Predicting the Air Quality Index in the National Capital Region of India using Statistical Learning Techniques

Hardik Shah
Manan Shah
Nikhil Soni

Purdue University

May 4, 2018

This page is intentionally left blank.

Contents

This page is intentionally left blank.

1 Abstract

The occurrences of smog in the National Capital Region of India has gone up, the concentration of PM 2.5 is through the roof [1]. It has been reported as one of the worst levels of air quality in Delhi since 1999 and probably one of the worst air quality in the world [2]. Low visibility has resulted in accidents across the city, notably a 24 vehicle pile-up on the Yamuna Expressway [3] . Deteriorating air quality has far reaching effects on health such as multiple sclerosis and lung cancer [4]. It also led to cancellation and delay of public transport, primarily trains and flights, causing much hindrance to the people [5]. The primary sources of smoke are stubble burning, lit garbage, road dust, power plants, factories, vehicles and a lack of vegetation across the city. It has been observed for the past five years that surface concentration levels of PM2.5 dramatically increase during the winter months in the national capital region of India. Major reasons for this trend include stubble burning during the harvest months in winter. This project aims to understand the increase in PM2.5 concentration using climatic variables using statistical learning techniques.

On the eve of new year 2018, the air quality was found to be the worst when compared to the last 2 years (2016 and 2017). The Air Quality Index during the period from 25 Dec'17 to 3rd Jan'18 remained roughly 50 notch points higher as compared to the last 2 years [6]. The lead pollutant was identified as PM2.5 [6] . Thus, it is very essential to understand the reasons behind the poor air quality index in Delhi and predict the air quality index using statistical learning techniques.

Our study has statistically established the association of environmental conditions such as rain, fog, temperature, relative humidity etc, with concentration of PM2.5. The inclusion of such predictors paves the way for implementing the resulting model in other regions of India as well.

2 Background

Much of the work in predicting PM2.5 is done in terms of multiple linear regression between the predictors such as aerosol optical depth, height of planetary boundary layer and temperature and PM 2.5. This project derives its inspiration from the work of Christopher et al. (2009).[4]. Gupta et al. have shown that aerosol optical depth has a strong correlation with PM2.5. A thorough description of the algorithm for retrieval of Aerosol Optical Thickness (AOT) from the NASA MODIS instruments has been provided by Gupta et al.[7]. The inaccuracies produced by MODIS Dark Target (MDT) algorithm AOT retrievals has been addressed and a surface parametrization valid for cities across the globe has been suggested by Gupta et al.[7]. These suggestions will lead to more accurate AOT retrievals which can be utilized for predicting the surface concentration level of PM2.5 in the national capital region of India.

The correlation between AOT and PM2.5 concentration is further improved by including environmental predictors such as temperature, humidity etc. The main issue with their approach is the assumption of a parametric relationship between the response and the predictors. Assuming a parametric relationship has several advantages such as an interpretable and simple model. However, it can lead to high bias and variance when predicting the response.

Therefore, we intend to apply the non-parametric approach to find an associative relationship between the predictors and the response.

Another reason for making such a model is to come up with a cost effective product, which can use publicly available data to give reasonably good predictions of any inhabited area in general. For instance, many of the industrial towns in India do not have infrastructure in place to predict the air quality locally. However, the air quality should be kept in check to plan for side affects of industrial activity on the socio-economic conditions of the said region. This study helps better understand the relationship of environment variables with surface concentration levels of PM2.5.

The scope of our study is very localized. During the course of our literature survey we did not encounter any prior work(s) which accounted for geo-spatial variable(s), environmental variables, natural events (eg. fog, thunderstorm etc.) and green cover. Hence, the extent of the literature survey is limited to identifying important input variables such as Aerosol_Type_Land (AOT) and the theoretical effect of input variables on the response. We do not have any pre-formed hypotheses regarding the predictors, this study is an exploratory exercise.

3 Data

We set the time span of data from 01/01/2015 to 01/01/2018 to get a more complete dataset as a larger time period would only make the proportion of missing values higher. Many stations were not operational until recently, which explains why a majority of missing data exists before 2015 and hence, we excluded that part.

To build this particular data set, datasets from 18 different weather stations spread across New Delhi were downloaded and collated according to dates and columns.

