Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter

Suleiman, Aminu; Tight, Miles; Quinn, Andrew

DOI:
10.1007/s10666-016-9507-5

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Suleiman, A, Tight, M & Quinn, A 2016, 'Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter', *Environmental Modeling & Assessment*, vol. 21, no. 6, pp. 731–750. https://doi.org/10.1007/s10666-016-9507-5

Environ Model Assess
DOI 10.1007/s10666-016-9507-5

Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter

A. Suleiman[1] · M. R. Tight[1] · A. D. Quinn[1]

Received: 18 March 2015 / Accepted: 7 March 2016

Abstract This paper examines the application of artificial neural network (ANN) and boosted regression tree (BRT) methods in air quality modelling. The methods were applied to developing air quality models for predicting roadside particle mass concentration (PM_{10}, $PM_{2.5}$) and particle number counts (PNC) based on air pollution, traffic and meteorological data from Marylebone Road in London. Elastic net, Lasso and principal components analysis were used as feature selection methods for the ANN models to reduce the number of predictor variables and improve their generalisation. The performance of the ANN with feature selection (ANN hybrid) and the BRT models was evaluated and compared using statistical performance metrics. The performance parameters include root mean square error (RMSE), fraction of prediction within a factor of two of the observation (FAC2), mean bias (MB), mean gross error (MGE), the coefficient of correlation (R) and coefficient of efficiency (CoE) values. The input variables selected by the elastic net produced the best performing ANN models. The ANN hybrid produced models performed only slightly better than the BRT models. The *R* values of the ANN elastic net and BRT models were 0.96 and 0.95 for PM_{10}, 0.96 and 0.96 for $PM_{2.5}$ and 0.89 and 0.87 for PNC, respectively. Their corresponding CoE values were 0.72 and 0.70 for PM_{10}, 0.74 and 0.76 for $PM_{2.5}$ and 0.81 and 0.71 for PNC respectively. About 80–99% of all the model predictions are within a factor of two of the observed particle concentrations. The BRT models offer more advantages regarding model interpretation and permit feature selection. Therefore, the study recommends the use of BRT over ANN where the model interpretation is a priority.

Highlights Three hybrid ANN and a BRT model for predicting roadside particulate matter were investigated.
Elastic-net regression was found to be a better feature selection method for ANN models than LASSO and PCA.
Hybrid models combining elastic-net and ANN methods were developed, and their performance was better than a standalone ANN model.
The BRT models gave far-reaching information about the relationships between the input variables and the target variables.
Both the ANN and BRT methods produced interpretable models for predicting particulate matter with good model—observation agreement.

Keywords Air quality · Boosted regression trees · Neural network · Particulate matter

1 Introduction

Air pollution from transport and other sources in urban areas is one of the major concerns that affect human health and the urban environment. Several studies have shown that there is a correlation between asthma, bronchitis, pneumonia, and respiratory infections and settlements located near major roads [1–5]. Brunekreef et al. [6] revealed that long-term exposure to traffic-related gaseous and particulate matter pollutants has a strong link to respiratory mortality. The long-term exposure to particulate matter, particularly fine particles ($PM_{2.5}$ or less), is often associated with premature death which accounts for most of the cost of air pollution [7–9].

The effects of traffic-derived air pollution can be effectively controlled by providing adequate and efficient air quality control and mitigation measures that can be designed and tested with the aid of air quality models. Air quality regulatory agencies have to complement measurements of air quality with models that can accurately predict pollutant

✉ A. Suleiman
axs526@bham.ac.uk; alaminsjam@gmail.com

[1] School of Engineering, University of Birmingham, Birmingham, West Midlands B15 2TT, UK

Published online: 30 March 2016

concentrations and determine the cause of the air quality problems. The models are calibrated using historical air pollution data and are used to forecast the likely air quality scenarios for the future. Air quality models currently used by regulatory agencies are mostly deterministic and are built on simple assumptions about atmospheric processes and involve high computational cost which limits their applications. They also require knowledge of the relationships between the variables involved and meteorological conditions. The deterministic models are not only constrained by the accurate characterisation of the dynamics of the natural phenomena but also on the model configuration options, e.g. default parameters and lack of real observations with the same spatial resolution with which to compare the model outputs [10, 11]. Steady-state Gaussian plume models are the most widely used air quality models and have been applied successfully in many air quality studies. However, despite their successful application, they are limited by assumptions regarding change of wind and source emission over time and do not include the detailed chemistry of particle pollutants [12]. In contrast, artificial neural network (ANN) and boosted regression tree (BRT) can be used to build air quality models with comparable prediction accuracy at a lesser computational cost and with no assumption of the atmospheric processes involved [13]. ANNs are capable of handling complex and robustly nonlinear relationships that exist between air quality variables [14] and produce models that can perform extremely well in predicting unseen data. ANN models can handle multivariate inputs, nonlinearity and uncertainty, but they require additional algorithms to perform feature selection. Also, they are regarded by many as black boxes because they estimate input(s)-output(s) relationships internally and give out the final results without revealing the contribution of the respective predictor variables as these are obtained with classical regression methods. However, several researchers have devised means of extracting information from trained neural networks. Olden and Jackson [15] reviewed and compared methods to study the relative importance of variables in the ANN they applied to ecological modelling. However, these methods have been rarely applied in air quality studies [16] and have not yet been incorporated into the widely used ANN toolboxes and packages. The BRT method, on the other hand, is far from being a black box as it provides partial dependence plots. These plots describe the interaction between the input variables and the target variables which are useful in interpreting the models. BRT is also equipped with algorithms that can perform feature selection and produce a plot that shows the relative influence of the input variables in the model development.

The aim of this paper is to examine the application of two machine learning methods (ANN and BRT) in modelling roadside particulate matter. This goal is achieved through (a) investigating the use of principal component analysis, Lasso and elastic-net regressions for determining suitable predictor variables for the ANN models; and (b) developing ANN and BRT models for predicting roadside particle concentrations including PM_{10}, $PM_{2.5}$ and particle number counts (PNC) and comparing their performance. The prediction performance of the models is evaluated using model evaluation functions provided through an open source software for air quality data analysis [17]. In the rest of this paper, Section 2 briefly describes the data, the feature selection techniques and the modelling methods used in the study. The findings of the study are discussed in Section 3, and Section 4 presents the conclusions.

2 Materials and Methods

2.1 Data Collection

The data for this study was obtained from two of the UK's Automatic Urban and Rural Network (AURN) air quality monitoring sites, London Marylebone Road and London Bloomsbury. The Marylebone Road monitoring station is located approximately 1.5 m to the southern side of a busy road in a street canyon aligned on an axis of 75°–255° [18] in Central London. The road consists of three lanes in each direction with an average traffic flow of about 80,000 vehicles per day. There is a light-controlled pedestrian crossing, and a junction located at about 50 and 150 m to the west of the monitoring site respectively [19]. The data comprises hourly traffic flow, air pollutants and meteorological variables for the period between 2000 and 2007. The traffic flow was aggregated into heavy-duty vehicles (HDV) and light-duty vehicles (LDV). The air pollutant variables include PM_{10}, $PM_{2.5}$, PNC, CO, SO_2, NOx, NO and NO_2 concentrations. Meteorological parameters included are wind speed, wind direction, temperature, solar radiation, relative humidity, barometric pressure and rainfall. The London Marylebone Road site traffic data was collected using induction loops buried in each lane, for counting and classification [19]. London Bloomsbury air quality monitoring site is located in the southeast corner of Russell Square Gardens in Central London [20] and approximately 2 km from the London Marylebone Road site. $PM_{2.5}$ and PM_{10} concentrations at the sites were monitored using two similar tapered element oscillating microbalances (TEOM), Model 1400AB with different sampling head designs [20]. The TEOM comprises a filter, tapered hollow glass tube and PM_{10} impactor inlet for measuring PM_{10} mass. Also, a sharp cut cyclone is attached to the TEOM for measuring $PM_{2.5}$ concentrations. The PNC data was collected using scanning mobility particle sizer spectrometer (SMPS) system that comprises electrostatic classifier (EC) model 3071A and a condensation particle counter (CPC) model 3022A for measuring the particle sizes and the particle concentration, respectively [20]. Data from these sites were made available through the London Air Archives [21] and UK Air Quality Archive [22] for download. A graphical representation and descriptive statistics of some of the data for the Marylebone Road site are shown in Fig. 1.

2.2 Air Quality Prediction Models

This section describes the modelling techniques and the general method followed in this study for developing the models for roadside particulate matter prediction. Figure 2 shows the flow chart of the modelling tasks carried out in this study. The data for the modelling was divided into 80 % training and 20 % testing subsets. The training data subset was first used to select the most relevant predictor variables for ANN models using principal components analysis, Lasso and elastic-net regression methods. Subsequently, the data for the selected variables was extracted from the training dataset and then used to train the ANN models. The BRT models have an inbuilt feature selection algorithm in their formulations. Therefore, the whole training data including all the variables was used for their training. When each model was sufficiently trained, it was then tested using the test data set and subsequently its performance was evaluated using various model performance evaluation functions. The final task was the comparison of the performances of the best ANN model and the BRT models for each target pollutant.

2.3 Feature Selection

Feature or variable selection is one of the most important steps in model building. It is required to remove irrelevant model input variables thereby reducing the computational complexity, learning difficulty, memory requirements and model

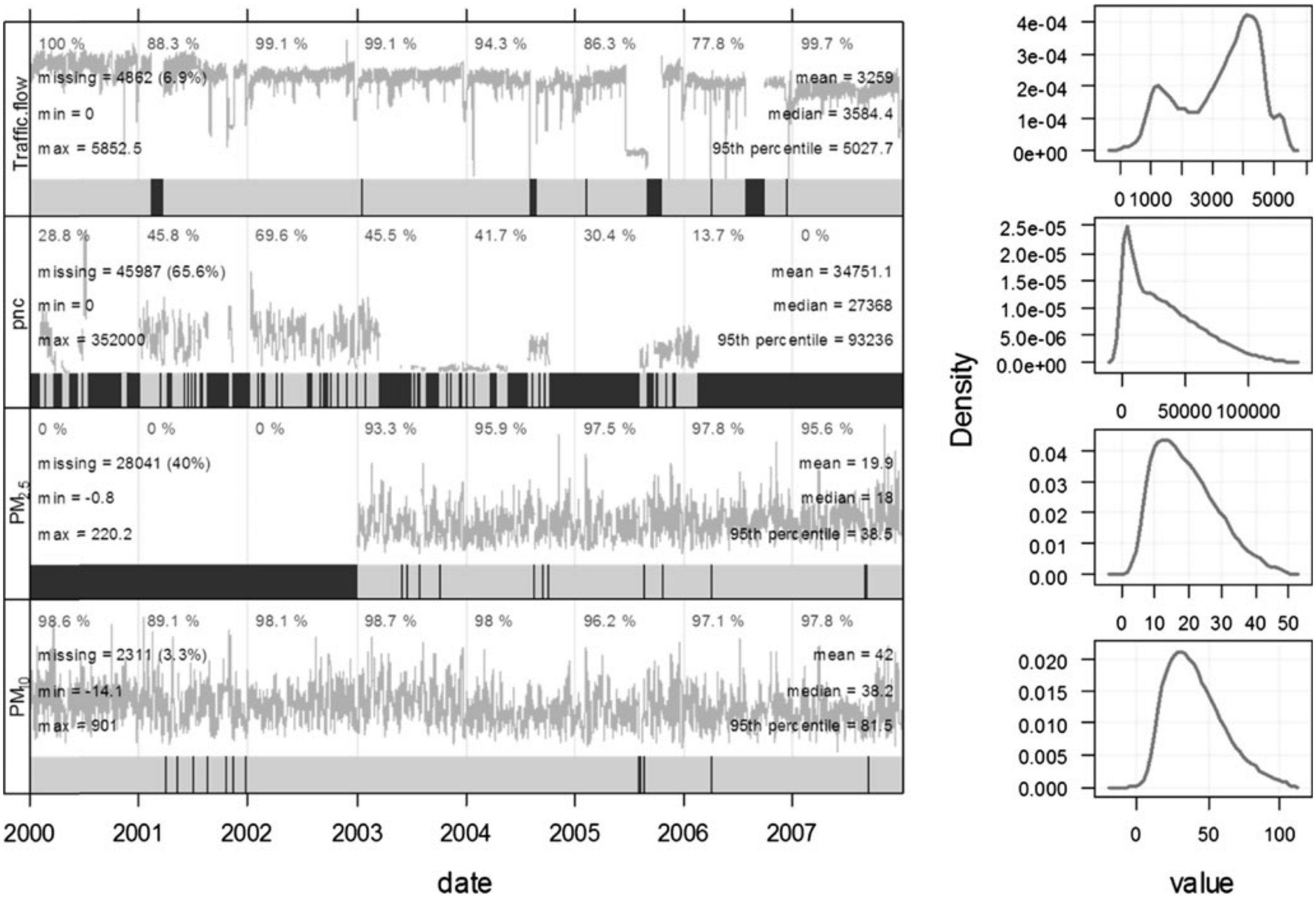

Fig. 1 Summary plot showing descriptive statistics and distribution of some of the Marylebone road data. The *plot* shows in each rectangle in the left panel a time series plot and descriptive statistics of a variable in the data. The *bar* at the lower part of each rectangle represents available and missing data using *grey* and *black colours*, respectively. The percentage of the data captured for every year or month is shown in *grey* in the upper part of each rectangle. The minimum, maximum, number and percent of missing data, mean, median and the 95th percentile for each variable plotted are also shown in *black*. The panels to the right are the density plots indicating the distribution of the data over the selected period. The PM_{10} and $PM_{2.5}$ concentrations were measured in μg/m^3, PNC in number/cm^3, and traffic flow in veh/h

