temporal variability of NO_2 and PM_{10} in Changsha, China. Atmos Environ 116:272–280

Lu WZ, Wang WJ, Wang XK, Yan SH, Lam JC (2004) Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong. Environ Res 96:79–87

Martinez-Munoz G, Suarez A (2010) Out-of-bag estimation of the optimal sample size in bagging. Pattern Recognit 43:143–152

McKendry IG (2002) Evaluation of artificial neural networks for fine particulate pollution (PM_{10} and $PM_{2.5}$) forecasting. J Air Waste Manage Assoc 52:1096–1101

Motevalli A, Naghibi SA, Hashemi H, Berndtsson R, Pradhan B, Gholami V (2019) Inverse method using boosted regression tree and k-nearest neighbour to quantify effects of point and non-point source nitrate pollution in groundwater. J Cleaner Prod 228:1248–1263. https://doi.org/10.1016/j.jclepro.2019.04.293

Navares R, Aznarte JL (2020) Predicting air quality with deep learning LSTM: towards comprehensive models. Ecol Inform 55:101019

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

Noor NM, Yahaya AS, Ramli NA, Abdullah MMAB (2014) Mean imputation techniques for filling the missing observations in air pollution dataset. Key Eng Mater 594-595:902–908

Noor NM, Yahaya AS, Ramli NA, Abdullah MMAB (2015) Filling the missing data of air pollutant concentration using single imputation methods. Appl Mech Mater 754–755:923–932. https://doi.org/10.4028/www.scientific.net/amm.754-755.923

Norazrin R, Yahaya AS, Hamid AH, Shukri A, Abdul H (2018) Predicting PM10 concentration using Bayesian regression with non-informative prior and conjugate prior model. Engineering Sci Res 3(2):59–65. https://doi.org/10.26666/rmp.jesr.2018.2.9

Park S, Kim M, Kim M, Namgung HG, Kim KT, Cho KH, Kwon SB (2018) Predicting PM10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN). J Hazard Mater 341:75–82. https://doi.org/10.1016/j.jhazmat.2017.07.050

Perez P (2012) Combined model for PM_{10} forecasting in a large city. Atmos Environ 60:271–276

Perimula Y (2012). HAZE: steps taken to reduce hot spots. New Strait Times. Online: http://www.nst.com.my/opinion/letters-to-the-editor/haze-steps-taken-to-reduce-hot-spots-1.98115. Accessed 10 October 2012

Popescu M, Ilie C, Panaitescu L, Lungu ML, Ilie M, Lungu D (2013) Artificial neural networks forecasting of the PM_{10} quantity in London considering the Harwell and Rochester stoke PM_{10} measurements. J Environ Prot Ecol 14:1473–1481

Reddington CL, Yoshioka M, Balasubramaniam R, Ridley D, Toh DY, Arnold SR, Spracklen DV (2014) Environ Res Lett 9:1–12

Ridgeway G (2007). Generalized boosted models: a guide to the gbm package

Ridgeway G (2010) GBM: generalized boosted regression models. R packages version 1:6–3.1

Ridgeway G (2012). gbm: Generalized Boosted Regression Models. R package. TRL, 2007. Primary NO2 Emissions from Road Vehicles in the Hatfield and Bell Common Tunnels. Published Project Report PPR262. TRL, 2011. The Highways Agency Roadside Air Pollution Monitoring Network Report 2010 1

Ridgeway G (2017). Gbm: generalized boosted regression models. R Package Version 2.1.3. https://CRAN.R-project.org/package=gbm

Ridgeway G (2020) Generalized boosted models: a guide to the gbm package. Compute 1:1–12

Sahani M, Zainon NA, Mahiyuddin WWR, Latif MT, Hod R, Khan MF, Tahir NM, Chan CC (2014) A case-crossover analysis of forest fire haze events and mortality in Malaysia. Atmos Environ 96:257–265

Sapini ML, Rahim NZBA, Noorani MSM (2015) The behaviour of PM10 and ozone in Malaysia through non-linear dynamical systems. AIP Conference Proceedings 1682. https://doi.org/10.1063/1.4932452

Sayegh A, Tate JE, Ropkins K (2016) Understanding how roadside concentrations of NOx are influenced by the background levels, traffic density, and meteorological conditions using boosted regression trees. Atmos Environ 127:163–175. https://doi.org/10.1016/j.atmosenv.2015.12.024

Schlink U, Thiem A, Kohajda T, Richter M, Strebel K (2010) Quantile regression of indoor air concentrations of volatile organic compound (VOC). Sci Total Environ 408:3840–3851

Shaziayani WN, Ul-saufie AZ, Ahmat H (2018). A 24-hour forecasting of PM10 concentration in urban area. doi:https://doi.org/10.1063/1.5054208

Ul-Saufie AZ, Yahaya AS, Ramli A, Hamid HA (2012a) Future PM10 concentration prediction using quantile regression models. Ipcbee 37:15–19

Ul-Saufie AZ, Yahaya AS, Ramli A, Hamid HA (2012b) Robust regression models for predicting PM10 concentration in an industrial area. Int J Eng Technol 2:364–370

Ul-Saufie AZ, Yahaya AS, Ramli A, Hamid HA (2015) PM10 concentrations short term prediction using feedforward backpropagation and general regression neural network in a sub-urban area. J Environ Sci Technol 8:59–73. https://doi.org/10.3923/jest.2015.59.73

Viotti P, Liuti G, Di Genova P (2002) Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. Ecol Model 148:27–46. https://doi.org/10.1016/S0304-3800(01)00434-3

Yahaya NZ, Ibrahim ZF, Yahaya J (2019) The used of the boosted regression tree optimization technique to analyse an air pollution data. Int J Recent Technol Eng 8:1565–1575. https://doi.org/10.35940/ijrte.b3807.118419

Zakri NL, Saudi ASM, Juahir H, Toriman ME, Abu IF, Mahmud MM, Khan MF (2018) Identification source of variation on regional impact of air quality pattern using chemometric techniques in Kuching, Sarawak. Int J Eng Technol 7:49