Atmospheric Environment 80 (2013) 426–437

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Identifying pollution sources and predicting urban air quality using ensemble learning methods

Kunwar P. Singh [a,b,*], Shikha Gupta [a,b], Premanjali Rai [a,b]

[a] Academy of Scientific and Innovative Research, Council of Scientific & Industrial Research, New Delhi, India
[b] Environmental Chemistry Division, CSIR – Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India

HIGHLIGHTS

- Developed tree ensemble models for seasonal discrimination and air quality prediction.
- PCA used to identify air pollution sources; air quality indices used for health risk.
- Bagging and boosting algorithms enhanced predictive ability of ensemble models.
- Ensemble classification and regression models performed better than SVMs.
- Proposed models can be used as tools for air quality prediction and management.

GRAPHICAL ABSTRACT

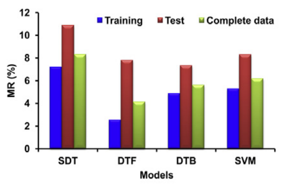
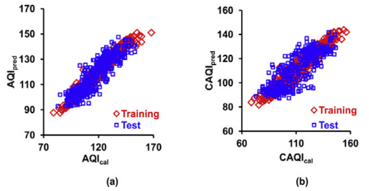


Figure shows misclassification rate in seasonal discrimination of air quality of Lucknow yielded by different models and suggest that the ensemble learning classification models (DTF and DTB) performed relatively better than SDT and SVM.



Figures show correlative distribution of calculated and model predicted values of (a) AQI, and (b) CAQI for Lucknow ambient air using DTB model.

ARTICLE INFO

Article history:
Received 23 June 2013
Received in revised form
5 August 2013
Accepted 12 August 2013

Keywords:
Ensemble learning
Decision tree forest
Decision treeboost
Air quality
Indices
Air pollution
Seasonal discrimination

ABSTRACT

In this study, principal components analysis (PCA) was performed to identify air pollution sources and tree based ensemble learning models were constructed to predict the urban air quality of Lucknow (India) using the air quality and meteorological databases pertaining to a period of five years. PCA identified vehicular emissions and fuel combustion as major air pollution sources. The air quality indices revealed the air quality unhealthy during the summer and winter. Ensemble models were constructed to discriminate between the seasonal air qualities, factors responsible for discrimination, and to predict the air quality indices. Accordingly, single decision tree (SDT), decision tree forest (DTF), and decision treeboost (DTB) were constructed and their generalization and predictive performance was evaluated in terms of several statistical parameters and compared with conventional machine learning benchmark, support vector machines (SVM). The DT and SVM models discriminated the seasonal air quality rendering misclassification rate (MR) of 8.32% (SDT); 4.12% (DTF); 5.62% (DTB), and 6.18% (SVM), respectively in complete data. The AQI and CAQI regression models yielded a correlation between measured and predicted values and root mean squared error of 0.901, 6.67 and 0.825, 9.45 (SDT); 0.951, 4.85 and 0.922, 6.56 (DTF); 0.959, 4.38 and 0.929, 6.30 (DTB); 0.890, 7.00 and 0.836, 9.16 (SVR) in complete data. The DTF and DTB models outperformed the SVM both in classification and regression which could be attributed to the incorporation of the bagging and boosting algorithms in these models. The proposed ensemble models successfully predicted the urban ambient air quality and can be used as effective tools for its management.

[*] Corresponding author. Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India. Tel.: +91 522 2476091; fax: +91 522 2628227.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).

1352-2310/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2013.08.023

1. Introduction

Air quality and its temporal and spatial variations in a region are largely determined by the nature of anthropogenic activities associated with various gaseous and particulate emissions and set of prevailing meteorological conditions there. Epidemiological studies have established the associations between the air pollutants and daily excess in mortality (Dockery and Pope, 1994) and morbidity (Kassomenos et al., 2008). Poor air quality has both acute as well as chronic health impacts (Nastos et al., 2010) and the severity of the impacts mostly depends upon two factors viz. ambient concentration of the air pollutants and its exposure time. Further, concentrations of air pollutants are subject to alter depending on the local topography, source emission, and surrounding meteorological conditions. However, among these variables, meteorological parameters are mostly responsible for causing variations in the ambient concentrations of air pollutants (Banerjee et al., 2011). Suspended particulate matter (SPM), respirable suspended particulate matter (RSPM), oxides of nitrogen (NO_x) and sulfur (SO_x), and ozone are among the major common air pollutants. RSPM refers to those SPM with nominal aerodynamic diameter of 10 μm or less. The environmental regulatory authorities (CPCB, 2009) have prescribed the guidelines for the maximum permissible levels of various air pollutants, such as SO_2 (80 $\mu g\ m^{-3}$), NO_2 (80 $\mu g\ m^{-3}$), and RSPM (100 $\mu g\ m^{-3}$), respectively.

Concern about air pollution in urban regions is receiving increasing importance worldwide (Chattopadhyay et al., 2010). The urban areas might be viewed as dense source of enormous anthropogenic emission of pollutants, which could alter the atmospheric composition, chemistry, and life cycles in its downwind regimes, extending over several hundred kilometers (Gupta et al., 2008). Petrol and diesel engines of motor vehicles were found to emit a wide variety of pollutants, principally oxides of nitrogen (NO_x) which had an increasing impact on urban air quality (Mage et al., 1996). Urban air pollution in India had increased rapidly with the population growth, number of motor vehicles, use of fuels with poor environmental performance, badly mentioned transportation system, poor land use pattern, and above all ineffective environmental regulations (Gupta et al., 2008). Developing appropriate strategies for air pollution prevention, understanding of the nature of sources and influences of meteorological conditions on their profiles are essentially required.

As for the health impact of air pollutants, air quality index (AQI) is an important indicator for general public to understand easily how bad or good the air quality is for their health and to assist in data interpretation for decision making processes related to pollution mitigation measures and environmental management. Basically, the AQI is defined as an index or rating scale for reporting the daily combined effect of ambient air pollutants recorded at the monitoring site (Kumar and Goyal, 2011). To ensure the safety of the humans and the archeological monuments, particularly in the urban areas, it is very much desired to monitor the ambient air quality on a regular basis and develop appropriate strategies for the control of the emission sources.

Accordingly, several urban air quality monitoring programs covering various major Indian cities have been initiated generating huge databases (Chattopadhyay et al., 2010) in recent past years. However, requirements of huge funds, dedicated manpower and instrumentation for such programs limit their viability. Therefore, to develop effective strategies for urban air quality management, it is essentially needed to develop appropriate methods which could be capable of predicting the air quality and enumerating the seasonal influences on ambient air in a region.

On the other hand, predictive modeling offers tools for forecasting the air quality based on past measurements. In recent years, several research efforts have been made in this direction. Atmospheric dispersion models used to predict the ground level concentration of the air pollutants around the sources (Cimorelli et al., 2005; EPA, 2005; Kesarkar et al., 2007; Bhaskar et al., 2008) require précised knowledge of several source parameters and the meteorological conditions (Collett and Oduyemi, 1997). The statistical models attempt to determine the underlying relationship between a set of input data and targets. Since, air quality data are generally very complex and exhibit nonlinear dependence, linear modeling approaches may not be suitable to model such data (Singh et al., 2012). The artificial neural networks (ANNs) are considered as standard nonlinear estimators and their predictive and generalization abilities have been well established through their successful applications in a variety of fields (Singh et al., 2012, 2013), but ANNs suffer with the problem of over-fitting in learning process.

Support vector machines (SVMs), exhibit excellent generalization abilities with non-linear systems (Singh et al., 2013), and also make use of limited data points in model building. In recent years ensemble learning methods (Snelder et al., 2009) have emerged as unbiased tools for modeling the complex relationships between set of independent and dependent variables and have been applied successfully in various research areas (Yang et al., 2010). In general, these methods are designed to overcome problems with weak predictors (Hancock et al., 2005) and have the advantage to alleviate the small sample size problem by averaging and incorporating over multiple classification models to reduce the potential for overfitting the training data (Dietterich, 2000). Decision trees (DTs) are commonly used as base predictors in building ensemble learning models (Zhang et al., 2008) and supplemented with bagging and stochastic gradient boosting techniques (Breiman, 1996; Friedman, 2002). The bagging aims minimizing of prediction variance by generating bootstrapped replica data sets, whereas, boosting creates a linear combination out of many models, where each new model is dependent on the preceding model (Friedman, 2002). Decision tree forest (DTF) and decision treeboost (DTB) implementing bagging and boosting techniques, respectively are relatively new methods for improving the accuracy of a predictive function (Yang et al., 2010). These techniques are inherently nonparametric statistical methods and make no assumption regarding the underlying distribution of the values of predictor variables and can handle numerical data that are highly skewed or multi-model in nature (Mahjoobi and Etemad-Shahidi, 2008). To our knowledge, ensemble learning methods have not yet been applied to the air quality prediction.