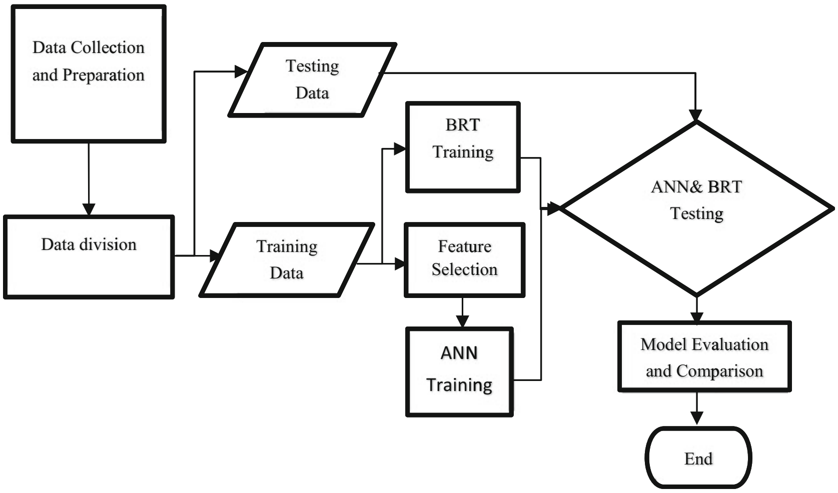

Fig. 2 Flow chart for the modelling process

complexity. It is also possible to improve the prediction accuracy as well as the ability to generalise the model [23, 24]. In this paper, three feature selection methods including principal components analysis (PCA), Lasso, and elastic-net regressions were used. The methods were applied to select the predictor variables with high predictive ability among the available variables prepared for the ANN modelling. ANN with PCA (ANNPCA), ANN with Lasso regression (ANNLASSO) and ANN with elastic-net regressions (ANNELASTICNET) are the three hybrid ANN models developed as a result of the application of the feature selection methods. BRT models have an inbuilt feature selection algorithm in their modelling process. Therefore, the three feature selection methods were applied only to ANN models.

2.3.1 Principal Component Analysis

PCA derives its strength in its ability to transform, using singular value decomposition (SVD), the input space into a set of orthogonal vectors called principal components (PCs). These PCs are then used as predictor variables for ANN models in place of the original input variables. The PCs are derived such that the first PC accounts for the highest variability in the input space followed by the second PC and the subsequent PCs, respectively. The first few PCs that explain most of the variation of the variables are retained, and the remaining PCs discarded. PCA has low noise sensitivity, and the reduction of the dimension of the input space decreases the requirements for capacity and memory. Also, it increases the efficiency of the model training given the processes are taking place in smaller dimensions. The main disadvantages of PCA are that it is hard to evaluate the covariance matrix with the desired accuracy, and even modest invariance could not be captured by the PCA unless the information is explicitly provided in the training data [25, 26]. This method is popular, and it has since been incorporated in many commercial and open source software platforms. Here, the *pcaNNet* function of the *caret* package [27] of R software [28] was used for developing the ANNPCA.

2.3.2 Lasso and Elastic-Net Regressions

Ridge regression is the backbone of Lasso and elastic-net regressions, and it shares similarities with the least square estimate, but its coefficients are estimated using a different quantity. For example, given the training data (x_{ij}, y_i) consisting of input variables x_{ij} and the target variables y_i, the method of least square estimates the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ using values that minimise the residual sum of squares (RSS) shown in Eq. 1,

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \tag{1}$$

where $i = 1, 2, 3\ldots n$, $j = 1, 2, 3\ldots, p$, β_0 is the intercept, and β_j are the coefficients.

The ridge regression coefficient estimates $\left(\hat{\beta}^{\text{ridge}}\right)$ are obtained using Eq. 2,

$$\hat{\beta}^{\text{ridge}} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2}$$

Where $\lambda \geq 0$ is a tuning parameter to be determined separately using cross-validation and the term $\lambda\sum_{j=1}^{p}\beta_j^2$, is called a shrinkage penalty. The shrinkage penalty is small when the coefficients $\beta_1, \ldots, \beta_p$ are close to zero thereby having the effect of shrinking the β_j estimates towards zero. The regularisation parameter λ regulates the impact of the two terms in Eq. 2. When $\lambda = 0$, the shrinkage penalty has no effect and the ridge regression is reduced to a least square. However, as $\lambda \to \infty$, the impacts of the penalty shrinkage grows and the ridge regression coefficient estimates will shrink towards zero [29]. Lasso is a shrinkage method which takes the advantages of ridge regression and feature selection techniques by shrinking the regression coefficients and forcing some to zero. In that way, the models so developed can be easily interpreted (less number of input variables). Lasso minimises the residual sum of squares, subject to a constraint that the sum of the absolute values of the coefficients should be less than or equal to a certain constant [24]. When applied to new data, Lasso would have smaller mean square error than ordinary least-squares estimates. The difference between Lasso and ridge regression is the penalty term, as the ridge regression uses L_2 norm, i.e. $\left(\sum_{j=1}^{p} \beta_j^2\right)$ as its penalty term, Lasso uses L_1 norm, i.e. $\left(\sum_{j=1}^{p} |\beta_j|\right)$ as shown in Eqs. 2 and 3, respectively. Mathematically, Lasso regression can be estimated using Eq. 3.

$$\hat{\beta}^{\text{lasso}} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right| \tag{3}$$

The elastic-net regression is the generalisation of Lasso where the advantages of the ridge and Lasso regressions are sought after as shown in Eq. 4. The elastic net performs shrinkage as well as subset selection where both the penalties are tuned to achieve optimal performance.

$$\hat{\beta}^{\text{elasticnet}} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p}\left|\beta_j\right| \tag{4}$$

where $i = 1, 2, 3, \ldots, n$, $j = 1, 2, 3, \ldots, p$.

β_0 is the intercept, β_j are the coefficients, λ_1 and λ_2 are tuning parameters. The methods are themselves modelling tools that could be used to develop the prediction models, but here, they were used as feature selection methods. In this study, the Lasso and elastic-net regressions shown in Eqs. 3 and 4 were implemented through *glmnet* package [30] of the R statistical software [28]. The *cv.glmnet* function uses k-fold cross-validation to calculate the regularisation path for the Lasso or elastic-net penalty at a grid of values for the regularisation parameter lambda (λ). The user specifies a value of alpha (the elastic-net mixing parameter, with $0 \le \alpha \le 1$) if elastic net was to be used to fit the model [30]. Lasso or elastic-net models with the lowest MSE were selected as the final models. After fitting the Lasso and elastic-net models, the input variables with zero or nearly zero coefficients in the resulting regression models were discarded and the data for the remaining variables were extracted in the training data and then used for the training of the ANN models.

2.4 Artificial Neural Network Models

ANN models are designed to mimic the behaviour of the human brain which comprises interconnected synaptic neurons capable of learning and storing information about their environment [31]. A neuron model comprises three elements, the connecting links characterised by their strength, a linear combiner which combines the weighted input signals and an activation function for limiting the amplitude range of the neuron's output to some finite value. Mathematically, a neuron can be represented by Eq. 5.

$$U_k = \sum_{j=1}^{m} w_{kj} x_j \qquad j = 1, 2, 3, \ldots, \tag{5}$$

$$y_k = f(U_k + b_k) \tag{6}$$

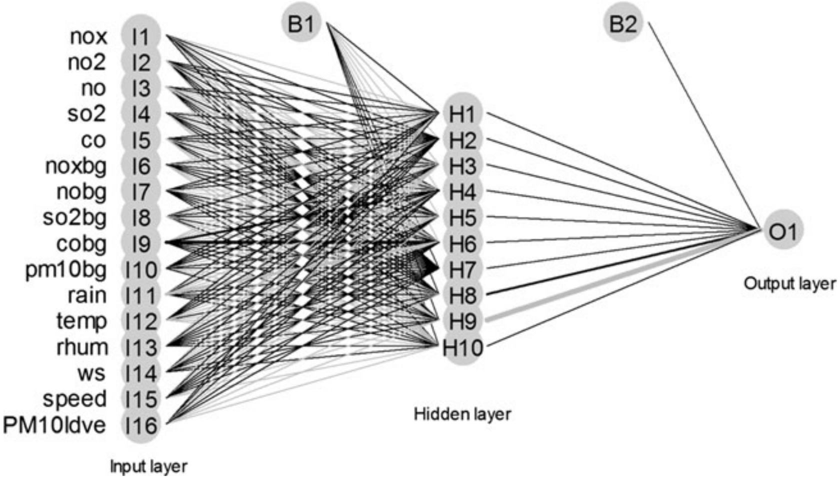

Fig. 3 The typical structure of a multilayer neural network. The suffix *bg* in some of the variable names indicates background concentration

where k represent a neuron, $x_1,\ldots,x_m$ are the input signals, w_{k1}, .., w_{km} are synaptic weights of neuron k, U_k is the linear combiner, b_k is the bias, $f(.)$ is an activation function (Eq. 7) and y_k is the output signal of the neuron k.

$$f(x) = \frac{1}{1+e^{-x}} \tag{7}$$

The neural network model architecture consists of three distinct and interconnected layers of neurons; input layer, hidden layer and output layer (Fig. 3). It processes information sequentially in the order in which the layers are arranged (i.e. from input layers to hidden layers and lastly to the output layer). The output of each layer serves as an input to the next layer [32]. The ANN models are designed to perform a certain task through training on historical data, and the goal of the training is not limited to learning and accurate representation of the sets of training data, but to model statistically the process that generates the data which is necessary for generalisation and accurate prediction [31]. There exist many variants of ANN models; however, in this paper, the multilayer perceptron network (MLP) was selected due to its popularity and availability on various commercial and open source software platforms. The training data set containing the variables selected during feature selection was used to train the ANN models using a supervised back-propagation algorithm. In this type of training, two major processes are involved (forward and backward passes). In a forward pass, the input variables, e.g. air pollutants, traffic and meteorological variables as shown in Fig. 3, are received by the input neurons and passed through connecting links of various weights with which the inputs are weighted. The weighted inputs are then summed up by a linear combiner (Eq. 5) and transmitted forward to the hidden layer neurons. The outputs of the hidden layer estimated by its activation function (Eq. 7) are then passed to the output layer where the final output of the network will be estimated using the output layer activation function as shown in Eq. 6.

The activation function limits the amplitudes of the outputs to certain threshold values to reduce the computational loads of the network and to keep the values of the outputs within a certain margin of the target variables, i.e. particulate matter in this case. The network outputs are then compared with the target output, and their difference is taken as the network error, which will then be propagated backward through the network, to update various weights within the network. The iteration continues until the minimum error is obtained using the gradient descent technique [33, 34]. In this study, the ANN training was implemented using *nnet* [35] and *caret* [27] packages of the R software for statistical analysis [28]. The *train* function of the *caret* package was used to search for the optimum model parameters (number of hidden neurons and weight decay) using 10-fold cross-validation. In this process, several ANN models were trained with three weight decay parameters (0.001, 0.01 and 0.1) while incrementing the number of hidden neurons up to a predetermined number of hidden neurons (50 in this case). The function then selects the ANN model with the minimum mean square error (MSE) as the final model. The final model for each target pollutant was then tested using the testing data, and the test results were evaluated using various model evaluation functions provided in the R openair package [17].

