Predicting Ozone Layer Concentration Using Multivariate Adaptive Regression Splines, Random Forest and Classification and Regression Tree

Sanjiban Sekhar Roy[1(✉)], Chitransh Pratyush[1], and Cornel Barna[2]

[1] School of Computer Science and Engineering, VIT University, Vellore, India
sanjibanroy09@gmail.com, c.pgitpit@gmail.com
[2] Automation and Applied Informatics, Aurel Vlaicu University of Arad, Arad, Romania
barna.cornel@gmail.com

Abstract. Air pollution is one of the major environmental worries in recent time. Abrupt increase in the concentration of any gas leads to air pollution. The cities are mostly affected due to the abundance of population there. One of the worst gaseous pollutants is OZONE (O3). In this paper, we propose three predictive models for estimation of concentration of ozone gases in the air which are Random Forest, Multivariate Adaptive Regression Splines and Classification and Regression Tree. Evaluation of the prediction models indicates that the Multivariate Adaptive Regression Splines model describes the dataset better and has achieved significantly better prediction accuracy as compared to the Random Forest and Classification and Regression Tree. A detailed comparative study has been carried out on the performances of Random Forest, Multivariate Adaptive Regression Splines and Classification and Regression Tree. MARS gives the result by considering less variables as compared to other two. Moreover, Random Forest takes a little more time for building the tree as the elapsed time was calculated to 45 s in this case. In addition, variable importance for each model has been predicted. Observing all the graphs Multivariate Adaptive Regression Splines gives the closest curve of both train and test set when compared. It can be concluded that multivariate adaptive regression splines can be a valuable tool in predicting ozone for future.

Keywords: Ozone · Multivariable adaptive regression spline · Random Forest · Classification and Regression Tree

1 Introduction

Urban atmospheric pollutants are increasing day by day. They are considered as one of the main causes of increased incidence of respiratory illness in citizens. It is now irrefutable that air pollution is being caused by large amount of Total Suspended Particulates (TSP) and respiratory particulate of Particulate Matter less than 10 μm in aerodynamic diameter that has numerous undesired consequences on human health [1]. Air pollutants in an area with good airflow quickly mixes with air and disperses

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 634, DOI 10.1007/978-3-319-62524-9_11

however when trapped in an area, pollutant concentration can increase rapidly which ultimately leads to degradation of air quality. In order to measure how polluted the air is Air Quality Index is examined while for properties of air we see Qualities of Air [2]. All these factors affect the ozone layer which is Earth's protective layer. It is a belt of naturally occurring ozone gas that sits 9.3–18.6 miles above Earth and serves as shield from harmful ultraviolet B radiation emitted by sun. Several steps like Montreal Protocol which declines emission of ODS (ozone depleting substances) have been taken. It is expected to result in a near complete recovery of ozone layer near the middle of 21st century. By 2012, the total combined abundance of anthropogenic ODS in the troposphere has decreased by nearly 10% from its peak value in 1994 [3, 5].