The main objectives of this study were (i) to construct ensemble learning (EL) based classification and regression functions to enumerate the influence of seasons on the air quality; and to predict the AQI in the study region using selected air quality (RSPM, NO_2, SO_2) and meteorological parameters (air temperature, relative humidity, wind speed, sunshine hours, evaporation rate) as the estimators, (ii) to compare the predictive and generalization abilities of these modeling approaches. Accordingly, EL models were developed and the performances of these models were evaluated in terms of several statistical criteria parameters and compared with the SVM approach as benchmark. This study has shown that the application of EL methods can be useful in predicting the air quality and enumerating the seasonal influences successfully for its effective management.

2. Methods

The basic aim of this study is (a) to identify the air pollution sources, (b) to find an accurate possible classification function $\overline{f}_c$ capable of discriminating the seasonal (summer, monsoon, winter) air quality to enumerate the responsible factors, influences of

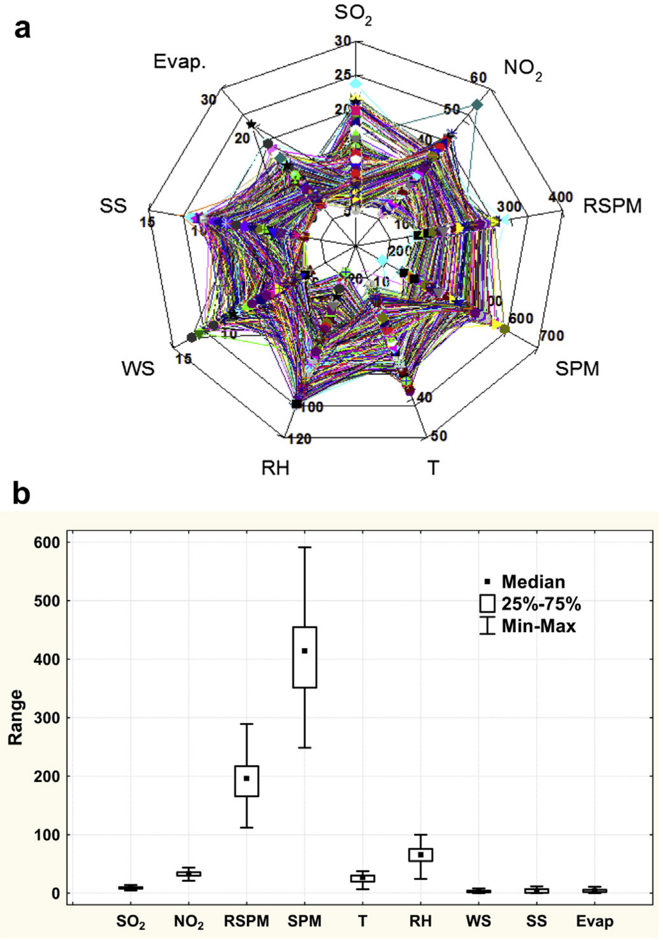


Fig. 1. The radar chart and Box–whisker plot of the ambient air quality variables and meteorological parameters in the Lucknow city.

seasonal variations on the ambient urban air composition, and (c) to find a regression function $\bar{f}_r$ capable of predicting the AQIs using the training data comprised of air quality and meteorological variables pertaining to the selected study area using the EL modeling approaches. Accordingly, in this study, we constructed the DT models (SDT, DTF, and DTB) for classification and regression. A conventional well known machine learning method, SVM is employed as the benchmark model.

2.1. The study area and data set

The data set considered in this study pertains to the Lucknow city (India) representing the ambient air quality (http//www.cpcbedb.nic.in) and meteorological conditions over a period of five years (January 2005–December 2009). Lucknow (26° 52′ N Latitude and 80° 56′ E Longitude) is located in the northern part of India. Demographic and other details of the study area are given elsewhere (Singh et al., 2012). The study region has a semi-arid climate and has three well demarcated seasons of winter (October–February), summer (March–June), and monsoon (July–September) and receives an average annual rainfall of about 100 cm. The mean temperature during the year varies between 3 °C and 45 °C. The common air pollution in the study area is mainly constituted by the emission from the domestic and vehicular sources. The concentrations of the air pollutants in the city represent the area background and are not affected by the local sources. The data set was comprised of total 1407 air quality measurements (24 h) for SO_2 (μg m^{-3}), NO_2 (μg m^{-3}), SPM (μg m^{-3}), and RSPM (μg m^{-3}) levels, and the meteorological parameters, such as the air temperature, T (°C), relative humidity, RH (%), wind speed, WS (km h^{-1}), evaporation (mm), and daily sunshine period, SS (h). Sampling, analysis and other measurements were accomplished as per standard protocols (Singh et al., 2012). The data set was analyzed statistically by generating the radar chart (Fig. 1) for all the variables. A radar chart is a graphical method that displays multivariate data in the form of a two-dimensional chart with several quantitative variables representing on axis starting from the same point. Further, the basic statistics of the daily concentration of air pollutants and metrological parameters during the summer, monsoon, and winter seasons are presented in Table 1

2.2. Data processing

In this study, both the single pollutant and combined pollutants based AQIs were calculated for all the data points. Single pollutant based AQI were calculated using the USEPA method (EPA, 1999).

$$I_p = \left[\frac{(I_{\mathrm{Hi}} - I_{\mathrm{Lo}})}{(\mathrm{BP}_{\mathrm{Hi}} - \mathrm{BP}_{\mathrm{Lo}})}\right](C_p - \mathrm{BP}_{\mathrm{Lo}}) + I_{\mathrm{Lo}} \tag{1}$$

where I_p is the AQI for the pollutant, p; C_p is the actual ambient concentration of the pollutant, p; $\mathrm{BP}_{\mathrm{Hi}}$ is the break point for the pollutant that is greater than or equal to C_p; $\mathrm{BP}_{\mathrm{Lo}}$ is the break point for the pollutant that is less than or equal to C_p; I_{Hi} is the sub-index value corresponding to $\mathrm{BP}_{\mathrm{Hi}}$; and I_{Lo} is the sub-index value corresponding to $\mathrm{BP}_{\mathrm{Lo}}$. The AQI values for each of the pollutants considered here were calculated using eq. (1) and the pollutant responsible for the highest index value was selected for modeling (EPA, 1999).

The combined air quality index (CAQI) was calculated by combining the individual pollutant's index, R_{i} (Senthilnathan, 2007);

$$R_i = \frac{\text{Concentration of pollutant}}{\text{Standard value of pollutant}} \tag{2a}$$

Table 1
Statistics of concentration of air pollutants and meteorological parameters during the summer, monsoon and winter seasons in the study area.

Parameter	Unit	Summer			Monsoon			Winter		
		Range	Mean	SD	Range	Mean	SD	Range	Mean	SD
SO_2	μg m^{-3}	6.6–20.0	9.9	2.1	4.8–13.3	7.7	1.1	6.7–23.7	10.5	3.2
NO_2	μg m^{-3}	24.0–55.0	33.0	3.7	20.0–42.8	29.8	4.6	23.5–43.3	34.0	3.6
RSPM	μg m^{-3}	136.0–272.0	199.8	21.1	112.0–228.0	156.8	14.8	124.0–289.0	208.5	26.7
SPM	μg m^{-3}	291.0–523.0	421.9	41.0	187.5–466.0	330.5	31.4	279.0–591.0	438.3	55.9
T	°C	10.6–37.8	29.0	4.3	13.3–32.8	29.3	1.8	6.7–30.6	19.1	4.4
RH	%	16.5–100.0	49.8	17.8	59.5–100.0	79.3	8.8	35.5–99.0	65.7	9.3
WS	km h^{-1}	0.0–12.2	4.1	2.1	0.5–10.0	3.0	1.8	0.2–13.0	2.5	2.0
SS	h	0.0–10.9	4.8	3.3	0.0–11.9	3.5	3.3	0.0–9.7	3.6	3.1
Evaporation	mm	0.1–16.5	6.7	2.7	0.4–8.2	3.8	1.6	0.2–20.9	2.5	1.4

SD-standard deviation.