2.5 Boosted Regression Trees Models

BRT derives its strength from two different algorithms; regression trees and gradient boosting. Regression trees are simple models that fit a response variable to predictor variables by partitioning the feature space using a series of partition rules, e.g. binary split, to identify regions in the data having the most consistent responses to predictors. A constant is then fitted to each region (e.g. mean response for observations in a particular region, in a regression problem). Gradient boosting, on the other hand, combines the output of weak learners (regression trees) to produce a more powerful and improved predictive performance. Therefore, the final model (BRT) would be a combination of several individual regression trees fitted in a forward stage-wise manner [36]. BRT for function approximation can be applied to a typical predictive learning system consisting of a set of predictor variables $\mathbf{X} = \{x_1,\ldots,x_n\}$ and a response variable y. For example, using a training sample $\{y_i, \mathbf{X}_i\}$, $i = 1,\ldots,N$ of known y and $\mathbf{X}$ values and we wish to find a function $F^*(\mathbf{X})$ that maps $\mathbf{X}$ to y, such that it minimises the expected value of a specified loss function (Eq. 8) over the joint distribution of all the values of $(y, \mathbf{X})$ [37, 38]. The Gradient boosting approximates $F(\mathbf{X})$ using Eq. 9,

$$F^*(\mathbf{X}) = \psi(y,\ F(\mathbf{X})) \tag{8}$$

$$F(\mathbf{X}) = \sum_{m=0}^{M} F_m(\mathbf{X}) \quad = \quad \sum_{m=0}^{M} \beta_m g(\mathbf{X};\boldsymbol{\alpha_m}) \tag{9}$$

Where $g(\mathbf{X};\boldsymbol{\alpha_m})$ represents a regression tree at a particular node, $\boldsymbol{\alpha_m}$ describes the tree parameters (i.e. splitting variables and split points), β_m are the expansion coefficients, $m = 1$, …, M. During each iteration m, the $\mathbf{X}$ space is split into **N**-disjointed regions $\{R_{nm}\}$, $n = 1,\ldots,N$ and predicts a separate constant in each one. In this study, the BRT algorithm was implemented using the following steps through the *gbm* package [39] of R software.

The user should first select the BRT tuning parameters including a loss function (*distribution*), the number of iterations, T (*n.trees*), the depth of each tree, K (*interaction.depth*), the learning rate parameter λ (*shrinkage*) and the subsampling rate, p(*bag.fraction*)

1. Initialise $F(\boldsymbol{X})$ to be a constant
2. For $m = 1$ *to* M do as follows:

   a. Calculate the residuals $r = -[\partial\psi(y_i, F(X_i)/\partial F(X_i)]_{F_m(\boldsymbol{X})} = F_{m-1}(\mathbf{X}), i = 1, \ldots, N$
   b. Randomly select $p \times N$ samples from the training data without replacement
   c. Fit a least squares regression tree to r in 2a with K terminal nodes to get the estimate $\boldsymbol{\alpha_m}$ of $\beta g(\mathbf{X};\alpha)$ using only randomly selected observations in (b)
   d. Get the estimates β_m by minimising the loss function $\psi(y, F_{m-1}(\mathbf{X})) + \beta g(\mathbf{X};\boldsymbol{\alpha_m})$
   e. Update $F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + \beta_m g(\mathbf{X};\boldsymbol{\alpha_m})$
3. Calculate $F(\mathbf{X}) = \sum_{m=0}^{M} F_m(\boldsymbol{X})$ [40]

Although boosting enhances the capabilities of regression trees, the BRT is susceptible to overfitting, therefore, a learning rate λ is introduced to control the situation by dampening the learning process as shown in Eq. 9.

$$F_m(\boldsymbol{X}) = F_{m-1}(\mathbf{X}) + \lambda\beta_m g(\mathbf{X};\boldsymbol{\alpha_m}) \tag{10}$$

Also, inspired by the concept of 'Bagging' [41], Friedman [37] modified the gradient boosting algorithm with random sub-sampling of the training data to improve its prediction accuracy and computational resource requirement. The *gbm* package allows for the selection of a suitable sub-sampling rate (*bag fraction*) to implement this modification.

The optimum BRT tuning parameters in this work were determined using the *train* function of the *caret* package. The function uses cross-validation to determine the optimum combination of the tuning parameters. For each pollutant, five different learning rates 0.001, 0.01, 0.05, 0.1 and 0.5, the number of trees from 1 to 10,000, tree complexities from 1 to 10 and a fixed bag fraction of 0.5 were tested. The model with the combination of tuning parameters that gives the lowest RMSE value was taken to be the final model in each case. The final models were then tested and evaluated with the same testing data used for testing ANN models.

2.6 Model Evaluation

The following model performance metrics provided in *openair* by Carslaw and Ropkins [17] were used for evaluating the ANN and the BRT prediction models developed in this study.

1. Root mean square error (RMSE): RMSE is a measure of the average error produced by a model, and is among the best measures of overall model performance that can be easily interpreted since they carry the same unit as the modelled and observed values. Although it is sensitive to extreme values, it reveals the actual size of the error produced by the model, unlike R^2 which is affected by the higher and low standard deviations of both observed and modelled values. However, it does not reveal the types or sources of the error which will assist greatly in refining the models [42, 43]. The RMSE is formulated as follows:

$$\text{RMSE} = \sqrt{{}^{1}/_{N}\sum_{i=1}^{N}(M_i - O_i)^2} \tag{11}$$

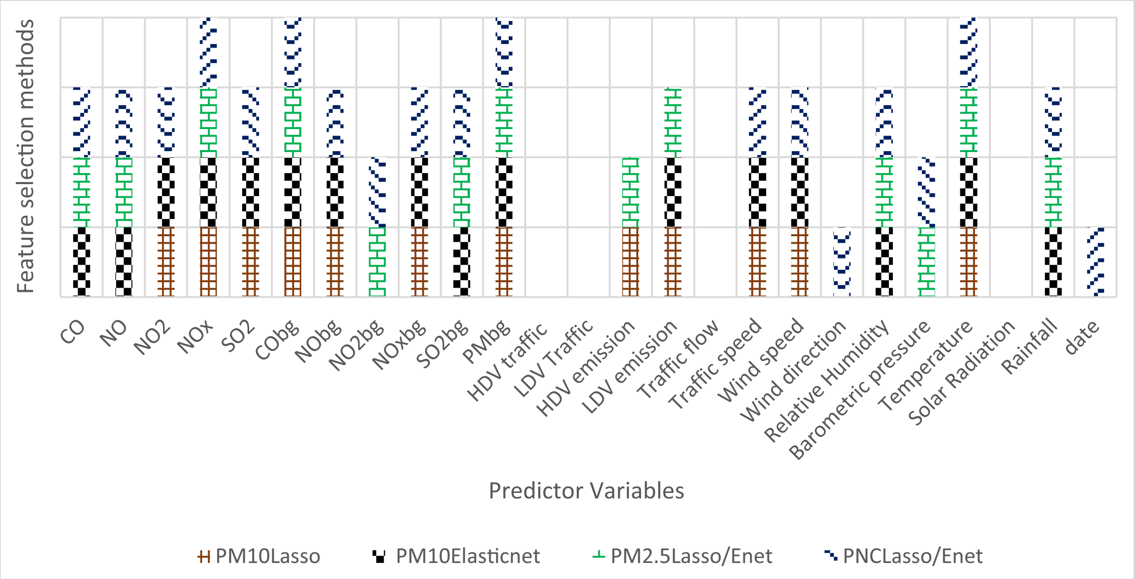

Fig. 4 Comparison of the input variables chosen by the Lasso and elastic-net regressions for PM_{10}, $PM_{2.5}$ and PNC ANN models. The *bar patterns* represent feature selection methods

2. The fraction of predictions within a factor of two of the observations (FAC2) measures the fraction of the model prediction that satisfies the condition in Eq. 11. A model which satisfies the condition would have a value of FAC2 equal to 1. Chang and Hanna [44] described FAC2 as a robust performance measure since it is not affected by outliers.

$$\text{FAC2} = 0.5 \leq M_i / O_i \leq 2.0 \tag{12}$$

3. Mean bias (MB): This is the measure of the model under or over prediction estimated as the difference between the mean observed and the mean predicted values. MB values range from $-\infty\ to + \infty$ with zero being the MB value for an ideal model. Although MB is being used as a model performance measure, its major weakness is that it does not provide more diagnostic value than the mean values of observed and predicted. Cort J. Willmott [42] suggested that the mean values of observed and predicted should be reported instead of MB since they are more familiar to researchers and contain little more information than MB. NMB is a normalised version of MB, and it is often used when comparing different pollutant concentration scales. MB is estimated using Eq. 12.

$$\text{MB} = \frac{1}{N} \sum_{i=1}^{N} (M_i - O_i) \tag{13}$$

4. Coefficient of efficiency (*CoE*):CoE is a measure of model efficiency that is robust and easy to interpret [45]. This measure has interpretation for zero and negative values. A perfect model has CoE value of one. Zero values of CoE indicate that the model's prediction accuracy is not more than the observed mean values of the data and negative CoE values indicate that the model's prediction accuracy is worse than the observed mean. CoE can be estimated using Eq. 13.

$$\text{CoE} = 1- \sum_{i=1}^{N} |(M_i - O_i)| / \sum_{i=1}^{N} \left| \left(O_i - \overline{O}_{i_i} \right) \right| \tag{14}$$

5. Graphical functions which include scatter plots, conditional quantile plots, time variation plots and polar annulus plots provided in the openair package were also used to explore the strengths and weaknesses of the ANN and BRT models developed.

3 Results and Discussion

3.1 ANN Modelling Results

The feature selection methods described in Section 2.3 were used in selecting the most relevant predictor variables for the training of the ANN models. Three hybrid ANN models each with a different feature selection method were expected as a result of using the feature selection methods. The hybrid models include ANN with principal components analysis (ANNPCA), ANN with Lasso regression (ANNLASSO) and ANN with elastic-net regression (ANNELASTICNET). However, it was only for the PM_{10} prediction that the three hybrid ANN models were obtained because Lasso and elastic net selected the same set of variables in the case of $PM_{2.5}$ and PNC prediction. Moreover, the ANNPCA and standalone ANN models require all the input variables for their training. Figure 4 shows the comparison between the variables selected by the feature selection methods. Lasso and elastic-net methods produce regression coefficients from which the decision to keep or remove the variables is taken based on their contribution to the models.

Table 1 Training performance of the neural network models

Models	Pollutants	Decay	NHN size	RMSE	*R*-squared
ANN	PM_{10} (μg/m^3)	0.01	19	12.64	0.69
ANNPCA	PM_{10} (μg/m^3)	0.001	13	11.24	0.76
ANNLASSO	PM_{10} (μg/m^3)	0.01	16	11.43	0.76
ANN ELASTICNET	PM_{10} (μg/m^3)	0.001	21	11.67	0.75
ANN	$PM_{2.5}$ (μg/m^3)	0.1	14	3.064	0.91
ANNPCA	$PM_{2.5}$ (μg/m^3)	0.1	16	3.226	0.90
ANNLASSO/ ELASTICNET	$PM_{2.5}$ (μg/m^3)	0.1	20	3.144	0.91
ANN	PNC (number/cm^3)	0.001	20	3937	0.97
ANNPCA	PNC (number/cm^3)	0.001	11	8155	0.91
ANNLASSO/ ELASTICNET	PNC (number/cm^3)	0.001	16	7141	0.90

NHN number of hidden neurons

Table 2 Test performance of the neural network models

Model	Pollutant	FAC2	MB	NMB	RMSE	*R*	COE	IOA
ANN	PM_{10}	0.99	0.23	0.01	12.46	0.85	0.68	0.84
ANNPCR	PM_{10}	0.99	0.20	0.00	9.61	0.90	0.69	0.84
ANNLASSO	PM_{10}	0.99	0.08	0.00	10.74	0.88	0.69	0.85
ANNELASTICNET	PM_{10}	0.99	0.11	0.00	10.02	0.89	0.70	0.85
ANN	$PM_{2.5}$	0.99	0.00	0.00	3.16	0.95	0.73	0.86
ANNPCA	$PM_{2.5}$	0.99	0.00	0.00	3.11	0.95	0.73	0.87
ANNELASTICNET	$PM_{2.5}$	1.00	0.02	0.00	3.03	0.96	0.74	0.87
ANN	PNC	0.92	30.17	0.00	5735.24	0.97	0.82	0.91
ANNPCR	PNC	0.89	−255.03	−0.01	8468.64	0.95	0.77	0.88
ANNELASTICNET	PNC	0.87	547.27	0.02	8081.09	0.96	0.81	0.91

The samples of the resulting regression models with the selected input variables for PM_{10}, $PM_{2.5}$ and PNC predictions are shown in Eqs. 14 to 17, respectively.