In the present era, there is a wide spread concern for ozone layer depletion due to the release of pollution. As particulate matter causes several kind of respiratory and cardiovascular disease, it also leads to ozone depletion which attracts more and more attention for air quality information. This shows the need for integration of different information system in a similar way as done by Birgersson et al. [6] for data integration using Machine Learning [6]. Prediction of air quality has thus become a necessary need to save the future. Machine learning has been applied in various fields [14, 15, 17]. Medical and other fields have also been covered by various classification techniques [18, 19, 22–25]. Just as application of rough set technique was done for data investigation by Roy et al. (2013), in this paper application of Random Forest, Multivariate Adaptive Regression Splines and Classification and Regression Tree techniques has been applied for predicting the concentration of ozone [13, 20, 21]. Chuanting Zhang and Dongfeng Yuan (2015) worked on Grained Air Quality Index Level Prediction Using Random Forest Algorithm on Cluster Computing of Spark [4]. Previously existing methods could not meet the demand of real time analysis so a distributed random forest algorithm is implemented using Spark on the basis of resilient distributed dataset and shared variable. Parallelized random forest is also used as prediction model. Estimation of benzene by on field calibration of an electronic nose has been carried by Vito et al. (2008) in which gas multi-sensor played an important role which helps to raise the density of the monitoring network. But their concentration estimation capabilities are seriously limited by the known stability and selectivity issues of solid-state sensors they often rely on. Sensor fusion algorithm used in regression need to be properly tuned via supervised learning. But this training was revealed to be unsuccessful [7, 12]. Forecasting and prediction of things has become an essential part for future life. Roy et al. (2015) worked on prediction of Stock Market Forecasting using Lasso Linear regression model [16]. Vito et al. (2009) worked on CO, NO_2 and NO_x urban pollution monitoring. Some authors have used gas multisensor devices as a tool for densening the urban pollution monitoring mesh due to the significantly low cost per unit [8]. But the drawback is that these sensors are not reliable for long term and selectivity issues. In this paper we concentrate on regression technique Multiple Adaptive Regression Spline (MARS) for Air Quality dataset. Hui et al. (2013) used this regression model for prediction of emission of CO_2 in ASEAN countries. A comparative study of multiple regression (MR) and multiple adaptive regression splines (MARS) was carried for statistical modelling of CO_2 over period of 1980–2007 [9]. MARS model was concluded as more feasible and with better predictive ability. This paper shows the comparison of regression techniques like Random Forest, Multivariate

Adaptive Regression Splines and Classification and Regression Tree on Air Quality data showing the prediction using Salford Predictive Modeller.

This paper is organised as follows. Section 2 overviews proposed techniques of Random Forest, Multivariate Adaptive Regression Splines and Classification and Regression Tree. Section 3 gives the experimental setup and the steps involved in performing the regression techniques on the given dataset. Section 4 displays the results and discussion. Section 5 concludes the paper.

2 Proposed Techniques

To work with Salford Modeller it is necessary to know the working of the regression techniques that are going to be used. All are type of machine learning like a computer program is said to learn from experience 'E' with respect to some class of tasks 'T' and performance measure 'P' if its performance at tasks in 'T' as measured by 'P' improves with experience 'E'. All these have been used for prediction of ozone concentration by extracting knowledge from dataset.

A. Random Forest *algorithm*

It is a tree-based ensemble learning method involving the combination of several models to solve a single prediction problem. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method. It may also be said as a collection of many CART trees that are not influenced by each other when it is constructed [4]. It works as a large collection of decorrelated decision trees. It comes under bagging technique (average noisy and unbiased models to create a model with low variance). Tress are based on random selection of data as well as variable. This develops lots of decision trees based on random selection of data and random selection of variables. After all the tree are built the data get fed in the tree and proximities are calculated for each pair of cases. If any two cases occupy the same terminal node, their proximity is changed and incremented by one. At last, the proximities get normalized by dividing it by the number of trees. Proximities can be used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data. It serves as one of the most useful tools in random forests. The proximities originally form an N $\times$ N matrix. After a tree is built, both training and test data are pulled down the tree. At the end, the proximities get normalized by dividing by the number of trees. Since the large data set could not fit an N $\times$ N matrix into fast memory, a modification reduced the required memory size to N $\times$ T where T stands for the number of trees in the forest. In order to speed up the computation-intensive scaling and iterative missing value replacement, the user is provided with the option of retaining only the nrnn largest proximities for each case. When the dataset is presented, the proximities of each case in the test set with each case in the training set can also be computed and compared. The amount of additional computing is moderate. The dataset contains thousands of data from which concentration of ozone is predicted. Thus Random Forest is useful in handling thousands of input variables without variable deletion. Hence Random Forest gives variable importance to each and every variable involved.

B. Classification and Regression Tree *algorithm*

Classification Regression Tree was introduced by Breiman et al. (1984) for classification or regression predictive modelling problems. It is often referred as 'Decision Tree' but now named as CART in modern software. It provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees. It is a binary tree that splits a node into two child nodes repeatedly beginning with the root that contains whole learning sample. Say for x being a nominal categorical variable of I categories, there are 2^{I-1}-1 possible splits for the predictor. If X is an ordinal categorical or continuous variable with K different values there are K − 1 different splits on X. At any node say t, the best split s is chosen to maximize a splitting criterion $\Delta i(s,t)$ [11]. There are 3 splitting criteria available.