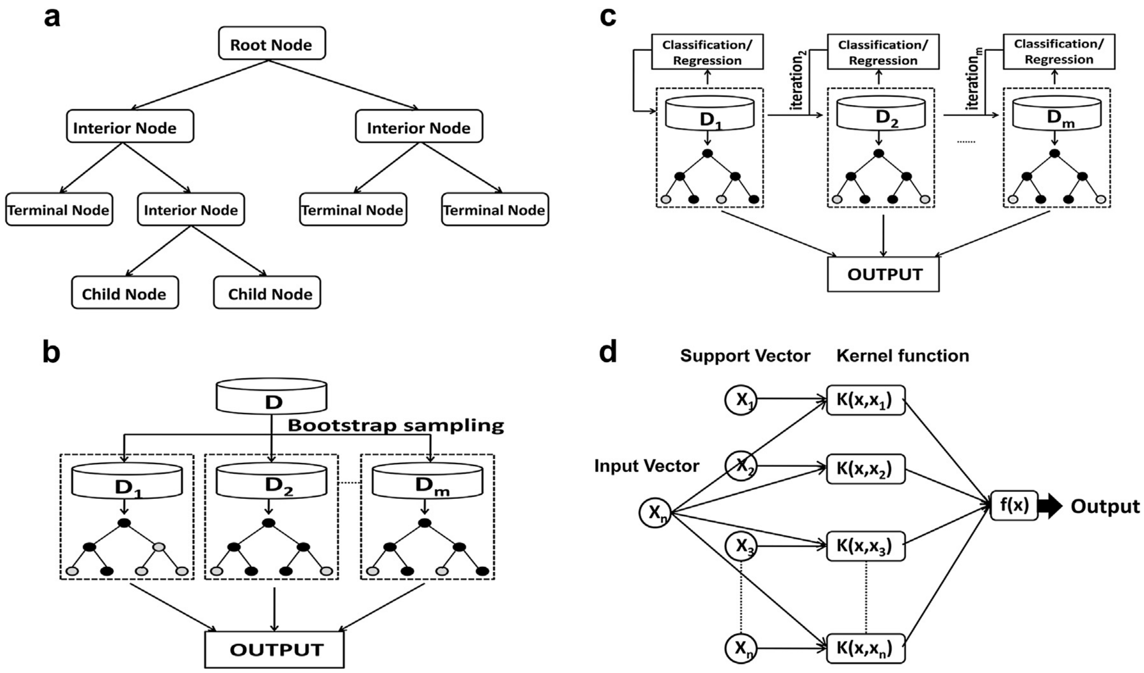


Fig. 2. Conceptual diagram of (a) SDT, (b) DTF, (c) DTB, and (d) SVM models.

The CAQI for *i* number of pollutant is given by the value obtained and defined as above R_i;

$$\text{CAQI} = \left(\sum_i \frac{R_i}{i}\right) \times 100 \tag{2b}$$

Here, both the AQIs and seasons were considered as dependent variables in regression and classification modeling, respectively, whereas the air quality and meteorological parameters were taken as independent variables. The data were then partitioned into two subsets; training and test, using the Kennard–Stone (K–S) approach. The K–S algorithm designs the model set in such a way that the objects are scattered uniformly around the training domain. Thus, all sources of the data variance are included into the training model (Basant et al., 2010). In the present study, the complete data set (1407 samples x 10 variables) was partitioned as training (985 samples × 10 variables) and test (422 samples × 10 variables) set. Thus, the training and test sets comprised of 70% and 30% samples, respectively. Since, all the independent variables (air quality and meteorological) in data exhibited significant correlations with the dependent variables, all of them were considered for classification and regression.

2.3. Nonlinearity in data

Nonlinearity in the data was tested using the Brock–Dechert–Scheinkman (BDS) statistics (Brock et al., 1996). BDS is a two-tail nonparametric method for testing the serial independence and nonlinear structure in a data based on the correlation integral. It tests the null hypothesis of independent and identically distributed (I.I.D.) data against an unspecified alternative. The BDS statistics is defined (Brock et al., 1996) as;

$$\text{BDS}_{\varepsilon,m} = \sqrt{N}\,\frac{\left[C_{\varepsilon,m} - \left(C_{\varepsilon,m}\right)^m\right]}{\sigma_{\varepsilon,m}} \tag{3}$$

where $\sigma_{\varepsilon,m}$ is the standard deviation of $C_{\varepsilon,m}$. If the computed BDS statistics exceeds the critical value at the conventional level, the null hypothesis of linearity is rejected, which reveals the presence of nonlinear dependence in the data (Anoruo, 2011).

2.4. Principal components analysis

PCA is a pattern recognition technique that attempts to explain the variance of a large set of inter-correlated variables. It indicates association between variables, thus, reducing the dimensionality of the data set. PCA extracts the eigenvalues and eigenvectors from the covariance matrix of original variables. The principal components (PCs) are the uncorrelated (orthogonal) variables, obtained by multiplying the original correlated variables with the eigenvector (loadings). The eigenvalues of the PCs are the measure of their associated variance, the participation of the original variables in the PCs is given by the loadings, and the individual transformed observations are called scores (Singh et al., 2005). PCA was performed on the complete data set standardized through *z*-scale transformations. Standardization tends to minimize the influence of variance difference of variables and eliminates the effect of different units of measurement and renders the data dimensionless (Singh et al., 2005). Here, PCA was performed to identify the pollution sources in the study area and to understand the influences of meteorological parameters on their levels.

2.5. Ensemble learning approaches

In ensemble learning (EL) multiple learners are trained to solve the same problem. An ensemble contains a number of base learners (Ishwaran and Kogalur, 2010) and their generalization ability is usually much stronger. Bagging and boosting are considered here for constructing the classification and regression decision tree models (DTF and DBT) to analyze the air quality data. Bagging uses different perturbed data sets and different feature sets for training base classifiers, whereas in boosting, diversity is obtained by increasing the weights of misclassified samples in an iterative manner (Yang et al., 2010). These methods use decision trees as base classifiers because DTs are sensitive to small changes on the training set (Dietterich, 2000). Simple averaging or majority voting is commonly used to aggregate the base classifiers. However, the primary disadvantage of DTF and DTB is that the models are complex and cannot be visualized like a single tree. Therefore, the SDT model was also considered here, as it provides intuitive understanding of how the predictor variables relate.

2.5.1. *Single decision tree*

SDT analysis can be used both for classification and regression problems and has number of features, including ability to deal with collinear data, to exclude insignificant variables, and to allow asymmetrical distribution of samples (Coops et al., 2011). A SDT is comprised of nodes and each node represents a set of records (rows) from the original data set. Nodes may further be divided into child nodes known as interior nodes, and those which do not have child nodes are known as terminals or leaf nodes (Fig. 2a). Unlike a real tree, a SDT starts with the root node on the top which represents all the rows in the data set (Swamy and Hanumanthappa, 2012). SDT algorithms considered here use the GINI index to evaluate the quality of splits for building the classification tree and minimum variance (least-square criteria) within nodes for regression tree (line and function fitting). The optimal split of a node is that ensuring the lowest GINI splitting index (ideally, zero). The impurity of each node is calculated by examining the distribution of categories of the target variable for the rows in the group. A pure node, where all rows have the same value of the target variable has an impurity value of zero (Gorunescu, 2011). A typical SDT learning algorithm adopts a top-down recursive divide-end-conquer strategy to construct a tree. Over-fitting of the training data may be prevented by the pruning technique that removes some branches of the tree after the tree is constructed. The maximum depth of tree and terminal (leaf) nodes are the method parameters which need to be optimized.

2.5.2. *Decision tree forest*

A DTF is an ensemble of SDTs whose predictions are combined to make the overall prediction for the forest (Fig. 2b). In DTF, a large number of independent trees are grown in parallel, and they do not interact until after all of them have been built. Bootstrap resampling method (Efron, 1979) and aggregating are the basis of bagging which is incorporated in DTF. Different training sub-sets are drawn at random with replacement from the training data set. Separate models are produced and used to predict the entire data from aforesaid sub-sets. Then various estimated models are aggregated by using the mean for regression problems or majority voting for classification problems. Theoretically in bagging, first a bootstrapped sample is constructed as (Erdal and Karakurt, 2013):

$$D_i^* = \left(Y_i^*, X_i^*\right) \tag{4a}$$

where D consists of data $\{(X_i, Y_i),\ i = 1,2,\ldots,n\}$, Y_i the real-valued response and X_i a p-dimensional predictor variable for the ith instance. Secondly, the bootstrapped predictor is estimated by the plug-in principle.

$$C_n^*(x) = h_n\left(D_i^*, \ldots\ldots, D_n^*\right)(x) \tag{4b}$$

where $C_n(x) = h_n\ (D_1, \ldots, D_n)\ (x)$ and h_n is the nth hypothesis Finally, the bagged predictor is;

$$C_{n,B}(x) = E^*\left[D_n^*(x)\right] \tag{4c}$$

Bagging can reduce variance when combined with the base learner generation with a good performance (Wang et al., 2011). The DTFs gaining strength from bagging technique use the out of bag data rows for model validation. This provides an independent test set without requiring a separate data set or holding back rows from the tree construction. The stochastic element in DTF algorithm makes it highly resistant to over-fitting. The DTF parameters are the size of each bag (as a percentage); the number of iterations (number of trees).