$$\begin{aligned}PM_{10} = {} & 0.027(nox) + 0.028(no2) + 0.053(no) \\ & + 0.157(so2) - 0.825(co) - 0.013(noxbg) - 0.054(nobg) \\ & - 0.019(so2bg) - 0.324(cobg) \\ & + 0.838(PMbg) - 1.701(rain) + 0.095(temp) \\ & + 0.023(Rhum) + 0.101(ws) + 0.075(T.\text{speed}) \\ & + 0.030(\text{LDV emission})\end{aligned} \tag{16}$$

$$\begin{aligned}PM_{2.5} = {} & 0.020(nox) + 0.026(no) - 1.156(co) \\ & - 0.011(no2bg) - 0.044(so2bg) - 2.817(cobg) \\ & + 0.972(PMbg) + 0.0270(rain) \\ & + 0.155(temp) - 0.0228(Bp) + 0.033(Rhum) \\ & + 0.019(ws) + 0.018(\text{LDV emission})\end{aligned} \tag{17}$$

$$\begin{aligned}PNC = {} & -294(no2) + 179 - 86(so2) + 5780(co) \\ & + 48(no2bg) - 141(nobg) + 364(so2bg) - 898(cobg) \\ & + 10(solRad) + 4616(rain) + 1128(temp) - 207(Bp) \\ & + 112(Rhum) - 782(ws) - 68(\text{LDV}) - 76(\text{HDV}) \\ & + 74(T.\text{speed}) + 834(\text{LDV emission})\end{aligned} \tag{18}$$

The most selected input variables were NOx, CObg, PMbg and the temperature, while the second most selected variables that were chosen in all cases except one were CO, NO, NO_2, SO_2, NObg, Noxbg, wind speed, rainfall, traffic speed and LDV emission. No2bg, HDV emission and barometric pressure were selected only in two cases while date and wind direction were only selected once. This selection shows that the roadside and background pollutants are the most significant predictors of the roadside particles, and temperature and wind speed were the most important meteorological variables identified by the methods. LDV and HDV emission rates are the traffic variables that were selected by at least two methods while traffic flow, HDV and LDV traffic were not selected in any case. The exclusion of the traffic flow and its composition might be because of their high correlation with their corresponding emission rates.

The hybrid ANN models were trained using 10-fold cross-validation repeated five times to determine appropriate model parameters including a number of hidden neurons (NHN), weight decay and the number of principal components for the PCA method. The training performance of the models and their corresponding parameters are shown in Table 1.

Considering the RMSE and *R*-squared values for PM_{10} prediction models, the ANN models with feature selection have shown similar training performances, but a slightly better training performance than the standalone ANN. The RMSE values for the hybrid ANN models are approximately 1.0 μg/m^3 less than the RMSE value of the standalone ANN. Also, the *R*-squared values of the hybrid ANN models are higher by approximately 0.6. The $PM_{2.5}$ prediction models show similar training performance while the standalone ANN for PNC performed better in training than the ANN models with feature selection. After the training, the models were tested with the holdout data set. Testing the models using the holdout data, provides confidence in their generalisation ability. The test results of the models are shown in Table 2.

The performance of the hybrid ANN models compares favourably with the standalone ANN models despite the reduction in the dimension of the variables in the case of the PCA method or reduction of the number of variables in the case of the Lasso and elastic-net methods. ANNLASSO/ELASTICNET models for the prediction of $PM_{2.5}$ and PNC produced slightly better predictions than the ANNPCA considering all the performance metrics while they show similar performance in the case of PM_{10} prediction. The ANNPCA models had similar performance to standalone ANN in the case of PM_{10} and $PM_{2.5}$ models but performed poorly in the case of PNC models (Fig. 5).

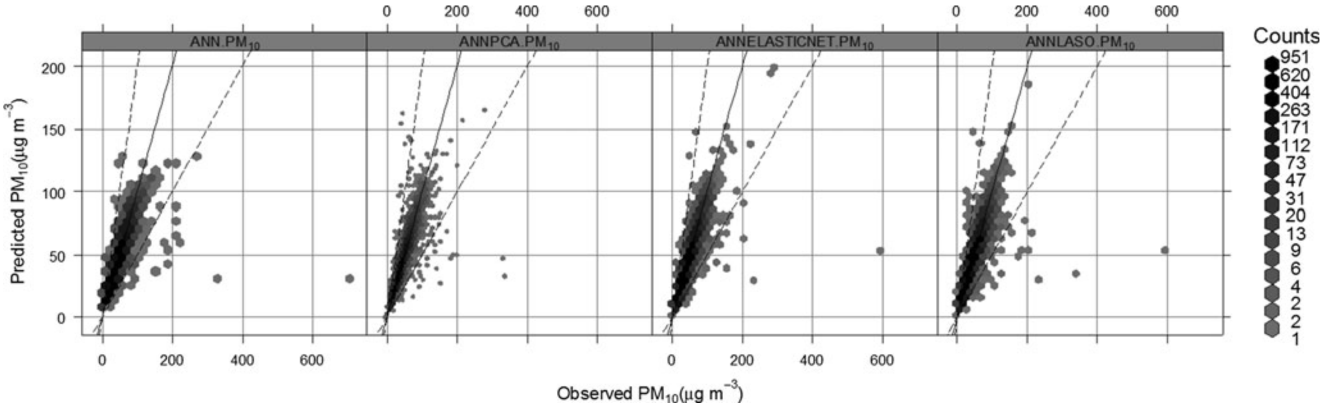

Fig. 5 Comparison between standalone ANN and hybrid ANN PM_{10} prediction models. The *outer dash lines* formed the boundaries of the FAC2 region, and the *key* to the right of the figure shows the frequency of the colour coded points

The ANNELASTICNET models for PM_{10} and $PM_{2.5}$ predictions have more of their predictions within a factor of two of the observed concentrations than the remaining models, and they captured the higher values of the concentrations more accurately (Fig. 5). In the case of PNC predictions, hybrid ANN models have shown similar scatter with the standalone ANN, but with the greater tendency of predicting the higher concentrations accurately. The ANNELASTICNET could be taken as the best performing ANN model because of its outstanding performance in all the cases considered.

3.2 BRT Modelling Results

The BRT models development began with the determination of the three model parameters (i.e. learning rate, the number of trees and tree complexity) using 10-fold cross-validation and the result is shown in Table 3. These parameters were then used to train the BRT models for predicting the concentrations of PM_{10}, $PM_{2.5}$ and PNC.

Table 3 shows the BRT model parameters and their training performances. The BRT model for PM_{10} prediction has shown better training performance than its corresponding ANN models as indicated by the RMSE and R-squared values of 8.9 μg/m^3 and 0.83 respectively. However, for the $PM_{2.5}$ and PNC predictions, the BRT models have shown slightly poorer performance than their corresponding ANN models (Table 4).

Table 3 Training parameters of the BRT models

Model parameter	PM_{10} (μg/m^3)	$PM_{2.5}$ (μg/m^3)	PNC (μg/m^3)
Number of trees	1000	1000	1000
Tree complexity	9	9	9
Learning rate	0.01	0.05	0.05
RMSE	8.90	4.78	9424
R-squared	0.83	0.88	0.88

The advantage of the BRT method over the standalone ANN method is that it can perform feature selection during the training and rank the input variables according to their contribution to the development of the BRT model. Here the BRT training was carried out in three stages: First, the whole set of variables were used for the training of the BRT models and, second, the six least contributing variables were tested for dropping in each case where four variables were finally dropped without compromising the predictive performance of the models. In the third step, the remaining variables were used to train the final models. Figure 6 shows the relative influence of the variables used in developing the final BRT models. The most important variable for PM_{10} prediction was roadside NOx and closely followed by background particle concentrations. However, in the prediction of $PM_{2.5}$, the background concentration of $PM_{2.5}$ was the most contributing variable followed by roadside NO.

For the PNC prediction, the date was the highest contributing variable, and this might be as a result of the sensitivity of PNC to temporal variation. In all the three cases, the roadside pollutant which contributed most was oxides of Nitrogen while the background pollutant contributing most was background concentrations of particles, and together they are more important than the traffic and meteorological variables. This behaviour is much expected since traffic is the major urban source of oxides of nitrogen and it also contributes to the formation of secondary particles. The contribution of traffic variables was less compared to the roadside oxides of nitrogen and the background particle concentrations, and this might occur due to the high correlation between these pollutants and the traffic variables. Traffic flow parameters are more important than the traffic speed in predicting PM_{10} concentration. Also, they have a comparable contribution to predicting $PM_{2.5}$ concentrations, and the least contributing variable in predicting PNC was the HDV traffic. Temperature and barometric pressure are the most contributing meteorological parameters for $PM_{2.5}$ and PNC predictions while Barometric

Table 4 Test performance of the BRT models

Pollutants	Models	FAC2	MB	MGE	NMB	NMGE	RMSE	*R*	CoE
PNC (number/cm^3)	BRT	0.80	−57.87	5369.95	0.00	0.23	8292.93	0.95	0.71
$PM_{2.5}$ (μg/m^3)	BRT	1.00	0.03	1.87	0.00	0.09	2.81	0.96	0.76
PM_{10} (μg/m^3)	BRT	0.99	−0.05	4.70	0.00	0.11	11.45	0.87	0.72

pressure and solar radiation are more important for PM_{10} predictions. The BRT algorithm is less sensitive to the correlated predictor variables than the Lasso and elastic-net algorithms, it assigns nearly equal contributions to the most correlated variables. However, Lasso and elastic-net methods can either drop the highly correlated variables or forced them to have zero contribution.

One important feature of the BRT method is the *partial dependence plot* that displays the effect of an input variable on the target variable while taking into account the average impact of all other variables in the BRT model [38, 46]. Although the integrity of the plots is affected by highly correlated variables, they provide a useful basis for interpreting the models [36, 47]. Figures 7, 8 and 9 show the partial dependence plots for the input variables used in the development of the BRT models for the prediction of roadside particles.

The partial dependence plots revealed that the roadside particle concentrations increase with the corresponding increase in roadside NOx concentrations. The approximate linear line graphs shown in Fig. 7 described this relationship. The pattern of the relationship is somewhat different with PNC concentration where the slope of the line in the plot flattened when the NOx concentration was around 300 μg/m3 and becomes steep again at around 1000 μg/m^3 where the PNC concentration increases without a corresponding increase in the NOx concentrations. The NO_2 concentrations show a parabolic relationship with the PM_{10} and $PM_{2.5}$ concentrations, and they increased with the commensurate increase in NO_2 concentrations up to around 200 μg/m^3 of the NO_2 and then decreased with further increase in the NO_2 concentrations. However, the NO_2 concentration shows a negative linear relationship with the PNC and all the particle concentrations show positive linear relationships with the NO and background particle concentrations. Moreover, the PM_{10} and PNC concentrations decreased with corresponding increases in the background concentrations of NOx, NO_2 and NO, while the $PM_{2.5}$ concentrations increase with a corresponding increase in their concentrations. The roadside SO_2 concentration shows a linear relationship with the PM_{10} and $PM_{2.5}$ when its concentration was between 0 and 20 μg/m^3 and then the relationship remained constant over the remaining range of the concentrations.