2.1 Gini Criterion

The impurity measure at a node t is defined as

$$i(t) = \sum_{i,j} c(i|j)p(i|t)p(j|t) \tag{1}$$

It is the decrease of impurity given by

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{2}$$

where p_L and p_R are probabilities of sending case to left child node and right

$$p_L = p(t_L)/p(t) \tag{3}$$

child node where
And

$$p_R = p(t_R)/p(t) \tag{4}$$

$$\Delta i(s,t) = p_L p_R \sum_j [|p(j|t) - p(j|t)|]^2 \tag{5}$$

2.2 Twoing Criterion

CART does not require any special data preparation other than a good representation of the problem.

C. Multivariate Adaptive Regression Splines *algorithm*

It is a form of regression analysis developed by Friedman in 1991 [10] with the aim to predict dependent variable from set of independent variable. It is simpler than other models like random forest and neural network. It is seen as an extension of linear models that automatically models non linearity interaction between in variables. Mars is not affected by outliners. It produces model that can be written as an equation. It

models both classification and regression tasks. It accepts a large number of predictor and chooses important predictor variable. The extensive use of MARS model can be done for prediction as it has been done for concentration of ozone in this paper. It is used for prediction and classification problems (Islam et al. 2015; [4]). The details of the MARS model can be observed through a website by Salford Systems. Also, this regression is influenced by the recursive partitioning method for which any criteria can be chosen for the selection of basis function of multivariate spline. One of the advantage of the mars model is that MARS can reduce the outliers. The proposed MARS forms the model with the use of two sided truncated functions of the predictor x which has the form below.

$$(\mathbf{x}-\mathbf{t})_+ = \begin{Bmatrix} \mathbf{x}-\mathbf{t}, & \mathbf{x} > \mathbf{t} \\ \mathbf{0}, & \mathbf{otherwise} \end{Bmatrix} \tag{6}$$

Equation (6) works as a basis function for linear and non-linear functions. Also, this Eq. (6) works to approximate any function $f(x)$. In MARS, let us assume that the dependent variable (output) is y and number of terms is M. The output variable can be represented as following,

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\beta}_0 + \sum_{\boldsymbol{m}-1}^{\boldsymbol{M}} \boldsymbol{\beta}_{\boldsymbol{m}} \boldsymbol{H}_{\boldsymbol{km}}(\boldsymbol{x}_{\boldsymbol{v}}(\boldsymbol{k}, \boldsymbol{m})) \tag{7}$$

In the Eq. (7) MARS works over M term. The terms $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_{\boldsymbol{m}}$ are the parameter. Hinge function i.e. the H can be written as the following equation,

$$H_{km}\left(x_{v(k,m)}\right) = \prod_{k} -1^{k} h_{km} \tag{8}$$

In the above Eq. (8) the product of k^{th} of the m^{th} term is given as, $x_{v(k,\ m)}$

The value of K = 1 and K = 2 gives additive and pairwise interaction respectively. For this work, the opted value of K is 2.

3 Experiment

The experiment to predict concentration of ozone is carried out by using software named 'Salford Predictor Modeller 7.0' founded in 1983 which supports the three techniques Random Forest, Multivariate Adaptive Regression Splines and Classification and Regression Tree.

A. DATASET

This dataset is given by Saverio De Vito in 2006 from UCI dataset containing the response of multisensor device deployed on the field in an Italian City. It has 9358 instances with 15 attributes which was recorded from March (2004) to February (2005). The attributes are Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NO_x) and Nitrogen Dioxide

(NO_2) and were provided by a co-located reference certified analyser. The attributes included Date, Time, True hourly averaged concentration CO in mg/m^3, PT08.S1 (tin oxide) hourly averaged sensor response, True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3, True hourly averaged Benzene concentration in microg/m^3, PT08.S2 (titania) hourly averaged sensor response, True hourly averaged NOx concentration in ppb, PT08.S3 (tungsten oxide) hourly averaged sensor response, True hourly averaged NO_2 concentration in microg/m^3, PT08.S4 (tungsten oxide) hourly averaged sensor response, PT08.S5 (indium oxide) hourly averaged sensor response, Temperature in Â°C, Relative Humidity (%), AH Absolute Humidity [7].