2.5.3. *Decision Treeboost*

Treeboost (TB) algorithm combines the strengths of regression tree and boosting. Boosting is a technique for improving the accuracy of a predictive function by applying function repeatedly in a series and combining the output of each function with weighting, so that the total error of prediction is minimized (Friedman, 2002). Gradient boosting is a sequential stage-wise forward iterative algorithm to find an additive predictor (Fig. 2c). The TB algorithm creates a tree ensemble and it uses randomization during the tree creations. The goal is to minimize the loss function in the training set, $\{\mathbf{x},y\}$. After each iteration, F represents sum of all trees built so far:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \mathrm{Tree}_m(\mathbf{x}) \tag{5a}$$

Regardless of the loss-function, the trees fitting the gradient on pseudo residuals are regression trees trained to minimize mean squared error (MSE). Regularization parameter is the number of gradient boosting iterations, m (the number of trees in the model when the base learner is a SDT). An important part of gradient boosting method is regularization by shrinkage which consists in modifying the update rule as;

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \upsilon \cdot \gamma_m h_m(\mathbf{x}). \quad 0 < \upsilon \leq 1 \tag{5b}$$

where parameter υ is called the learning rate and $h_m(\mathbf{x})$ is the base learner. The size of the tree is an important consideration for model accuracy. Large size trees are pruned using backward pruning procedure, thus rendering optimal size trees in the model. Optimal size of the tree was decided using the criteria of minimal cross-validation error. The TB model for classification is essentially the same as for regression except logit (probability) values are fitted rather than raw target values. At the end of the process, the category that minimizes the misclassification cost is chosen as the predicted value. The DTB uses the Huber M − regression loss function which makes it highly resistant to outliers (Huber, 1964). The number of terminal nodes and depth of the trees are the method's parameter which can be adjusted for a data set at hand. It controls the maximum allowed level of interaction between the variables in the model.

2.6. *Support vector machines*

SVMs are based on the concepts from statistical learning theory (Vapnik, 1995) and can be used both in classification and regression. For a data set, D comprised of N pairs, $(\mathbf{x}_i, y_i)$, in which $\mathbf{x}_i \in R^N$ and $y_i \in \{-1, +1\}$, SVMs construct a hyper plane $w.\mathbf{\Phi}(\mathbf{x}) + b = 0$ able to separate the data in D with minimum error maximizing the margin of separation between the classes. Here, $\mathbf{\Phi}$ is a function that maps the data in D a higher dimension space, making the classes linearly separable. In SVM training and predictions, the mapping function appears as dot products in the form $\mathbf{\Phi}(\mathbf{x}_i)$. $\mathbf{\Phi}(\mathbf{x}_j)$, which can be computed by the kernel functions. In case of regression, SVM gives a decision function:

$$f(x) = \mathrm{sign}\left(\sum\nolimits_{i=1}^{N} \propto_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{6}$$

where α_i is the coefficient to be learned and K is the kernel function. Parameter α_i is trained through maximizing the Lagrangian expression, such that $C \geq \propto_i \geq 0, i = 1, \ldots, N$, where C is a positive constant (Singh et al., 2011). Here, we used Gaussian radial kernel function for which the most important parameter is the width γ, controlling the amplitude of the kernel function and, hence, the generalization ability of SVM (Noori et al., 2011). A schematic diagram of the SVM model constructed here is shown in Fig. 2d. The main deficiency of SVMs that concerns is the difficulty of interpreting the generated model and their sensibility to a proper parameter tuning.

2.7. Modeling performance criteria

The performance of the classification models was assessed in terms of the misclassification rate (MR), sensitivity, specificity, accuracy of prediction, and Matthew's correlation coefficient (MCC) computed using the confusion matrix (Cheng et al., 2011; Singh et al., 2011). It may be mentioned that accuracy of test depends on how well the test separates the group being tested. Sensitivity is the most important parameter in a classification model. In fact, the low sensitivity value indicates the low ability of a model to recognize the true positives. The specificity is another important indicator. High specificity value indicates the high ability of the model to recognize the false positives (Cheng et al., 2011). MCC value equal to 1 is regarded as a perfect prediction, whereas, 0 is for a completely random prediction.

The performance of each of the regression models constructed here was evaluated using different statistical criteria parameters: the root mean square error (RMSE), mean absolute error (MAE), and the correlation coefficient (R) (Singh et al., 2010, 2013) between the measured and predicted values of the response. Each performance criteria term described above conveys specific information regarding the predictive performance efficiency of a specific model. Residuals analysis was also performed to assess the adequacy of the regression models.

3. Results and discussion

A radar chart of the considered variables is shown in Fig. 1. The radar chart shows that the variables used here are distributed over a sufficiently large space. Table 1 summarizes the basic statistics of the daily concentration of the air pollutants and meteorological parameters during the summer ($n = 475$), monsoon ($n = 359$), and winter ($n = 573$) seasons. Among the air pollutants, the respective mean concentrations of SPM, RSPM, NO_2 and SO_2 were 421.90 μg m^{-3}, 199.81 μg m^{-3}, 33.04 μg m^{-3}, and 9.86 μg m^{-3} (summer), 330.52 μg m^{-3}, 156.76 μg m^{-3}, 29.84 μg m^{-3}, and 7.70 μg m^{-3} (monsoon), and 438.28 μg m^{-3}, 208.47 μg m^{-3}, 34.03 μg m^{-3}, and 10.53 μg m^{-3} (winter).

Histograms of the calculated AQI and CAQI values for summer, monsoon and winter seasons are shown in Fig. 3. Frequencies of the calculated values of the AQI and CAQI in various sub-ranges (bins) are represented as vertical bars. The histograms show nearly normal distribution of calculated AQI and CAQI values. AQIs are related to the overall status of air pollution via pre-defined set of clearly identified criteria. These criteria should be universal and irrespective of the level of pollution. It should be sufficiently flexible to account for different level of population exposure, variable meteorological and climatic conditions occurring in the area as well as the sensitivity of flora and fauna (EPA, 1998). The computed values of AQIs were compared with the rating scale to assess the degree of pollution in the ambient air (EPA, 1999; Senthilnathan and Rajan, 2002). The rating scale of AQI is defined as range to know the effect of pollutants on human population. The calculated AQI and CAQI values for the ambient urban air of Lucknow ranged between 91–159 and 82–145 for summer, 79–137 and 69–130 for monsoon and 85–168 and 77–156 for winter, respectively. It is evident that according to the EPA rating (EPA, 1999), the urban air quality of Lucknow city is categorized as unhealthy during summer and winter seasons, whereas according to the CAQI values, the air quality in the study area is severely polluted in all the three seasons and it is unhealthy to children, asthmatics and people suffering with bronchial disease (Senthilnathan, 2007).

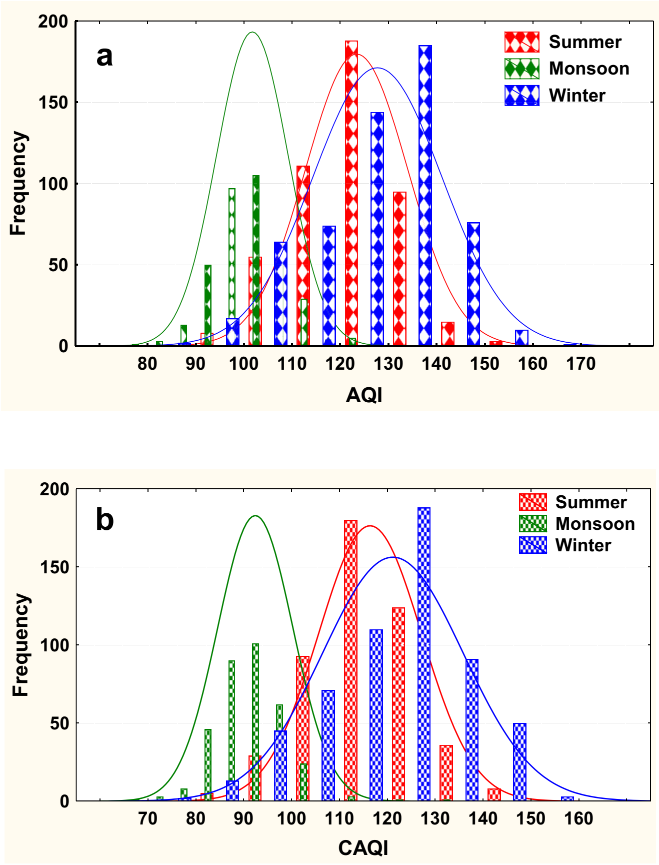


Fig. 3. Histogram of the (a) Seasonal AQI, and (b) Seasonal CAQI values corresponding to the ambient urban air of Lucknow.