3.1 Response Variable

We have identified the following as our response variable:

$$PM2.5\ (\mu gm^{-3})$$

Particulate matter with aerodynamic diameters less than $2.5\mu m$ ($PM2.5$ *or* $PM_{2.5}$) can cause respiratory and lung diseases and even premature death[4].

Completeness	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
90.42%	0.00	0.00	0.00	71.14	114.56	3897.31

Table 1: Summary Statistics for the Response variable $PM2.5$

Please refer the section, *"Data Cleaning" (sec.3.3, pg.3)* to read about the data cleaning operations in detail.

3.2 Input Variables

Acronyms	***Description***	***Value type***	***Source***
Station	Abbreviation of the station name	Factor with 18 levels	CPCB
WS	Wind speed in m/s	Numeric	CPCB
WD	Wind direction in degrees	Numeric	CPCB
AT	Ambient temperature in degree C	Numeric	CPCB
RH	Relative Humidity in %	Numeric	CPCB
SR	Solar radiation in Wm^{-2}	Numeric	CPCB
BP	Barometric pressure in mmHg	Numeric	CPCB
Aerosol_Type_Land	Aerosol optical depth (dimensionless)	Numeric	NASA MODIS
TempN	Temperature in degree C	Numeric	AccuWeather
Humid	Humidity in %	Numeric	AccuWeather
Precip	Precipitation	Numeric	AccuWeather
Events	Natural phenomena	Factor with 10 levels	AccuWeather
GC	Green cover around the station	Factor with 4 levels	Delhi.gov

Table 2: Data description

CPCB: Central Pollution Control Board of India is a statutory organization under the Ministry of Environment, Forest and Climate Change, Government of India.

Most of the input variables listed in the table are self-explanatory. Some of the uncommon input variables are listed below with a brief description:

- **Aerosol Optical Thickness** is a measure of the extinction of the solar beam by dust and haze. In other words, particles in the atmosphere (dust, smoke, pollution) can block sunlight by absorbing or by scattering light. AOT tells us how much direct sunlight is prevented from reaching the ground by

these aerosol particles. It is a dimensionless number that is related to the amount of aerosol in the vertical column of atmosphere over the observed location.

- **Wind Speed** affects the concentration of precursors in a swath of land, thus wind speed affects the presence of precursors horizontally, which cannot be measured by a satellite.
- **Ambient Temperature** can enhance the photochemical reactions in the atmosphere and hence production of PM2.5 particles. Temperature inversion can also reduce the vertical mixing and therefore increase chemical concentration of precursors[4].
- **Relative Humidity** The absorption of water by the deposited particles may increase the PM2.5 reading of the gauge, that suggests RH affects the measurement of PM2.5[8].
- **Solar Radiation** indicates the available sunlight for photochemical reactions due to clouds and season[4].
- **Barometric Pressure** indicates the wind conditions, as in high pressure is related to low winds and vice versa. Therefore, the transportation of PM2.5 is related to Barometric Pressure.
- **Green-cover**[9] or vegetation in and around the city in terms of density, sorted as a categorical variable with 4 levels.
  - 4: Very Dense Forest
  - 3: Moderately Dense Forest
  - 2: Open Forest
  - 1: Scrub
  - 0: Non-Forest
- **TempN & AT**:These two variables are technically different in the sense that AT measures the ambient temperature for the region surrounding the weather station. And, TempN is the average temperature of the city over a large region, such as for the entire National Capital Region, and hence allows us to pick up on the local temperature which might affect weather across a larger region.
- **Stations**: Refer table5, pg.27 which lists the names of ground based monitoring stations corresponding with station codes.
- **Events**: For full description of all the events accounted in this study please refer to table 6, pg.27.

3.3 Data Cleaning

1. For all the variables (dependent and independent), legible values were identified based on historically recorded values. By legible we mean historically accurate and observed values. These values do include outliers and leverage points (if any). For instance, the response PM2.5 has the maximum value as $3897.31 \mu g^{-3}$, this value is recorded by one of the 18 stations considered in our study.
2. Unrealistic values eg. negative values in WS, SR etc. were replaced with NA. Because, these values were obviously bad measurements.
3. For the data-frame in R, we dropped rows with NA values in all the columns.