B. Air Quality Prediction Steps

This describes the various steps taken for prediction by Salford Modeller.

Step 1: The database is opened in the software as it supports all type of file.

Step 2: The model is designed by selecting the predictor. A total of 12 predictors are selected for MARS/Random Forest/CART for this dataset. Date and time are not chosen as they don't have any effect. PT08_S5_O3_ is set as the target in all the cases.

Step 3: The analysis method is selected as MARS/Random Forest/CART with analysis type being 'regression' in all three cases.

Step 4: It's time to separate the dataset as learning set and test set. This is done by selecting Fraction of cases selected at random for testing by assigning any value. Remember the values are in terms of percent. Here we put the test set as 0.30.

Step 5: Now the model is started and resulting graph pops up showing the information required for future prediction of target variable. It also provides summary for all other details which is discussed in the next section.

4 Result and Discussion

In this section we compare the results given for the target variable PT08_S5_O3_ through Random Forest, Multivariate Adaptive Regression Splines and Classification and Regression Tree by Salford Predictive Modeller. Out of 15 attributes, 12 are being used for used as predictors while 1 is selected as the targeted variable and the targeted variable being PT08.S5 (O3). 30% of the 9358 instances are selected for test case while rest go in for the learn case. This paper contains the graphical representation of the learn and test value, summary of important terms, list of variable importance by the three models used. On applying Multivariate Adaptive Regression Splines, we get Fig. 1 which shows the graph shows result of learn and test case where the Y-axis represents MSE with an interval of 50,000 and X-axis representing Basis functions which was taken as 15 initially. From the graph conclusion can be made that there are least error as both the learn and test cases are same. Initially the MSE value starts from 200,000 drops till 5000 and gradually becomes constant.

There are several important parameters that give the model error measure. These important parameters have been listed in Table 1 showing their value for both learn and test. The variables include RMSE, MSE, GCV, MAD, MRAD, SSY, SSE, R^2, R^2 Norm, GCV R-Sr. Out of 15 attributes, 12 were set as predictors but after the regression

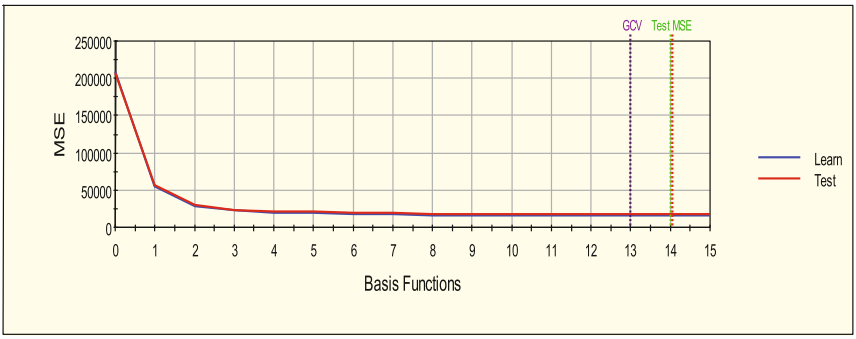

Fig. 1. MSE vs. basis functions

model was prepared it was deducted that only 8 variables were important for prediction of PT08.S5 (O3). The most important variable was found to be PT08_S2_NMHC. The scores of all the variables are given in decreasing order of their importance in predicting the predictor in Table 2. The number of basis function was set as 15 initially. The model assigns special variables to make a new equation to cover all points of non-linearity. These variables are termed as basis variables. The model is a weighted sum of basis function. Each basis function takes one of the following forms

Table 1. MARS result for learn and test

Name	Learn	Test
RMSE	126.46071	131.97553
MSE	15,992.31082	17,417.54040
GCV	16,138.69982	n/a
MAD	97.47557	100.28940
MRAD	0.11444	0.11721
SSY	1,376,184,805.13191	57,713,8537.64731
SSE	105,309,366.83668	48,281,422.13143
R^2	0.92348	0.91634
R^2 Norm	0.92348	0.91634
GCV R-Sq	0.92280	n/a

(1) Constant. Only one term i.e. the intercept
(2) Hinge function
(3) Product of 2 or more Hinge function. Table 3 consists all the basis function and their combination to give the final equation of Y.