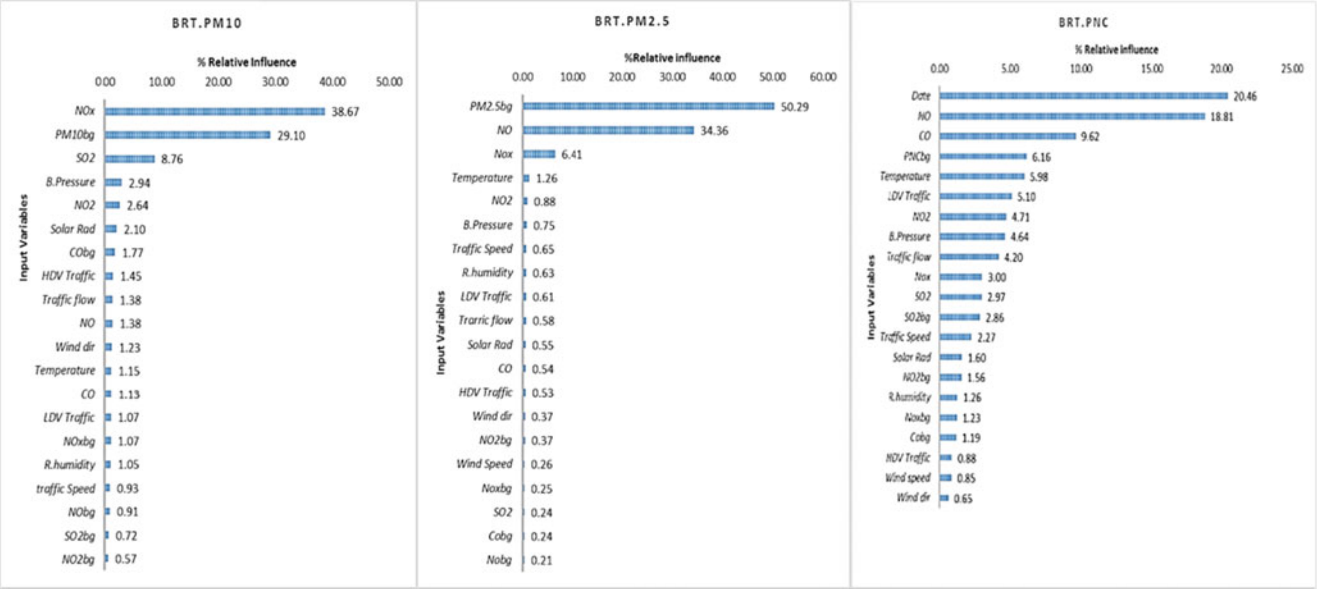

Fig. 6 Relative influence of the input variables in BRT models

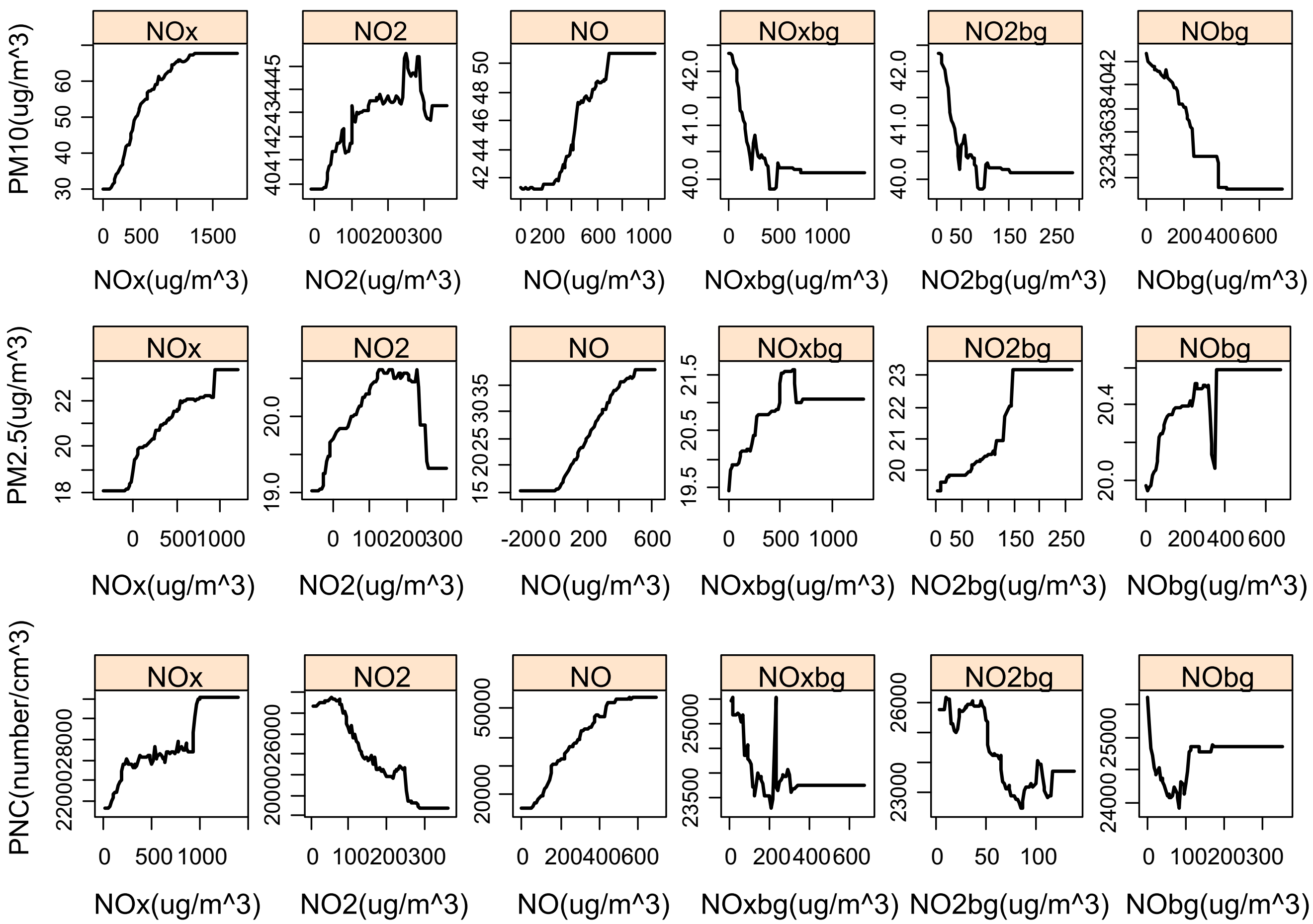

Fig. 7 Partial dependence plots showing the effects of pollutants and wind variables on the BRT model predictions of the roadside particle concentrations

However, the PNC concentrations decrease with corresponding increases in SO_2 concentrations up to 20 μg/m^3 of SO_2 and then the relationship changes to positive linear up to around 35 μg/m^3 and then remained constant for the rest of the values. The BRT model shows that the PM_{10} and $PM_{2.5}$ concentrations have negative linear relationships with the CO concentrations while having a positive linear relationship with the PNC. The positive relationships between the particles and most of the gaseous pollutants show that the gaseous pollutants play a vital role in the formation of the particles, or they share common sources. This information could give a clue to the intricate relationship between gaseous and the particle pollutants and will help in taking an urgent decision before conducting a detailed laboratory analysis of the relationships.

The BRT models also show that the higher particle concentrations are more associated with the winds coming from the south, southwest and southeast. These are the directions of the dominant winds at the site where the data was collected. Also, these directions coincide with the side of the road where the monitoring unit is located which suggests that Canyon recirculation vortices delivered most of the particle concentrations to the monitoring unit. This information is also useful as it provides a clue to whether the monitoring unit was in a right position or not. The relationship between the wind speeds and the concentrations of $PM_{2.5}$ and PNC was shown to be negative linear. This relationship is expected because when the wind speed is high, the ventilation in the street increases and then most of the particle concentrations are removed from the street. However, the PM_{10} concentrations show the opposite where they increased with the corresponding increase in wind speed. The possible explanation for this relationship is that the higher winds might carry dust and other larger size particles especially non-exhaust particles which could have raised the concentrations of the PM_{10}.

The temperature and relative humidity show a nearly positive linear relationship with the PM_{10} and $PM_{2.5}$ while the PNC concentrations show nearly linear relationships with temperature and an almost constant relationship with relative humidity. The positive association between the particle

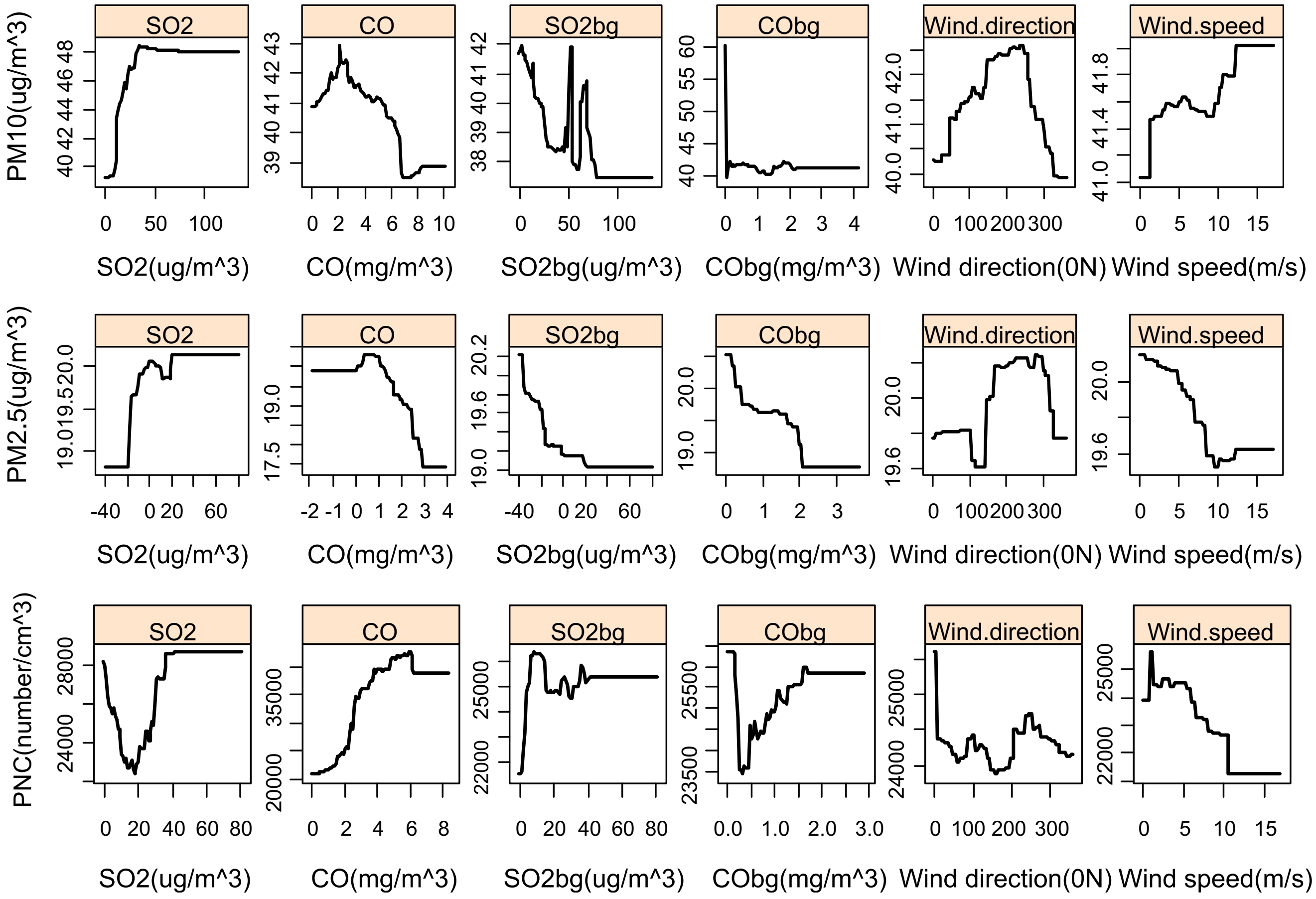

Fig. 7 (continued)

concentrations and the temperature did not agree with the findings of the previous studies by Dos Santos-Juusela et al. [48]. However, Barmpadimos et al. [49] reported positive relationships between the temperature and PM_{10} concentrations in the summer and Tai et al. [50] found a positive correlation between most of the components of $PM_{2.5}$ except for nitrate which shows a negative association. A linear correlation between the temperature and the particle concentrations was also estimated, and the coefficients of correlations between them were found to be 0.15, 0.14 and 0.26 for PM_{10}, $PM_{2.5}$ and PNC, respectively. Moreover, the elastic-net models also show a positive relationship with the temperature. This relationship needs to be further investigated especially to find out the seasonal relationship between the temperature and the particles and the levels at which the relationship changes.

The traffic flow and the HDV traffic show a negative linear relationship with the concentrations of $PM_{2.5}$ which is not in agreement with the fact that the concentration increases with the corresponding increase in traffic flow. However, it might explain the stop and go situation at the site, where the emission is high when the vehicles are not moving and during acceleration and then reduces as the flow becomes normal. However, in the case of PM_{10} and PNC, the concentrations remain relatively constant when the traffic flow was between 2000 and 4000 veh/h and then suddenly increased to higher concentrations and then remained constant as shown in Fig. 9. This behaviour could explain the situation when the road reaches its capacity where the concentration increases as a result of high numbers of vehicles. The HDV and LDV traffic captured the hourly variation of the concentrations of PM_{10} and PNC. HDV traffic shows strong association with the average PNC concentration, and it has a bimodal distribution which suggests that it keeps track of the temporal variation of the PNC in the model.