4 Exploratory Data Analysis

From fig.1, fig.2 and fig.3, we see that the response variable is highly skewed. One of the primary reasons for the skewness is the large number of missing values in the response variable PM2.5. Also, the response is not normally distributed and has outliers.

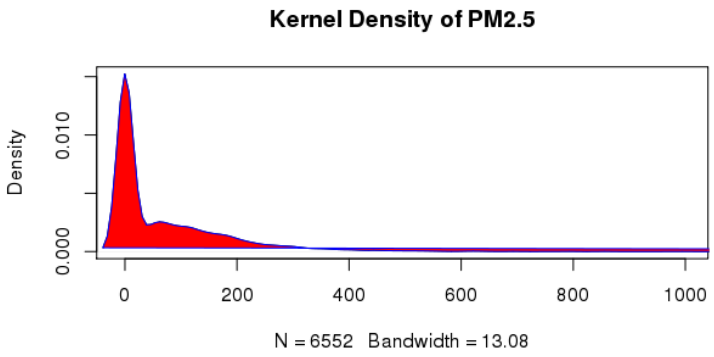

Figure 1: Kernel Distribution of PM2.5

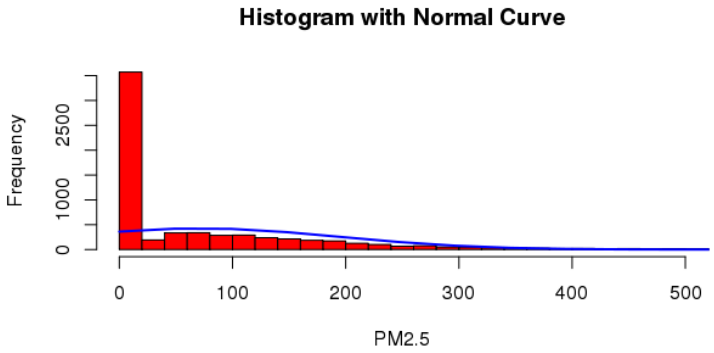

Figure 2: Histogram of PM2.5

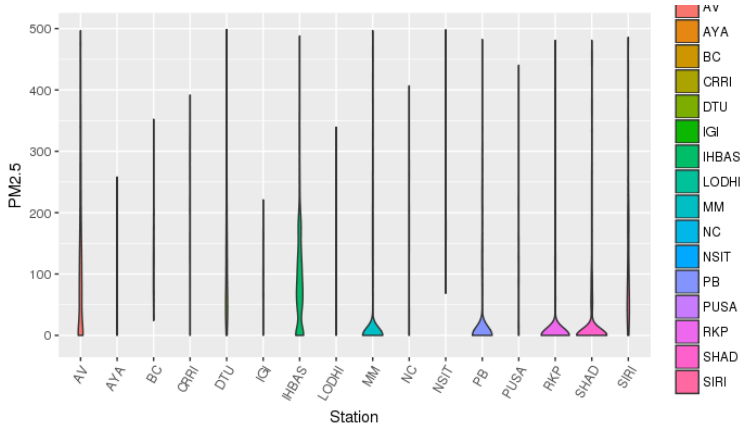

Figure 3: Distribution of PM2.5 according to the station

4.1 Correlation Plot

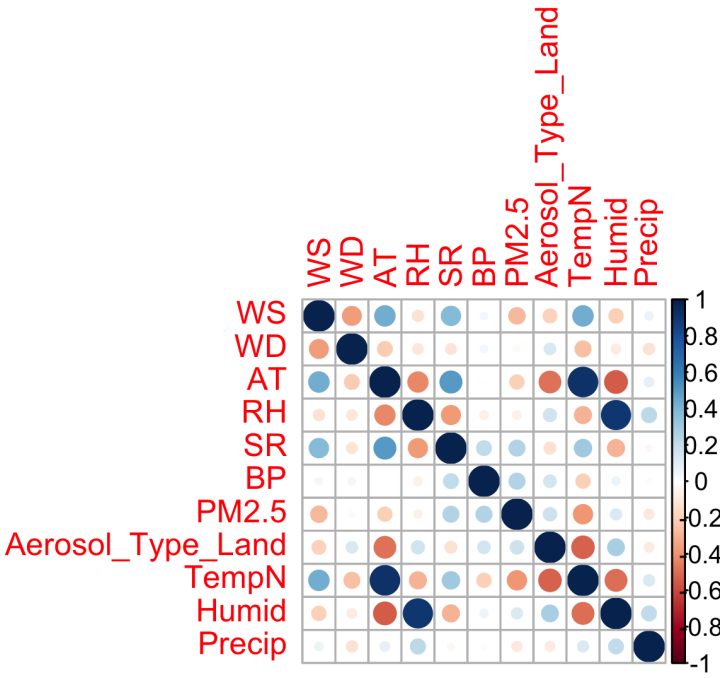

Figure 4: Correlation Plot

The correlation plot was generated for pairwise complete values. We see that some variables such as TempN and AT and Humid and RH are highly correlated, which makes sense, but no variables are highly correlated with PM2.5 and hence, we can use all the variables without having to remove them or change them further. We choose to include all variables in the final model despite correlations.

4.2 Principle Components Analysis