Random Forest was started by setting the number of trees to be built as 500 and number of predictors for each node as 3. The frequency of the report was set 10 along with parent minimum case as 2. The elapsed time was nearly 45 s for creating the trees.

Table 2. Variable importance in MARS

Variable	Score
PT08_S2_NMHC_	100.00
PT08_S1_CO_	96.61
RH	55.18
NO2_GT_	46.47
PT08_S3_NOX_	44.38
PT08_S4_NO2_	18.28
T	16.76
NOX_GT_	4.64

Table 3. MARS basis functions

BF6	max(0, 11 - NOX_GT_)
BF7	max(0, PT08_S2_NMHC_ - 609.25)
BF8	max(0, 609.25 - PT08_S2_NMHC_)
BF9	max(0, T - 2.1)
BF11	max(0, RH - 63.15)
BF12	max(0, 63.15 - RH)
BF13	max(0, NO2_GT_ - 98)
BF14	max(0, 98 - NO2_GT_)
BF15	max(0, PT08_S4_NO2_ + 200)
Y	769.987 + 0.584446 * BF1 + 0.860364 * BF2 − 0.0722947 * BF3 + 0.317973 * BF4 + 0.0305292 * BF5 − 0.42804 * BF6 + 0.747153 * BF7 − 1.56998 * BF8 − 2.74185 * BF9 − 5.13236 * BF11 − 3.58688 * BF12 + 1.15626 * BF13 + 0.359236 * BF14 − 0.102166 * BF15

Separate graphs were obtained showing the comparison of train as in Fig. 2 and test cases as in Fig. 3 with 500 trees having maximum terminal node of 3293. Observing the curve in both the cases, a clear difference can be seen in the curve. Both start with MSE 40,000 but the train set is less steep when compared with test set. Train set shows a turn at 16th tree (18,502.625) while test set shows a turn at 8th tree (16,911.168). Both the graph for train and test are given in Figs. 2 and 3 respectively.

There are several important parameters that give the model error measure. These important parameters have been listed in Table 4 showing their value for both learn and test. The variables include RMSE, MSE, MAD, MRAD, SSY, SSE, R^2, R^2 Norm. Unlike MARS in random forest all 12 variables have their own importance and show their contribution for building trees. Importance of each variable has been shown in Table 5.

Classification and Regression Tree model also leads in building up of trees. It gives a graph where the Y-axis shows Relative Error while X-axis shows the number of nodes. The graph value shows the relative error value as 0.083 at 150^{th} node. So by examining the graph directly we can get the relative error of test from the train set as shown in Fig. 4.