The nonlinear dependence of the data was estimated using the BDS statistics. BDS extracts linear structure in the data by use of an estimated linear filter. The BDS statistics was calculated using eq. 3 ($m = 2$ to 5 and $\varepsilon = 0.5$) and the null hypothesis of linearity is rejected if the computed test statistics exceeds the critical value at the conventional level. In our case, the BDS statistics exceeded the p value ($p < 2.2 \times 10^{-16}$), thus suggesting for severe nonlinear data structure and hence, a nonlinear model is required for developing an appropriate regression function.

3.1. Source identification

PCA was performed to determine correlations between pollutants and meteorological parameters, and to identify the source profiles of various air pollutants in the study area. PCA yielded three PCs with eigen value greater than 1 and accounting for 77.5% of the variance. Plots of PC1 versus PC2 and PC3 are shown in Fig. 4. PC1 explaining 35.2% of data variance indicates significant correlation among traffic and emission variables, which may be attributed to their matching diurnal cycle resulting in high PC loadings (>0.70). An inverse relationship of air temperature (T) with SO_2, RSPM and SPM may be attributed to the fact that consumption of fuels

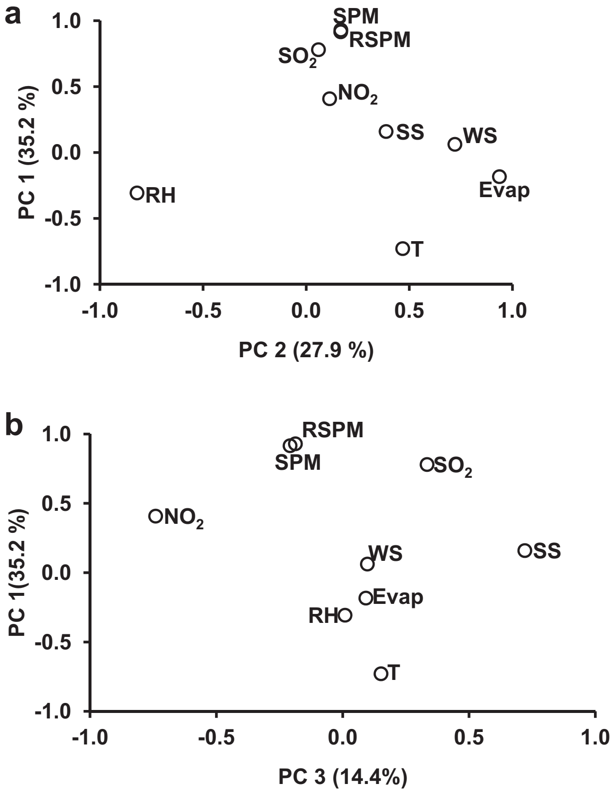


Fig. 4. Loading plots of (a) PC1 versus PC2, and (b) PC1 versus PC3.

depends on the air temperature and it is not the primary parameter that affects the diffusion condition of pollution. Thus, temperature is considered a pollution control parameter (Akpinar et al., 2009). PC2 capturing 27.9% of variance shows a high correlation among meteorological variables i.e. RH, WS, and evap. PC3 explaining 14.4% of variance correlated NO_2 and SS inversely. This may be due to their matching diurnal behavior – elevated during daytime and depleted during night. Main sources of oxides of sulfur and nitrogen arise from the road traffic, emission from fuel fired equipments and industries. There are total 1.2 million vehicles registered in the city with annual consumption of about 12 million liters petrol and 12.6 million litres diesel (Singh et al., 2012).

3.2. Air quality prediction modeling

Classification and regression models were constructed using the SDT, DTF, DTB and SVM approaches. Classification models were developed to discriminate the ambient urban air quality during different seasons (summer, monsoon, and winter) in Lucknow, and to identify the factors responsible for discrimination. Regression models were constructed to predict the AQIs for Lucknow city using set of independent variables with a view to enumerate health risk associated with prevailing air quality. Architecture and parameters of the constructed models were optimized using the V-fold cross-validation (CV) procedure. To perform a V-fold CV, data were first partitioned into V-folds. After making V-folds, iterations of training and validation are performed such that, within each iteration, a different fold of data held out for validation, while the remaining V-1 folds are used for learning and subsequently the learned models are used to make predictions about the data in the validation fold. Thus, each time, a model is constructed and tested with an unseen data set. The advantage of this method is that it performs reliable and unbiased testing on data sets (Singh et al., 2013).

3.2.1. Classification modeling

SDT, DTF, DTB, and SVM approaches were used to construct classification functions to discriminate the ambient air during different seasons in Lucknow using the air quality and meteorological parameters. Here, season was the category variable. Such a discriminatory tool would help to evaluate the influences of seasonal variations on the ambient air quality and to understand the responsible factors for discrimination. The optimal parameters for these models were determined using 10-fold CV procedure and misclassification rate (MR) in training and validation data as the criteria. The average values of MR (ten runs) in training and CV for different models were: 7.21% and 12.49% (SDT), 2.54% and 11.37% (DTF), 4.87% and 12.18% (DTB), 5.28% and 11.88% (SVM). The optimal parameters for SDT such as maximum depth of tree, total group splits, and terminal (leaf) nodes were 10, 48, and 34, respectively. In DTF and DTB, number of trees, Depth, and average group split were 195, 14, 66.3, and 395, 4, 113, respectively. The value of C, γ and number of support vectors (SVs) for optimal SVM model were 0.98, 11, and 330, respectively. The optimal models were then applied to the test and complete data arrays. The sensitivity, specificity, accuracy, and MCC values corresponding to the constructed models in training, test and complete data sets are given in Table 2.

The optimal seasonal discriminatory models rendered MR of 8.32% (SDT), 4.12% (DTF), 5.62% (DTB), 6.18% (SVM), respectively in complete data. All the classification models identified all the meteorological and pollutant parameters (T, RH, RSPM, Evap, SO_2, SS, NO_2, WS) as the discriminating variables between the seasons in the study area. Fig. 5 shows the relative contribution of these variables in various discriminating functions. It is evident that contribution of these variables in DT models ranged between 1.9% and 100%. The discriminating variables in each model were determined in view of their importance in corresponding model. The importance of independent variables in each model was determined using the difference in MR calculated using actual data values of all predictors and those computed through randomly rearrangement values of the each predictor. From Table 1, it is evident that all the four pollutants exhibited relatively higher mean concentrations during the winter followed by summer, and monsoon seasons. It may be attributed to the fact that during the winters, the pollutants emitted from various anthropogenic and natural sources are trapped in the boundary layer due to frequent temperature inversions, while in the summer months, this polluted air mixes well with the free tropospheric air causing dilution of the pollutants and the greater photochemical reactions due to higher solar radiation. In the monsoon season, the precipitation washes out the atmospheric pollutants. Seasonal variations in pollutants concentrations are associated with the change in meteorological parameters (T, RH, WS, SS, and Evap), which reflect the change in solar radiation and boundary layer stability (Hassan et al., 2013). Further, the classification results (Table 2) show that accuracy, sensitivity and specificity values rendered by different models in complete data were more than 93%, 85% and 94%, respectively, whereas the MCC values ranged between 0.85 and 0.96.

The average gain values in training and validation data for the selected season-based classification models of ambient air in Lucknow ranged 1.742–2.216 and 1.638–2.084, respectively. The gain values show how much improvement the model provides in picking out the best of the cases. The gain of 1 means we are not doing any selective targeting (Abdelaal et al., 2010). The performance criteria parameters (sensitivity, specificity, accuracy, MCC), and values of average gain for the classification models suggest that all the DT and SVM modeling approaches are capable to discriminate the seasonal influences on ambient air quality and to identify the responsible variables in a quantitative manner. However, the

Table 2
Performance parameters of seasonal air quality classification models.