PCA is generally carried out in order to reduce dimensions by obtaining a direction in which to project the entire data onto the direction so that it maximizes the variance explained in the data. In the case of regression, we use this to assess the variables which create maximum variance amongst the data set from the selected variables. PCA analysis also gives us the variable importance plot which tells us the amount of variance explained by that principal component. If the dimensions are reduced significantly, such that very few principal components explain more than 80 percent of the data, it would greatly help in analysis and prediction as the model would become much more simpler.

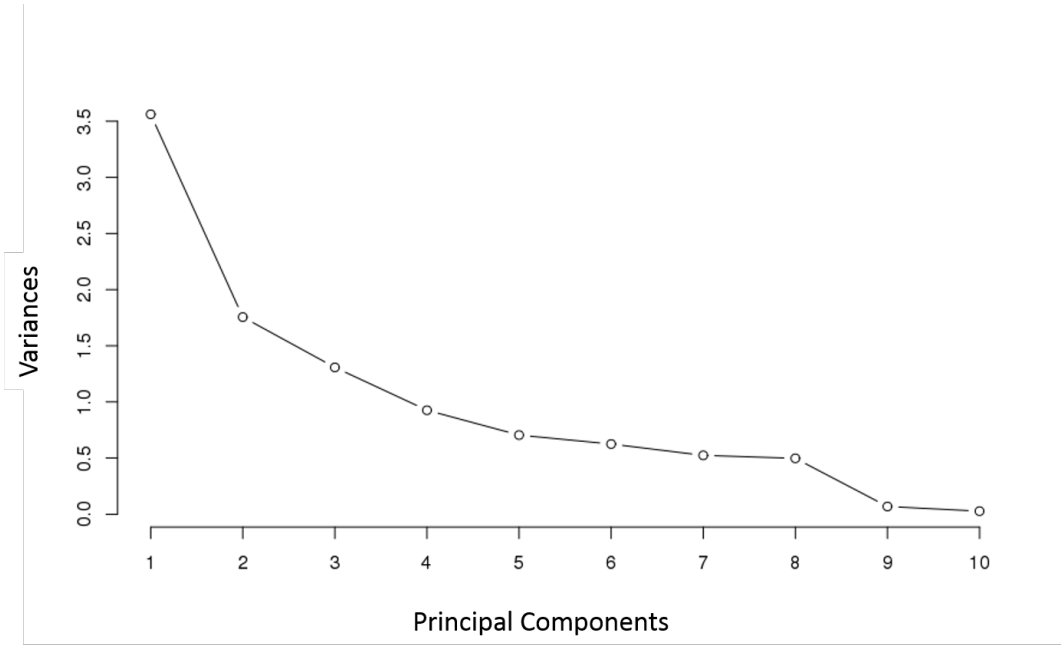

Figure 5: Distribution of variances by PC's

Based on the plots above, we see that the variance is explained by 4 principal components for just 80 percent of the data. Hence, this proves that using PCA does not help predictions in our case since we still end up needing 4 PC's instead of using the original 13. It however, does explain the important variables which explain majority of the variance in the data set.

Also, based on the biplot below, we see the principal components and the distribution of station points around these.