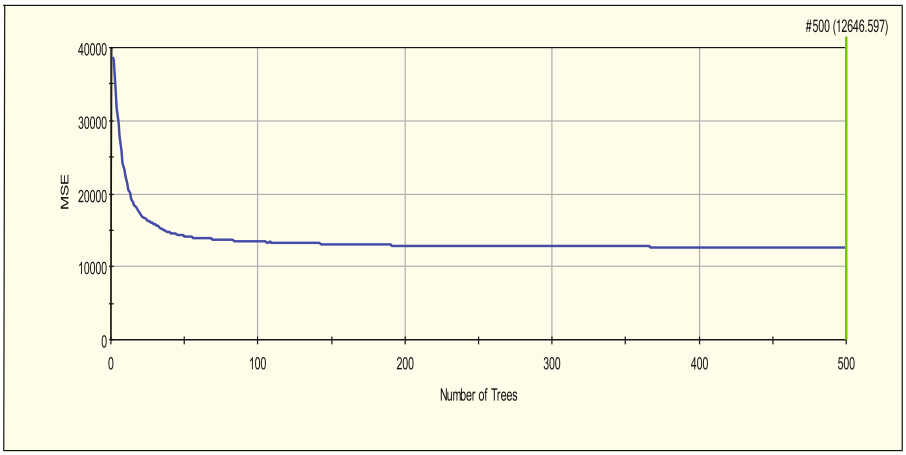

Fig. 2. Train set plot in random forest

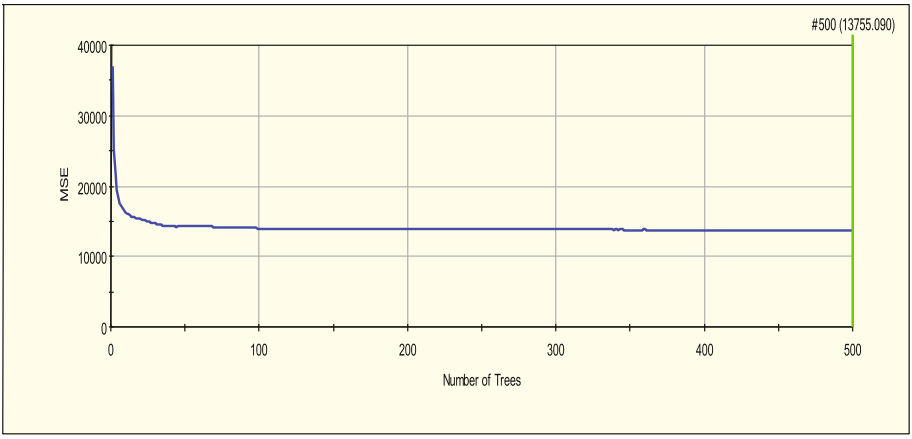

Fig. 3. Test set plot in random forest

Table 4. Random forest result for learn and test

Name	OOB	Test
RMSE	112.45649	117.28208
MSE	12646.46110	13755.08634
MAD	84.08904	86.55029
MRAD	0.10388	0.10746
SSY	1,376,184,805.13192	577,138,537.64731
SSE	83,276,946.34398	38,129,099.33380
R^2	0.93949	0.93393
R^2 Norm	0.94676	0.94078

Table 5. Variable importance in random forest

Variable	Score
PT08_S1_CO_	100.0000
C6H6_GT_	59.0625
PT08_S2_NMHC_	57.9778
PT08_S3_NOX_	28.8258
PT08_S4_NO2_	19.8222
NOX_GT_	17.0169
RH	13.3625
T	12.6999
NO2_GT_	11.0866
AH	10.2157
CO_GT_	5.6714
NMHC_GT_	1.0818

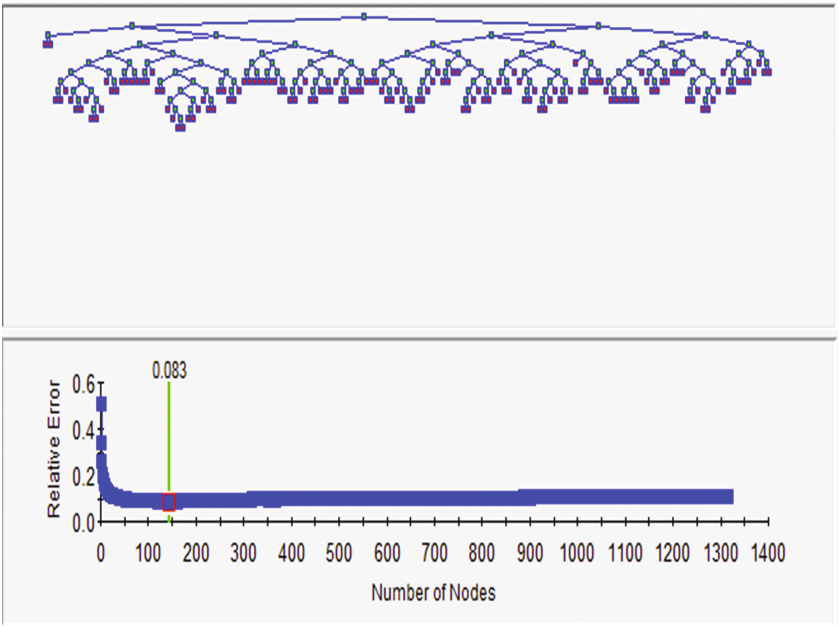

Fig. 4. Relative error through CART

There are several important parameters that give the model error measure. These important parameters have been listed in Table 6 showing their value for both learn and test. The variables include RMSE, MSE, MAD, MRAD, SSY, SSE, R^2, R^2 Norm, AIC, AICc, BIC, Relative Error. CART also requires the use all 12 predictor variables just as Random Forest does. Table 7 lists the variable according to their importance.