Model	Class	Total cases	Correct assignations	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Training set							
SDT	Summer	392	364	92.86	95.28	94.31	0.88
	Monsoon	218	204	87.93	98.14	95.74	0.88
	Winter	375	346	95.84	95.35	95.53	0.90
	Total	985					
DTF	Summer	392	377	98.43	97.51	97.87	0.96
	Monsoon	218	214	95.11	99.47	98.48	0.96
	Winter	375	369	97.88	99.01	98.58	0.97
	Total	985					
DTB	Summer	392	368	95.34	95.99	95.74	0.91
	Monsoon	218	209	93.72	98.82	97.66	0.93
	Winter	375	360	95.74	97.54	96.85	0.93
	Total	985					
SVM	Summer	392	365	95.55	95.52	95.53	0.91
	Monsoon	218	207	91.59	98.55	96.95	0.91
	Winter	375	361	95.76	97.70	96.95	0.94
	Total	985					
Test set							
SDT	Summer	83	58	84.06	92.92	91.47	0.72
	Monsoon	141	136	82.93	98.06	92.18	0.84
	Winter	198	182	96.30	93.13	94.55	0.89
	Total	422					
DTF	Summer	83	61	93.85	93.84	93.84	0.80
	Monsoon	141	139	85.28	99.23	93.84	0.87
	Winter	198	189	97.42	96.05	96.68	0.93
	Total	422					
DTB	Summer	83	63	94.03	94.37	94.31	0.81
	Monsoon	141	139	86.88	99.24	94.55	0.89
	Winter	198	189	96.92	96.04	96.45	0.93
	Total	422					
SVM	Summer	83	60	93.75	93.58	93.60	0.79
	Monsoon	141	135	86.54	97.74	93.60	0.86
	Winter	198	192	95.05	97.27	96.21	0.92
	Total	422					
Complete set							
SDT	Summer	475	422	91.54	94.40	93.46	0.85
	Monsoon	359	340	85.86	98.12	94.67	0.87
	Winter	573	528	96.00	94.75	95.24	0.90
	Total	1407					
DTF	Summer	475	438	97.77	96.14	96.66	0.93
	Monsoon	359	353	90.98	99.41	97.09	0.93
	Winter	573	558	97.72	98.21	98.01	0.96
	Total	1407					
DTB	Summer	475	431	95.14	95.39	95.31	0.89
	Monsoon	359	348	90.86	98.93	96.73	0.92
	Winter	573	549	96.15	97.13	96.73	0.93
	Total	1407					
SVM	Summer	475	425	95.29	94.80	94.95	0.89
	Monsoon	359	342	89.53	98.34	95.95	0.90
	Winter	573	553	95.51	97.58	96.73	0.93
	Total	1407					

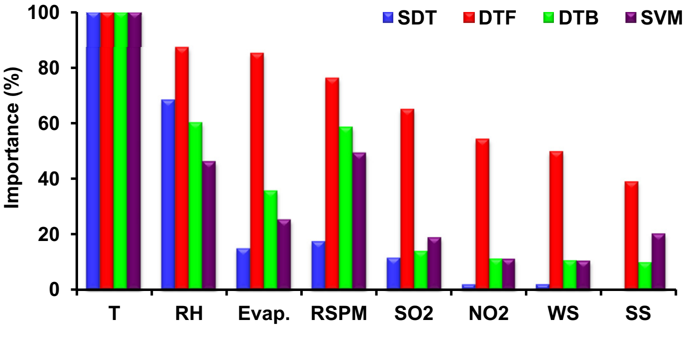


Fig. 5. Contribution of the predictor variables in the various classification models.

Table 3
Optimal parameters for regression models.

Model parameters	Regression model	
	AQI	CAQI
Single decision tree (SDT)		
Max. depth of tree	10	10
Total group split	120	129
Terminal (leaf) nodes	121	130
Decision tree forest (DTF)		
Number of tree in forest	200	200
Max. depth of any tree	28	19
Average group split	216.7	214.0
Decision treeboost (DTB)		
Tree in full series	400	400
Max. depth of any tree	9	10
Average group split in each tree	827.8	421.4
Support vector machine (SVM)		
C	73.44	11.85
γ	1.48	5.19
Support Vectors	660	807

ELs performed relatively better than SDT and SVM. Further, the performances of both the DTF and DTB models are comparable.

3.2.2. *Regression modeling*

Regression functions were developed for predicting the EPA method based AQI as well as the combined AQI (CAQI) using the SDT, DTF, DTB, and SVM approaches. The prediction models for AQI were constructed using SO_2, NO_2, and meteorological parameters, whereas, models for CAQI were based on meteorological parameters alone. A 10-fold CV method was used to derive the optimal model parameters. Correlation between predicted and measured response and corresponding RMSE in training and validation were the criteria parameters. The average (10 runs) values of R and RMSEs in CV for different regression models ranged between 0.831–0.874 and 7.37–8.58 (AQI), and 0.780–0.810 and 9.64–10.31 (CAQI), respectively. The optimal parameters of different models constructed for predicting AQI and CAQI of ambient air are provided in Table 3.

The respective contributions of selected variables in different models (AQI and CAQI) are shown in Fig. 6a,b. The models were then applied to the test and complete data arrays. The corresponding values of various statistical performance parameters computed from the predicted response values (AQI and CAQI) yielded by DTs and SVM models for the urban air quality in training, validation, and complete data arrays are presented in Table 4. The constructed SDT, DTF, DTB, and SVR models for AQI prediction explained data variance of 81.11%, 90.03%, 91.88% and 79.19% in complete data, whereas, for CAQI prediction, the respective models captured 67.90%, 84.55%, 85.73%, and 69.86% variance. Proportion of variance explained by model variables is the best single measure of how well the predicted values match the actual values. A model predicting exactly matching values with measured ones would explain 100% variance in data.

Table 4 shows that the relative correlation coefficient (R) differences between the SDT ($R_{\text{training}} = 0.911$, $R_{\text{test}} = 0.877$), DTF ($R_{\text{training}} = 0.972$, $R_{\text{test}} = 0.903$); DTB ($R_{\text{training}} = 0.984$, $R_{\text{test}} = 0.902$) and the SVR ($R_{\text{training}} = 0.887$, $R_{\text{test}} = 0.894$) models for predicting AQI in ambient air of Lucknow is 2.7%, 9.6%, and 10.9% in training and −1.9%, 1.0%, and 0.9% in the testing phase, respectively, whereas, in case of CAQI prediction models, the relative correlation coefficient differences were −0.7%, 15%, 16.5% in training and −2.5%, 0.4%, 0% in test set. This suggests that the ensemble regression models are superior to the corresponding SDT and SVM models. It may further be noted that there is a direct relationship between R, MAE, and RMSE in training phase and ELs are superior to SVM for minimizing RMSE and MAE (Table 4). The RMSE and MAE statistics are also consistent with the R-statistics in testing phase. RMSE is a quadratic scoring rule which measures the average magnitude of the error. It gives a relatively high weight to large errors, hence most useful when large errors are particularly undesirable. MAE measures the average magnitude of the error in a set of predictions, without considering their direction. It is a linear score which means that all the individual differences between predictions and corresponding measured values are weighted equally in the average. Further, the calculated and model predicted AQIs are plotted in Figs. 7 and 8. A closely followed pattern of variation by the calculated and model predicted AQIs values both in the training and test sets is evident suggesting that both the EL models performed reasonably well. Plots of the residuals and model predicted values of the response variable are known to provide more information regarding the model fitness to a data set. A random distribution of

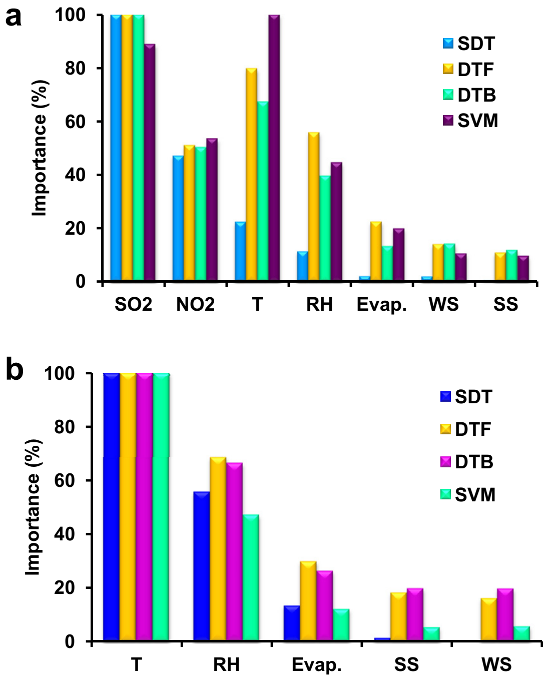


Fig. 6. Contribution of the predictor variables in the various regression models for prediction of (a) AQI, and (b) CAQI.

Table 4a
Statistical performance of the AQI prediction models.