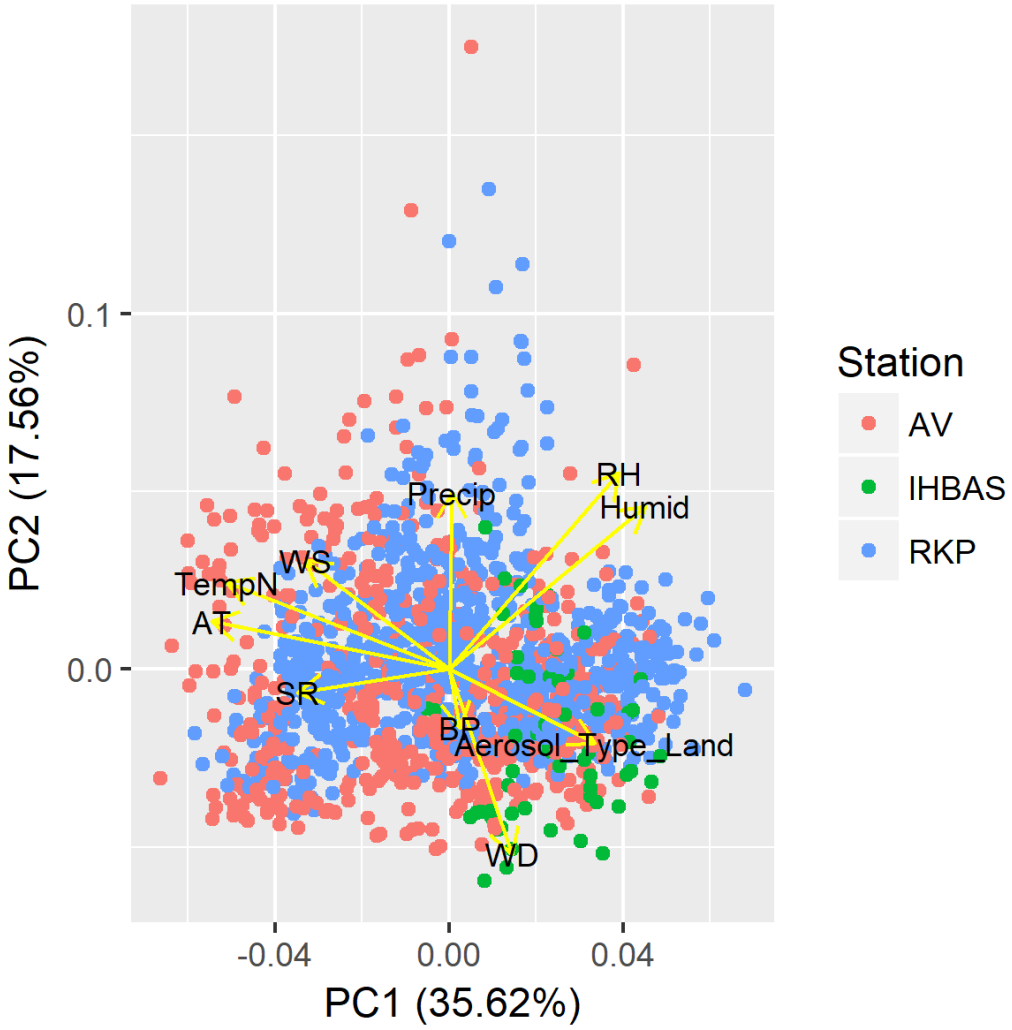

Figure 6: PCA Bi-plot

This Biplot is essential in giving us information related to the three stations which end up explaining most of the variance in the response variable. Clearly, solar radiation and wind direction are important in explaining the variance in the data and strongly so. We can see a cluster of points forming for the three separate stations.The readings of station Aya Nagar are more towards the positive axis of solar radiation, indicating a high but uniform distribution of solar radiation at this station. Also, stations RK Puram and IHBAS recieve little solar radiation which explains the differences in climate conditions at these stations. The centering of cluster to the origin shows a lot of readings shows normal behaviour in terms of AOT, humidity, precipitation etc. and there a few outlier readings which show an abnormally large deviation from average at for certain days. There are some points which are away from the origin, which show a that they have high values of solar radiation and low values of wind speed. More points are clustered towards the center indicating wind speed and solar radiation as primary indicators of affecting variables. As informative as this plot is at confirming our observations and indicating the variables which explain the most variance, we choose not to reduce the model down to just 4 variables since we already use just 13 variables. We choose to go ahead and use all our variables in our model in order to not skip out important variables.