The positive linear relationship shown by most of the variables is an indication of the sign of their contribution in determining the appropriate prediction. For example, the positive correlation might be excitatory while the negative correlation might be inhibitive in deciding the final predicted value; therefore, both the input variables with the positive and negative relationships are vital in determining the final prediction of the model. The analysis of the partial dependence plots could help the model user to have a fair understanding of the type of the relationship between the predictor variables and the particle concentrations. The information gained could inform several management decisions related to the control of

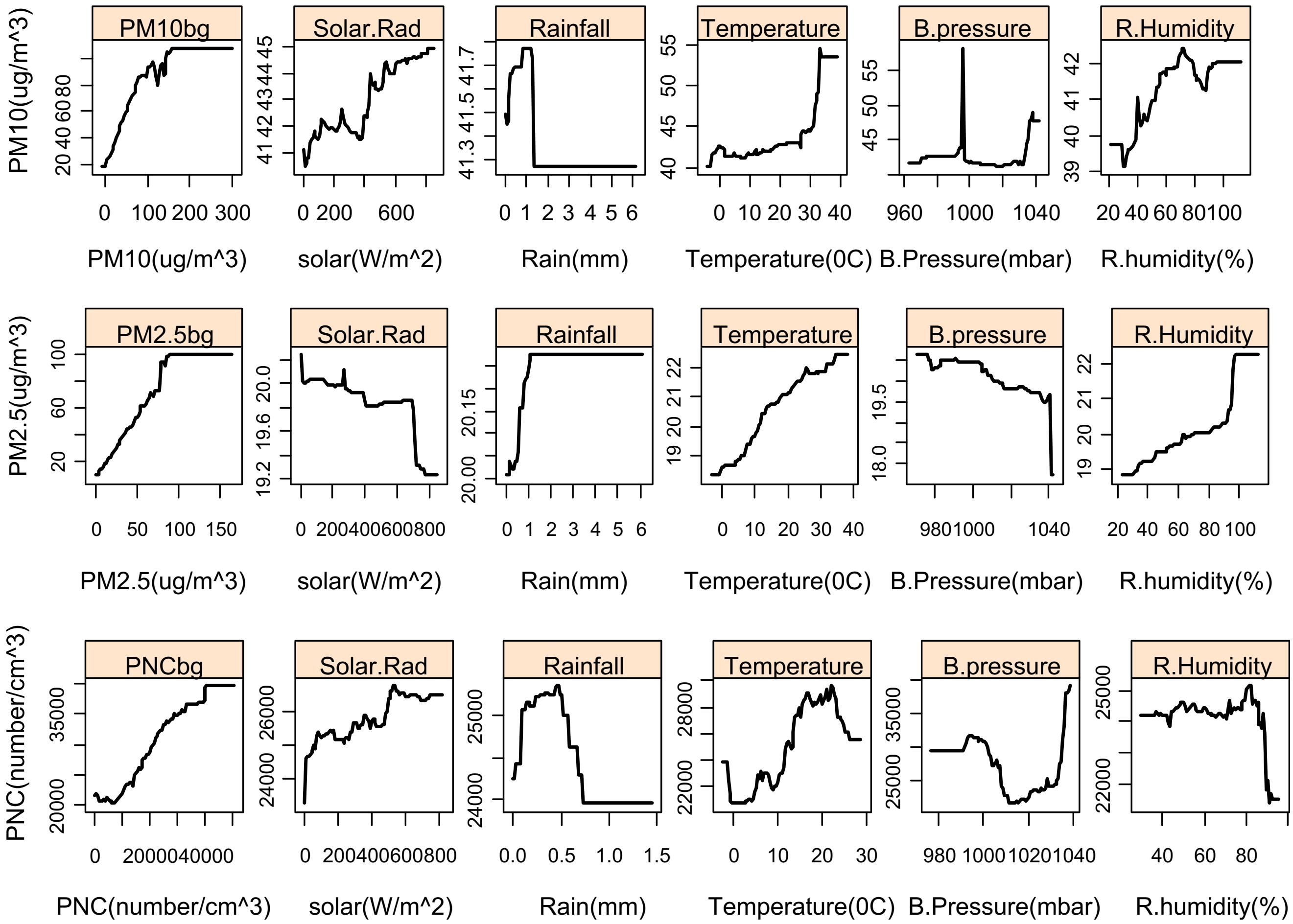

Fig. 8 Partial dependence plots showing the effects of background particle concentrations and meteorological variables on the BRT model predictions of the roadside particle concentrations

air quality. For example, any control measure taken to reduce the roadside oxides of nitrogen will have a significant impact on the particle concentrations due to their strong relationship explained by the BRT models. Moreover accurate determination of the levels of oxides of nitrogen could yield better BRT models for the prediction of the roadside particles.

3.3 Comparison of the performance of ANN and BRT prediction models

In this section, the performance of the ANN and BRT models are compared to allow for drawing conclusions on whether one method could be recommended over the other. Statistical performance metrics (FAC2, MB, MGE, NMB, NMG, RMSE, R and CoE), conditional quantile plots and polar annulus plots were used for the comparison.

The ANN and BRT models show similar performance as indicated by their performance statistics in Table 5. The main difference observed was in the PNC prediction models where the ANN model performed better than the BRT model. The models have their normalised mean bias ranging between 0.00 and 0.02, which is an indicator of whether the models are over-predicting or under predicting the particle concentrations. The prediction errors of the models measured by the RMSE values were found to be 10 and 11.5 µg/m^3, 3.03 and 2.81 µg/m^3 for ANN and BRT models for the prediction of PM_{10} and $PM_{2.5}$ respectively. About 99 to 100 % of the PM_{10} and $PM_{2.5}$ model predictions were within a factor of two of their respective observed concentrations while the ANN and BRT models for PNC prediction have 87 and 80 % of their predictions within a factor of two of the observed PNC concentrations respectively. The Coefficient of Efficiency (CoE) that indicates the accuracy of the model prediction is between 0.7 and 0.81. CoE is the measure of the model efficiency that is robust and easy to interpret [45], and it has an interpretation for zero and negative values. A perfect model has CoE value of one, and zero values of CoE show that the model's prediction accuracy is not more than the mean values of the observed concentrations. Negative CoE values show that the model's

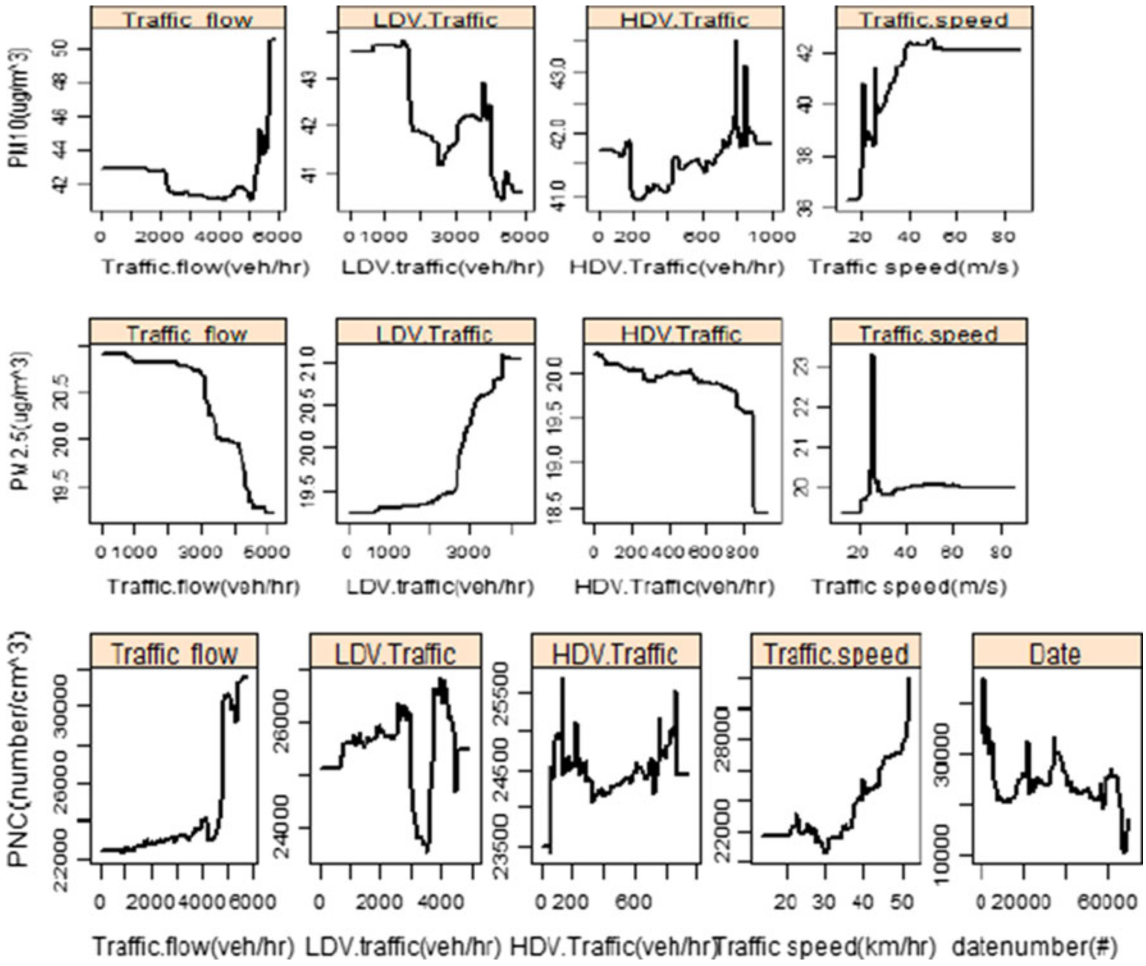

Fig. 9 Partial dependence plots, showing the effects of traffic variables on the BRT model predictions of roadside particle concentrations

prediction accuracy is worse than the observed mean. Here, the ANN and BRT models performed well above their respective observed mean values. The conditional quantile plots shown in Fig. 10 compare the performance of the standalone ANN, ANNELASTICNET and BRT models for the prediction of PM_{10} respectively. The standalone ANN could only predict PM_{10} concentrations up to 100 μg/m^3 accurately and deviates significantly from the perfect model line (smooth line) towards the region of higher concentrations. However, its predictions improved when the elastic-net regression was used for selecting the most relevant variables for the modelling as shown in the middle panel. The ANNELASTICNET is better regarding the agreement between the observed and predicted values. Its predictions (bumpy line) matched the perfect model line more accurately than the remaining models. However, the BRT model shows better data coverage though not so accurate in the prediction of

Table 5 ANN and BRT model performance statistics

Pollutants	Models	FAC2	MB	MGE	NMB	NMGE	RMSE	*R*	CoE
PM_{10} (μg/m^3)	BRT	0.99	−0.05	4.70	0.00	0.11	11.45	0.87	0.72
	ANN ELASTICNET	0.99	0.11	4.98	0.00	0.12	10.01	0.89	0.70
$PM_{2.5}$ (μg/m^3)	BRT	1.00	0.03	1.87	0.00	0.09	2.81	0.96	0.76
	ANN ELASTICNET	1.00	0.02	2.08	0.00	0.10	3.03	0.96	0.74
PNC (number/ cm^3)	BRT	0.80	−57.87	5369.95	0.00	0.23	8292.93	0.95	0.71
	ANN ELASTICNET	0.87	547	4376	0.02	0.18	8081.09	0.96	0.81

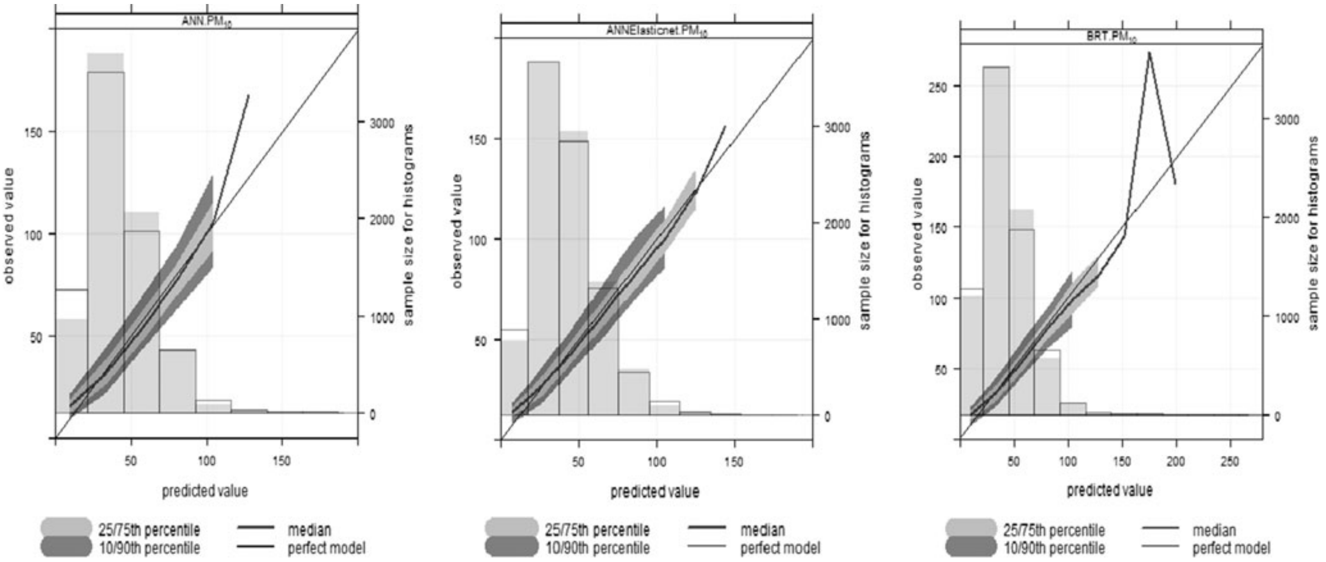

Fig. 10 Comparison between ANNELASTICNET, ANN and BRT PM_{10} model predictions

the higher concentrations. The same conclusion could be made for the $PM_{2.5}$ and PNC models as shown in Figs. 11 and 12 respectively.