Model	Sub-set	Mean	SD	MAE	RMSE	R
Calculated	Training	120.38	15.30	–	–	–
	Test	116.95	15.25	–	–	–
	Complete	119.35	15.36	–	–	–
SDT	Training	120.38	13.94	4.94	6.31	0.911
	Test	117.38	14.70	5.91	7.46	0.877
	Complete	119.48	14.23	5.23	6.67	0.901
DTF	Training	120.40	13.45	3.01	3.88	0.972
	Test	117.65	13.65	5.24	6.58	0.903
	Complete	119.58	13.56	3.68	4.85	0.951
DTB	Training	120.40	13.96	2.25	2.96	0.984
	Test	117.24	14.15	5.26	6.59	0.902
	Complete	119.45	14.09	3.15	4.38	0.959
SVM	Training	120.34	13.44	5.39	7.07	0.887
	Test	117.21	13.98	5.42	6.84	0.894
	Complete	119.40	13.67	5.40	7.00	0.890

SD-standard deviation.

Table 4b
Statistical performance of the CAQI prediction models.

Model	Sub-set	Mean	SD	MAE	RMSE	R
Calculated	Training	112.98	16.43	–	–	–
	Test	109.51	17.06	–	–	–
	Complete	111.94	16.69	–	–	–
SDT	Training	112.98	13.56	7.31	9.27	0.825
	Test	110.72	15.42	7.80	9.87	0.822
	Complete	112.30	14.18	7.46	9.45	0.825
DTF	Training	112.92	13.98	3.89	5.10	0.956
	Test	110.95	14.69	7.24	9.10	0.846
	Complete	112.03	14.25	4.89	6.56	0.922
DTB	Training	113.08	14.02	3.55	4.55	0.968
	Test	109.95	14.61	7.25	9.18	0.843
	Complete	112.14	14.27	4.66	6.30	0.929
SVM	Training	112.91	13.37	6.92	9.14	0.831
	Test	110.07	15.08	7.23	9.22	0.843
	Complete	112.06	13.96	7.01	9.16	0.836

SD-standard deviation.

the residuals suggests that the model fits the data well, whereas, a non-random distribution shows that the model does not fit the data adequately (Singh et al., 2013). Plots of the model-predicted response and the corresponding residuals for the training and test sets show almost complete independence and random distribution (Figures not shown due to brevity).

The obtained results indicate that the tree based ensemble models (DTF and DTB) are superior alternatives to the conventional machine learning (SVM) model both for classification and regression problems. Performance of SVM is largely dependent on the kernel function selection and parametric setting is vital for its forecasting accuracy. Moreover, SVM considering a limited data points (support vectors) in learning process, is a rigid method which may increase the prediction error of the model (Singh et al., 2011; Erdal and Karakurt, 2013). The tree-based ensemble models can reasonably increase their accuracy by generating many replica data sets and in creating various models, which has a lower bias, and then integrating them in building an ensemble model with higher performance (Chou et al., 2011). A relatively better performance of DTF and DTB models may be attributed to incorporation of bagging and stochastic gradient boosting algorithms (Grunwald et al., 2009). In bagging and boosting multiple version of SDTs are formed by making bootstrapped replicas of the learning set and subsequently using these as new learning sets. In bagging, each new independent resample is drawn at random with replacement from the entire learning data, however in boosting; the resampling for the next SDT depends on the performance of the previous SDT. In general, tree based ensembles can inherit almost all advantages of tree based models while overcoming their primary problem of inaccuracy (Chou et al., 2011; Erdal and Karakurt, 2013).

4. Conclusions

In this study, ambient urban air quality and meteorological data of Lucknow city was used. PCA approach identified the vehicular emissions and fuel combustion as the major air pollution sources in Lucknow. Air quality indices suggested that the urban air quality during summers and winters was unhealthy to humans. Tree based ensemble learning models were constructed to develop classification and regression functions for discriminating the air quality during different seasons and to predict the AQIs for the city using the five years databases. Several statistical parameters were computed to evaluate the generalization and predictive abilities of the proposed models. Ensemble learning methods (DTF and DTB) implementing bagging and stochastic gradient boosting algorithms noticeably enhanced the accuracy of SDT model and performed relatively better than SVM. The proposed models successfully predicted the seasonal influences identifying the discriminating variables and AQIs. These models can be used as tools in air quality prediction and management.

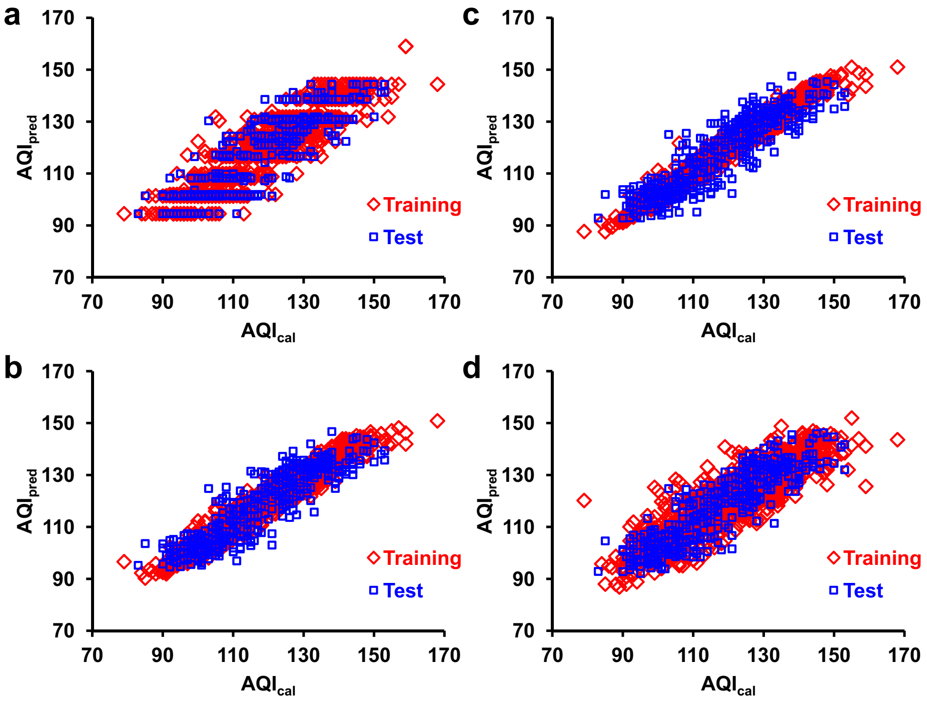


Fig. 7. Plot of the calculated and model predicted values of the AQI in training and test sets (a) SDT, (b) DTF, (c) DTB, and (d) SVM models.

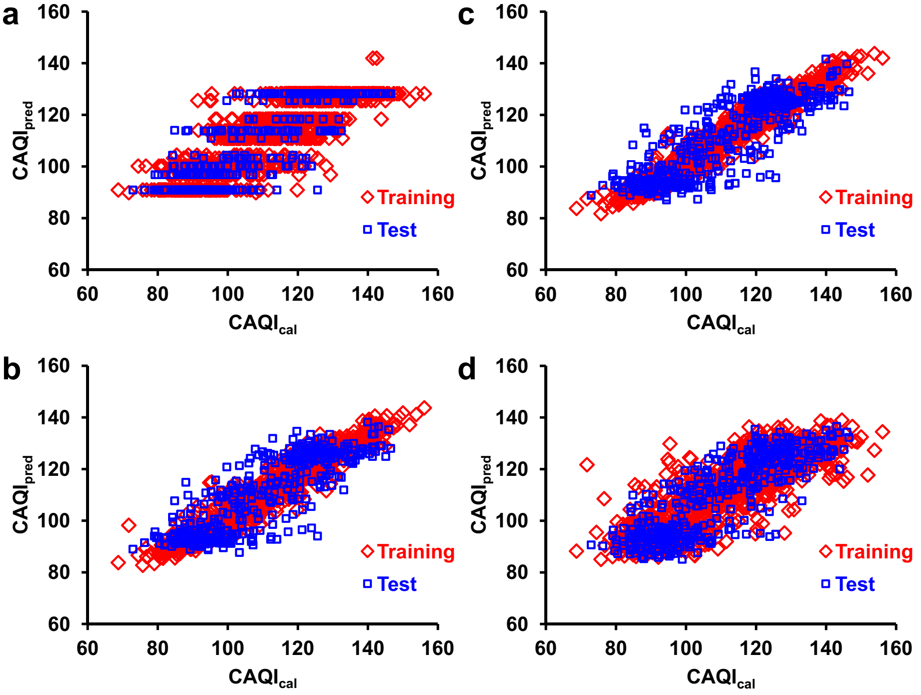


Fig. 8. Plot of the calculated and model predicted values of the CAQI in training and test sets (a) SDT, (b) DTF, (c) DTB, and (d) SVM models.

Acknowledgments

The authors thank the Director, Indian Institute of Toxicology Research, Lucknow (India) for his keen interest in this work.