Table 6. Random forest result for learn and test

Name	Learn	Test
RMSE	105.96244	131.25001
MSE	11,228.03897	17,226.56562
MAD	80.66569	96.96731
MRAD	0.08853	0.10464
SSY	1,376,184,805.13191	577,138,537.64731
SSE	73,936,636.62739	47,752,039.90055
R^2	0.94627	0.91726
R^2 Norm	0.94627	0.91736
AIC	61,436.82555	27,062.66453
AICc	61,436.87303	27,062.77762
BIC	61,518.33615	27,133.79242
Relative Error	0.05373	0.08274

Table 7. Variable importance in CART

Variable	Score
PT08_S1_CO_	100.0000
PT08_S2_NMHC_	97.3907
C6H6_GT_	97.3598
PT08_S3_NOX_	60.6814
PT08_S4_NO2_	53.5883
CO_GT_	44.9134
T	22.8821
NOX_GT_	7.2299
NO2_GT_	3.2165
AH	1.9955
RH	1.4960
NMHC_GT_	0.2262

5 Conclusion

In this paper we have proposed to show the prediction of ozone concentration by using three regression model. By keeping the train and test in the ratio 7:3 we compare the result from all three cases. Evaluation of the prediction models indicates that the Multivariate Adaptive Regression Splines model describes the dataset better and has achieved significantly better prediction accuracy as compared to the Random Forest and Classification and Regression Tree. Multivariate Adaptive Regression Splines gives the result by considering less variables as compared to other two. It evaluates on basis of 8 variables while other two require all variables. Moreover, Random Forest takes a little more time for building the tree as the elapsed time was calculated to 45 s in this case. PT08_S2_NMHC_ is the most important variable as given by Multivariate Adaptive Regression Splines while PT08_S1_CO_ is most important variable as given

by Random Forest and Classification and Regression Tree. Observing all the graphs Multivariate Adaptive Regression Splines gives the closest curve of both train and test set when compared. It can be concluded that multivariate adaptive regression splines can be a valuable tool in predicting ozone for future.

References

1. Ozer, P., Laghdaf, M.B.O.M., Lemine, S.O.M., Gassan, J.: Estimation of air quality degradation due to Saharan dust at Nouakchott, Mauritania, from horizontal visibility data. Water Air Soil Pollut. **178**(1–4), 79 (2007)
2. Zhang, W.Y., Han, T.T., Zhao, Z.B., Zhang, J., Wang, Y.F.: The prediction of surface layer ozone concentration using an improved AR Model. In: International Conference of Information Technology, Computer Engineering and Management Sciences, Nanjing, Jiangsu, pp. 72–75 (2011)
3. Assessment for Decision-Makers: Scientific Assessment of Ozone Depletion: 2014, World Meteorological Organization, Global Ozone Research and Monitoring Project—Report No. 56, Geneva, Switzerland (2014)
4. Zhang, C., Yuan, D.: Fast fine-grained air quality index level prediction using random forest algorithm on cluster computing of spark. In: 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing and 2015 IEEE 12th International Conference on Autonomic and Trusted Computing and 2015 IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), Beijing, pp. 929–934 (2015)
5. Wang, Y., Yue, W.: Target data association in communication constrained environment using CART: compressed adaptive reference topology. In: 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), Beijing, China, pp. 333–338 (2016)
6. Birgersson, M., Hansson, G., Franke, U.: Data integration using machine learning. In: 2016 IEEE 20th International Enterprise Distributed Object Computing Workshop (EDOCW), Vienna, Austria, pp. 1–10 (2016)
7. De Vito, S., Massera, E., Piga, M., Martinotto, L., Di Francia, G.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sens. Actuators B Chem. **129**(2), 750–757 (2008)
8. De Vito, S., Piga, M., Martinotto, L., Di Francia, G.: CO, NO_2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. Sens. Actuators B Chem. **143**(1), 182–191 (2009). ISSN: 0925-4005
9. Hui, T.S., Rahman, S.A., Labadin, J.: Comparison between multiple regression and multivariate adaptive regression splines for predicting CO_2 emissions in ASEAN countries. In: 2013 8th International Conference on Information Technology in Asia (CITA), Kota Samarahan, pp. 1–5 (2013)
10. Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**, 1–67 (1991)
11. Breiman, L., Friedman, J.H., Olshen, R., Stone, C.J.: Classification and Regression Tree. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California (1984)
12. Abbasi, R., Moradi, M.H., Molaeezadeh, S.F.: Long-term prediction of blood pressure time series using multiple fuzzy functions. In: 2014 21th Iranian Conference on Biomedical Engineering (ICBME), Tehran pp. 124–127 (2014)