The polar annulus plots shown in Figs. 13 and 14 were drawn to explore how accurate the ANNELASTICNET and BRT models for PM_{10} predictions captures the trend, seasonal and temporal variation that exists in the observation data. Considering the upper left panel of Fig. 13 (Trend), the ANNELASTICNET model accurately captures the reduction in the level of PM_{10} concentrations between 2003 and 2007 and the high concentrations associated with northerly winds in 2006. Also, the high concentrations associated with the southerly winds in summer and winter are captured by the model prediction though with slight under-prediction. The PM_{10} concentrations were higher when the wind was coming from the south-east, south and southwest, this has also been adequately taken care of by the model predictions. The daily and hourly variation in the test data has been altered, but the model adequately reflects the alteration in its prediction. For example, the analysis of the data not shown here indicates that the particle concentrations are higher on weekdays and in the daytime while lower at weekends and night times. These properties have not been captured in the test data, and the model also did not show them despite the fact that they existed in the training data. For the BRT method, the models captured accurately the higher concentrations associated with the north and south

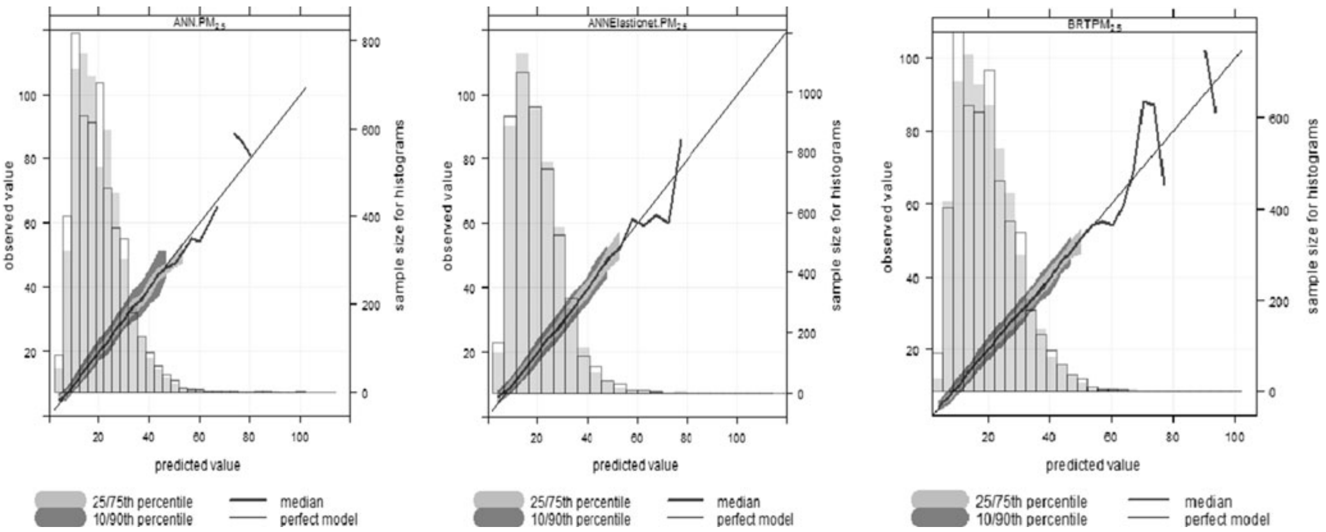

Fig. 11 Comparison between ANNELASTICNET, ANN and BRT $PM_{2.5}$ model predictions

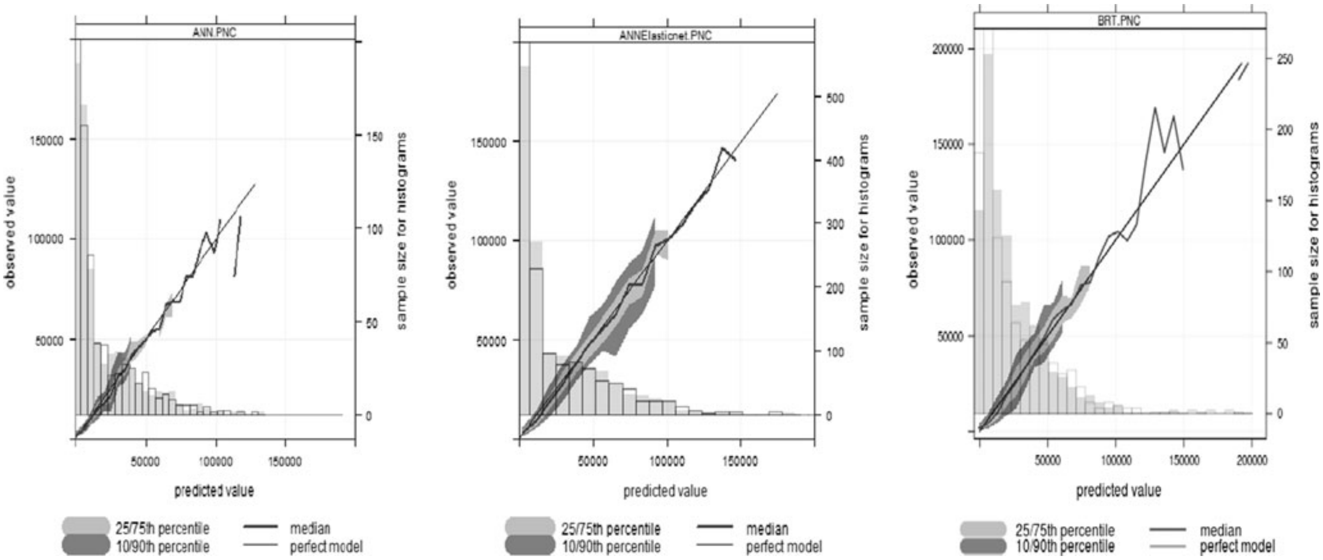

Fig. 12 Comparison between ANNELASTICNET, ANN and BRT PNC model predictions. Figures. 10, 11 and 12 show conditional quantile plots indicating the agreement between the models prediction and the observed PMs concentrations. The model prediction and observation values were divided into bin pairs of equal length, and the median, 25/75th and 10/90th percentiles of each bin, was estimated and then plotted. The *smooth line* represents an ideal model, and the *bumpy line* represents the median of the predicted values, the *shading* shows the quantile intervals of the predictions and the *histograms* display the counts of the predicted values

winds in 2003 and the missing data for the rest of the year (see Fig. 14). Moreover, they show the seasonal variations indicated by the observed data especially the higher concentrations associated with easterly and southerly winds in the winter. The daily and hourly variations were seriously altered by the random nature of the data division but still the BRT models capture the properties of the test data. The most valuable property that was captured accurately by the model was the association of the higher concentrations with the winds coming from south and southeast which indicates the effect of Canyon recirculation since the monitoring station was located in a street

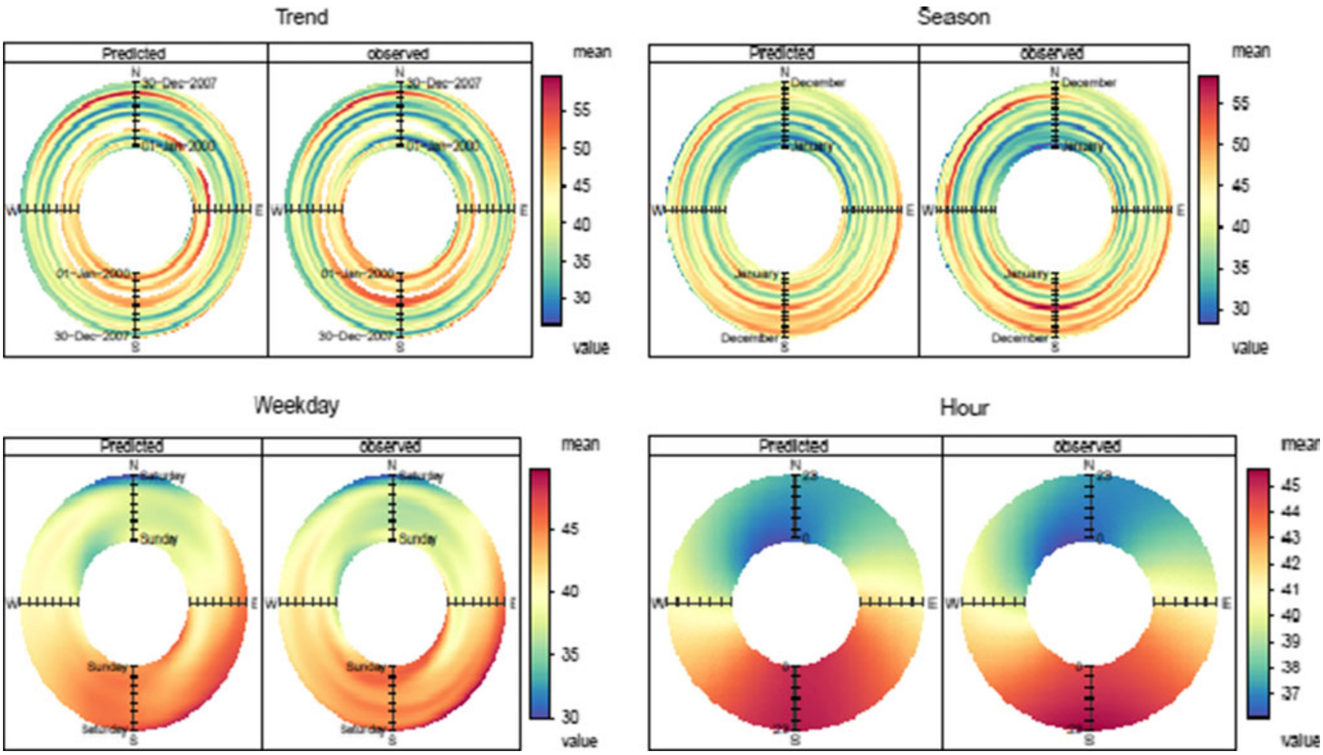

Fig. 13 Comparison between the observed and ANNELASTICNET-predicted PM_{10} concentrations

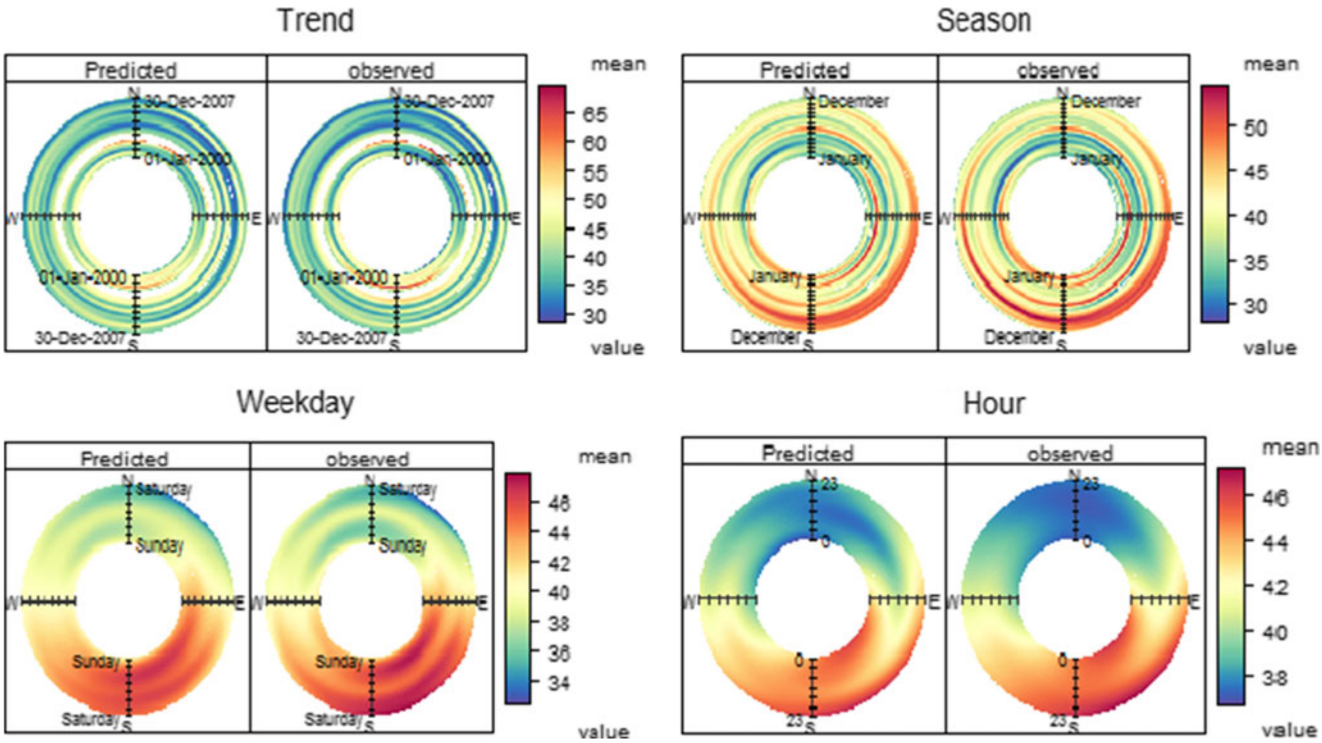

Fig. 14 Comparison between the observed and BRT-predicted PM_{10} concentrations. Figures 13 and 14 show the polar annulus plots comparing the agreement in trend and seasonal, weekday and hourly variation between the model predictions and the observed data. The *graduated marks* on north, south, east and west indicates years, months, days and hours in trend, season, weekday and hour plots, respectively. The *plots* also show the variation of the particle concentrations with wind directions. The *colour scale* shows the PM_{10} concentration levels from low to high

canyon aligned on an axis of 75°–255° to the southern side of the road.