References

Abdelaal, M.M.A., Farouq, M.W., Sena, H.A., Salem, A.B.M., 2010. Applied classification support vector machine for providing second opinion of breast cancer diagnosis. The Online Journal on Mathematics and Statistics 1, 1–7.

Akpinar, E.K., Akpinar, S., Öztop, H.F., 2009. Statistical analysis of meteorological factors and air pollution at winter months in elaziğ, Turkey. Journal of Urban and Environmental Engineering 3, 7–16.

Anoruo, E., 2011. Testing for linear and nonlinear causality between crude oil price changes and stock market returns. International Journal of Economic Sciences and Applied Research 4, 75–92.

Banerjee, T., Singh, S.B., Srivastava, R.K., 2011. Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India. Atmospheric Research 99, 505–517.

Basant, N., Gupta, S., Malik, A., Singh, K.P., 2010. Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water – a case study. Chemometrics and Intelligent Laboratory Systems 104, 172–180.

Brock, W.A., Dechert, W.D., Scheinkman, J.A., LeBaron, B., 1996. A test for independence based on the correlation dimension. Econometric Review 15, 197–235.

Bhaskar, B.V., Rajasekhar, R.V.J., Muthusubramaian, P., Kesarkar, A.P., 2008. Measurement and modeling of respirable particulate (PM_{10}) and lead pollution over Madurai, India. Air Quality, Atmosphere and Health 1, 45–55.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Cheng, F., Shen, J., Yu, Y., Li, W., Liu, G., Lee, P.W., Tang, Y., 2011. In silico prediction of *Tetrahymena pyriformis* toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. Chemosphere 82, 1636–1643.

Chattopadhyay, S., Gupta, S., Saha, R.N., 2010. Spatial and temporal variation of urban air quality: a GIS approach. Journal of Environmental Protection 1, 264–277.

Chou, J.S., Chiu, C.K., Farfoura, M., Al-Taharwa, I., 2011. Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques. Journal of Computing in Civil Engineering 25, 242–253.

Cimorelli, A.J., Perry, S.G., Venkatram, A., Weil, J., Paine, R., Wilson, R.B., Lee, R.F., Peters, E.D., Brode, R.W., 2005. AERMOD: a dispersion model for industrial source applications. Part I: general model formulation and boundary layer characterization. Journal of Applied Meteorology 44, 682–693.

Collett, R.S., Oduyemi, K., 1997. Air quality modeling: a technical review of mathematical approaches. Meteorological Applications 4, 235–246.

Coops, N.C., Waring, R.H., Beier, C., Roy-Jauvin, R., Wang, T., 2011. Modeling the occurrence of 15 coniferous tree species throughout the Pacific Northwest of North America using a hybrid approach of a generic process-based growth model and decision tree analysis. Applied Vegetation Science 14, 402–414.

CPCB, 2009. National Ambient Air Quality Standards. CPCB No. B-29016/20/90/PCI-L.

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Multiple Classifier Systems LBCS-1857, pp. 1–15.

Dockery, D.W., Pope, C.A., 1994. Acute respiratory effects of particulate air pollution. Annual Review of Public Health 15, 107–132.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7, 1–26.

EPA, 1998. National Air Quality and Emissions Trends Report 1997. Environmental Protection Agency 454:R98–016. Environmental Protection Agency, Oûce of Air Quality Planning and Standards, Research Triangle Park.

EPA, 1999. Air quality index reporting: final rule. Federal Register. Part III, 40 CFR Part 58.

EPA, 2005. Guideline on Air Quality Models (revised). US Environmental Protection Agency, Research Triangle Park: NC (p. 40 CFR 51).

Erdal, H.I., Karakurt, O., 2013. Advancing monthly stream flow prediction accuracy of CART models using ensemble learning paradigms. Journal of Hydrology 477, 119–128.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics and Data Analysis 38, 367–378.

Gorunescu, F., 2011. Data Mining Concepts, Models and Techniques, Intelligent System Reference Library. Springer-Verlag, Heidelberg. http://dx.doi.org/10.1007/978-3-642-19721-5.

Grunwald, S., Daroub, S.H., Lang, T.A., Diaz, O.A., 2009. Tree-based modeling of complex interactions of phosphorus loadings and environmental factors. Science of the Total Environment 407, 3772–3783.

Gupta, A.K., Karar, K., Ayoob, S., John, K., 2008. Spatio-temporal characteristics of gaseous and particulate pollutants in an urban region of Kolkata, India. Atmospheric Research 87, 103–115.

Hancock, T., Put, R., Coomans, D., Vander Heyden, Y., Everingham, Y., 2005. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. Chemometrics and Intelligent Laboratory Systems 76, 185–196.

Hassan, S.K., El–Abssawy, A.A., Khoder, M.I., 2013. Characteristics of gas–phase nitric acid and ammonium–nitrate–sulfate aerosol, and their gas–phase precursors in a suburban area in Cairo, Egypt. Atmospheric Pollution Research 4, 117–129.

http//www.cpcbedb.nic.in.

Huber, P., 1964. Robust estimation of a location parameter. Annals of Mathematical Statistics 53, 73–101.

Ishwaran, H., Kogalur, U.B., 2010. Consistency of random survival forests. Statistics & Probability Letters 80, 1056–1064.

Kassomenos, P., Papaloukas, C., Petrakis, M., Karakitsios, S., 2008. Assessment and prediction of short term hospital admission: the case of Athens, Greece. Atmospheric Environment 42, 7078–7086.

Kesarkar, A.P., Dalvi, M., Kaginlkar, A., Ojha, A., 2007. Coupling of the weather research and forecasting model with AERMOD for pollutant dispersion modeling. A case study for PM_{10} dispersion over Pune, India. Atmospheric Environment 41, 1976–1988.

Kumar, A., Goyal, P., 2011. Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research 2, 436–444.

Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., Gwynne, M., 1996. Urban air pollution in mega cities of the world. Atmospheric Environment 30, 681–686.

Mahjoobi, J., Etemad-Shahidi, A., 2008. An alternative approach for the prediction of significant wave heights based on classification and regression trees. Applied Ocean Research 30, 172–177.

Nastos, P.T., Paliatsos, A.G., Anthracopoulos, M.B., Roma, E.S., Priftis, K.N., 2010. Outdoor particulate matter and childhood asthma admissions in Athens, Greece: a time-series study. Environmental Health 45, 1–9.

Noori, R., Karbassi, A.R., Moghaddamnia, K., Han, D., Zokaei-Ashtiani, M.H., Farokhnia, A., Ghafari Gousheh, N., 2011. Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction. Journal of Hydrology 401, 177–189.

Senthilnathan, T., Rajan, R.D., 2002. Status of suspended particulate matter concentration in Chennai city during the year 2000. Indian Journal of Environmental Protection 22, 1383–1387.

Senthilnathan, T., 2007. Analysis of concentration of air pollutants and air quality index levels in the ambient air in Chennai city. Journal of Institution of Engineers 87, 3–7.

Singh, K.P., Malik, A., Mohan, D., Sinha, S., Singh, V.K., 2005. Chemometric data analysis of pollutants in wastewater – a case study. Analytica Chimica Acta 532, 15–25.

Singh, K.P., Basant, N., Malik, A., Jain, G., 2010. Modeling the performance of "up-flow anaerobic sludge blanket" reactor based wastewater treatment plant using linear and nonlinear approaches—A case study. Analytica Chimica Acta 658, 1–11.

Singh, K.P., Basant, N., Gupta, S., 2011. Support vector machine in water quality management. Analytica Chimica Acta 703, 152–162.

Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P., 2012. Linear and nonlinear modeling approaches for urban air quality prediction. Science of the Total Environment 426, 244–255.

Singh, K.P., Gupta, S., Ojha, P., Rai, P., 2013. Predicting adsorptive removal of chlorophenol from aqueous solution using artificial intelligence based modeling approaches. Environmental Science and Pollution Research 20, 2271–2287.

Snelder, T.H., Lamouroux, N., Leathwick, J.R., Pella, H., Sauquet, E., Shankar, U., 2009. Predictive mapping of the natural flow regimes of France. Journal of Hydrology 373, 57–67.

Swamy, M.N., Hanumanthappa, M., 2012. Predicting academic success from student enrolment data using decision tree technique. International Journal of Applied Information Systems 4, 1–6.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Wang, G., Hao, J., Ma, J., Jiang, H., 2011. A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications 38, 223–230.

Yang, P., Yang, Y.H., Zhou, B.B., Zomaya, A.Y., 2010. A review of ensemble methods in bioinformatics. Current Bioinformatics 5, 296–308.

Zhang, C.X., Zhang, J.S., Wang, G.W., 2008. An empirical study of using rotation forest to improve regressors. Applied Mathematics and Computation 195, 618–629.