13. Roy, S.S., Viswanatham, V.M., Krishna, P.V., Saraf, N., Gupta, A., Mishra, R.: Applicability of rough set technique for data investigation and optimization of intrusion detection system. In: International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness, pp. 479–484. Springer, Berlin Heidelberg, January 2013
14. Roy, S.S., Viswanatham, V.M., Krishna, P.V.: Spam detection using hybrid model of rough set and decorate ensemble. Int. J. Comput. Syst. Eng. **2**(3), 139–147 (2016)
15. Roy, S.S., Viswanatham, V.M.: Classifying spam emails using artificial intelligent techniques. Int. J. Eng. Res. Afr. **22**, 152–161 (2016)
16. Roy, S.S., Mittal, D., Basu, A., Abraham, A.: Stock market forecasting using LASSO linear regression model. In: Afro-European Conference for Industrial Advancement, pp. 371–381. Springer International Publishing (2015)
17. Basu, A., Roy, S.S., Abraham, A.: A Novel diagnostic approach based on support vector machine with linear kernel for classifying the Erythemato-Squamous disease. In: 2015 International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 343–347. IEEE, February 2015
18. Mittal, D., Gaurav, D., Roy, S.S.: An effective hybridized classifier for breast cancer diagnosis. In: 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), pp. 1026–1031. IEEE, July 2015
19. Roy, S.S., Gupta, A., Sinha, A., Ramesh, R.: Cancer data investigation using variable precision Rough set with flexible classification. In: Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology, pp. 472–475. ACM, October 2012
20. Samui, P., Roy, S.S., Kurup, P., Dalkiliç, Y.: Modeling of Seismic Liquefaction Data Using Extreme Learning Machine, Nova publishers, Natural Disaster Research, Prediction and Mitigation, pp. 6x9-(NBC-R). ISBN: 978-1-53610-356-4
21. Roy, S.S., Krishna, P.V., Yenduri, S.: Analyzing intrusion detection system: an ensemble based stacking approach. In: 2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 000307–000309 (2014)
22. Popescu-Bodorin, N., Balas, V.E., Motoc, I.M.: Iris codes classification using discriminant and witness directions. arXiv preprint arXiv:1110.6483 (2011)
23. Azar, A.T., Balas, V.E., Olariu, T.: Artificial neural network for accurate prediction of post-dialysis urea rebound. In: 2010 4th International Workshop on Soft Computing Applications (SOFA), pp. 165–170. IEEE, July 2010
24. Roopaei, M., Balas, V.E.: Adaptive gain sliding mode control in uncertain MIMO systems. In: 3rd International Workshop on Soft Computing Applications, SOFA 2009, pp. 77–82. IEEE, July 2009
25. Ponalagusamy, R., Senthilkumar, S.: Investigation on time-multiplexing cellular neural network simulation by RKAHeM (4, 4) technique. Int. J. Adv. Intell. Paradig. **3**(1), 43–66 (2011)