4 Conclusions

This paper used air pollution, meteorological and traffic data collected at Marylebone Road and Bloomsbury sites in London to study the use of ANN and BRT methods for modelling roadside particulate matter. The effects of feature selection on the prediction accuracy of the ANN models were also investigated. Elastic-net regression selected predictor variables that yielded the best roadside particle prediction models among the three methods investigated. The prediction accuracy of the ANN and BRT models was compared using FAC2, MB, MGE, NMB, NMGE, RMSE, R, and CoE values. About 87–99% percent of the model predictions are within a factor of two of the observed data which shows good agreement between the model predictions and the particle observations. The CoE values of the models were found to be between 0.70 and 0.81 showing that the models can predict particle concentrations much more accurately than the mean of the observed concentrations. The ANN models were found to be only slightly more accurate than the BRT models. For example, the prediction errors of PM_{10} and $PM_{2.5}$ models measured by the RMSE values differed by only 1 $\mu g/m^3$ and the RMSE values of the PNC models differed with only about 200 number/cm^3. Also, both the ANN and BRT models show nearly zero bias in their predictions as indicated by NMB values. They also show similar R values ranging between 0.86 and 0.96 showing high correlation with the particle observations. However, despite their similarities in performance statistics, BRT models can perform feature selection and give valuable information about the relationships between the input variables and the target variables. The analysis of the BRT relative influence and partial dependence plots revealed that the background particle concentrations and the oxides of nitrogen are the most relevant predictor variables and have strong positive relations with the particle concentrations. They also show that Temperature is more important in predicting $PM_{2.5}$ and PNC than PM_{10}. It was also discovered that the traffic variables keep track of the particles temporal variations in the models. The study concluded that both ANN and BRT methods can be used to develop air quality models for the prediction of roadside particles with good model—observation agreement with preference to BRT models when model interpretation is more important than the prediction accuracy.

Acknowledgements The authors would like to thank Petroleum Technology Development Fund (PTDF) for funding this research and the London Air Quality Network and Environmental research group at Kings College London for their help in providing data and Openair package for the data analysis.

References

1. Brauer, M., Hoek, G., Van Vliet, P., Meliefste, K., Fischer, P. H., Wijga, A., et al. (2002). Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children. *Am J Respir Crit Care Med, 166*(8), 1092–1098. **Research Support, Non-U S Gov't**.
2. Kim, J. J., Smorodinsky, S., Lipsett, M., Singer, B. C., Hodgson, A. T., & Ostro, B. (2004). Traffic-related air pollution near busy roads: the East Bay Children's Respiratory Health Study. *Am J Respir Crit Care Med, 170*(5), 520–526.
3. McConnell, R., Berhane, K., Yao, L., Jerrett, M., Lurmann, F., Gilliland, F., et al. (2006). Traffic, susceptibility, and childhood asthma. *Environmental Health Perspectives, 114*(5), 766–772.
4. Lindgren, A., Stroh, E., Nihlen, U., Montnemery, P., Axmon, A., & Jakobsson, K. (2009). Traffic exposure associated with allergic asthma and allergic rhinitis in adults. A cross-sectional study in southern Sweden. International Journal of Health Geographics, 8, doi:10.1186/1476-072x-8-25.
5. Heinrich, J., Topp, R., Gehring, U., & Thefeld, W. (2005). Traffic at residential address, respiratory health, and atopy in adults: the National German Health Survey 1998. *Environmental Research, 98*(2), 240–249. doi:10.1016/j.envres.2004.08.004.
6. Brunekreef, B., Beelen, R., Hoek, G., Schouten, L., Bausch-Goldbohm, S., Fischer, P., et al. (2009). Effects of long-term exposure to traffic-related air pollution on respiratory and cardiovascular mortality in the Netherlands: the NLCS-AIR study. *Research report (Health Effects Institute), 139*, 5–71. **discussion 73–89**.
7. USEPA (2011). The Benefits and Costs of the Clean Air Act: 1990–2020. In U.S. Environmental Protection Agency Office of Air and Radiation (Ed.), Final Report.
8. Yim, S. H., & Barrett, S. R. (2012). Public health impacts of combustion emissions in the United Kingdom. *Environ Sci Technol, 46*(8), 4291–4296. doi:10.1021/es2040416.
9. COMEAP (2010). *Committee on the Medical Effects of Air Pollutants (COMEAP):the mortality effects of long-term exposure to particulate air pollution in the United Kingdom*. https://www.gov.uk/government/publications/comeap-mortality-effects-of-long-term-exposure-to-particulate-air-pollution-in-the-uk. Accessed 1 May 2014.
10. Chave, J., & Levin, S. (2003). Scale and scaling in ecological and economic systems. *Environmental and Resource Economics, 26*(4), 527–557.
11. National Research Council Committee on Models in the Regulatory Decision Process. (2007). *Models in environmental regulatory decision making*. Washhington, DC, USA: National Academies Press.
12. Lagzi, I., Mészáros, R., Gelybó, G., & Leelőssy, Á. (2013). *Atmospheric chemistry*. Hungary: Eötvös Loránd University.
13. Gardner, M. W., & Dorling, S. R. (2000). Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment, 34*(1), 21–34. doi:10.1016/S1352-2310(99)00359-3.
14. Esplin, G. J. (1995). Approximate explicit solution to the general line source problem. *Atmospheric Environment, 29*(12), 1459–1463. doi:10.1016/1352-2310(94)00348-O.
15. Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling, 154*(1), 135–150.
16. Yan Chan, K., & Jian, L. (2013). Identification of significant factors for air pollution levels using a neural network based knowledge discovery system. *Neurocomputing, 99*, 564–569. doi:10.1016/j.neucom.2012.06.003.
17. Carslaw, D. C., & Ropkins, K. (2012). openair—an R package for air quality data analysis. *Environmental Modelling & Software, 27–28*, 52–61. doi:10.1016/j.envsoft.2011.09.008.
18. Jones, A. M., & Harrison, R. M. (2005). Interpretation of particulate elemental and organic carbon concentrations at rural, urban and kerbside sites. *Atmospheric Environment, 39*(37), 7114–7126. doi:10.1016/j.atmosenv.2005.08.017.
19. Jones, A., & Harrison, R. (2006). Estimation of the emission factors of particle number and mass fractions from traffic at a site where mean vehicle speeds vary over short distances. *Atmospheric Environment, 40*(37), 7125–7137. doi:10.1016/j.atmosenv.2006.06.030.
20. Aurelie, C., Harrison, R. M. (2005). Comparison between SMPS, nano-SMPS and epiphaniometer data at an urban background site (Bloomsbury) and a roadside site (Marylebone road). http://uk-air.defra.gov.uk/reports.
21. LondonAir (2013). London Air quality Network. http://www.londonair.org.uk/london/asp/datadownload.asp. Accessed 03/04/2013 2013.
22. UK-AIR (2013). Department fro Environment Food and Rural Affairs Data Archive. http://uk-air.defra.gov.uk/data/maryleboneroad. Accessed 03/04 2013.
23. Bowden, G. J., Maier, H. R., & Dandy, G. C. (2005). Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. *Journal of Hydrology, 301*(1–4), 93–107. doi:10.1016/j.jhydrol.2004.06.020.
24. Hastie, T., Tibshirani, R., & Friedman, J. (2008). The elements of statistical learning data Mining, inference, and prediction. Springer. Second Edition.
25. Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An overview of principal component analysis. *Journal of Signal and Information Processing, 4*, 173.
26. Chunming, L., Yanhua, D., Hongtao, M., & YuShan, L. A Statistical PCA Method for face recognition. In Intelligent Information Technology Application, 2008. IITA '08. Second International Symposium on, 20–22 Dec. 2008 2008 (Vol. 3, pp. 376–380). doi:10.1109/IITA.2008.71.
27. Kuhn, M. (2012). *Caret: Classification and Regression Training*. R package version 5.15-044. http://CRAN.R-project.org/package=caret.
28. R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.
29. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R. (Vol. 103)*. New York Heidelberg Dordrecht London: Springer.
30. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1.
31. Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.
32. Mlakar P., & Božnar, M. Z. (2011). Artificial neural networks - a useful tool in air pollution and meteorological modelling. In Nejadkoorki, F. (Ed.), *Advanced air pollution*. InTech. Available from: http://www.intechopen.com/books/advanced-air-pollution/

artificial-neural-networks-auseful-tool-in-air-pollution-and-meteorological-modelling.

33. Haykin, S. (2005). *Neural networks - a comprehensive foundation* (2nd ed.). Delhi: Pearson Printice Hall Publication.
34. Nagendra, S. M. S., & Khare, M. (2006). Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. *Ecological Modelling, 190*(1–2), 99–115. doi: 10.1016/j.ecolmodel.2005.01.062.
35. Ripley, B. (2013). Feed-forward neural networks and multinomial log-linear models. nnet package, version 7.3-6. *URL http://www.stats.ox.ac.uk/pub/MASS4*.
36. Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802–813.
37. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378. doi: 10.1016/S0167-9473(01)00065-2.
38. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. doi: 10.2307/2699986.
39. Ridgeway, G., Southworth, M. H., & RUnit, S. (2013). Package 'gbm'. *Viitattu, 10*, 2013.
40. Ridgeway, G. (2007). Generalized boosted models: a guide to the gbm package. *Update, 1*(1), 2007.
41. Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
42. Willmott, C. J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society, 63*(11), 1309–1313.
43. Willmott, C. J. (1981). On the validation of models. *Physical geography, 2*(2), 184–194.
44. Chang, J. C., & Hanna, S. R. (2004). Air quality model performance evaluation. *Meteorology and Atmospheric Physics, 87*(1–3), 167–196. doi:10.1007/s00703-003-0070-7.
45. Legates, D. R., & McCabe, G. J. (2013). A refined index of model performance: a rejoinder. *International Journal of Climatology, 33*(4), 1053–1056
46. Carslaw, D. C., & Taylor, P. J. (2009). Analysis of air pollution data at a mixed source location using boosted regression trees. *Atmospheric Environment, 43*(22–23), 3563–3570. doi:10.1016/j.atmosenv.2009.04.001.
47. Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Stat Med, 22*(9), 1365–1381. doi:10.1002/sim.1501.
48. Dos Santos-Juusela, V., Petäjä, T., Kousa, A., & Hämeri, K. (2013). Spatial-temporal variations of particle number concentrations between a busy street and the urban background. *Atmospheric Environment, 79*, 324–333.
49. Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., & Prévôt, A. (2011). Influence of meteorology on PM 10 trends and variability in Switzerland from 1991 to 2008. *Atmospheric Chemistry and Physics, 11*(4), 1813–1835.
50. Tai, A. P. K., Mickley, L. J., & Jacob, D. J. (2010). Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: implications for the sensitivity of PM2.5 to climate change. *Atmospheric Environment, 44*(32), 3976–3984. doi:10.1016/j.atmosenv.2010.06.060.