5 Methodology and Models

5.1 General Linear Model

General linear models[10] assume a parametric relationship between the response and the predictors. These types of models are highly affected by phenomena such as non-linearity of the response-predictor relationships ,collinearity, heteroscedasticity, outliers and collinearity[11]. For our model, certain predictors such as but not limited to AT, TempN and Humidity have high correlations. Also, the response has outliers. Therefore, this model will not be representative of the actual underlying relationship function.

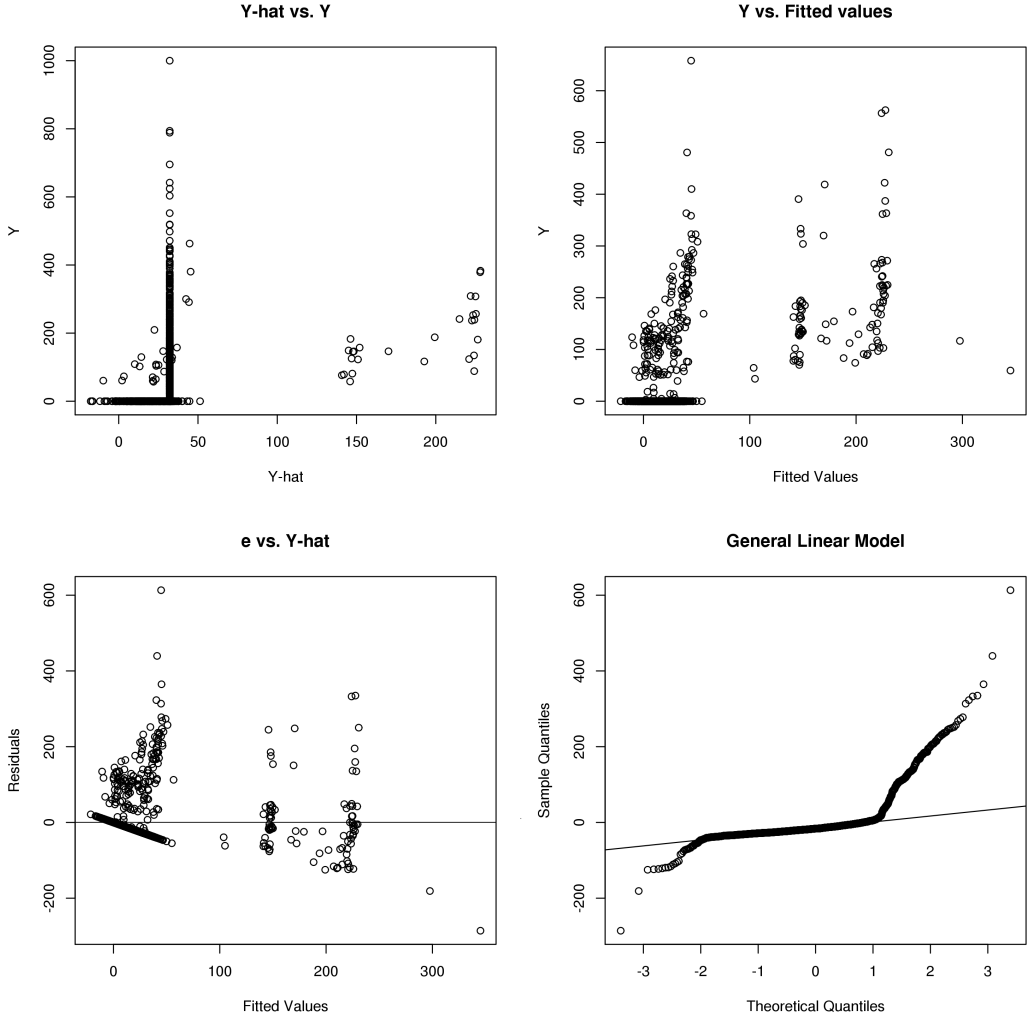

Figure 7: GLM Diagnostics and Residual Analysis

The diagnostics show that the model is not a good fit. The Y-hat vs. Y and Y vs. Fitted Values show that the predictions are not linear with respect to the actual values. This model does not account for the variance in the training dataset. Therefore, it is safe to conclude that this model is inadequate.

5.2 General Additive Model

Generalized additive models[12] do not assume any parametric relationship between the response and the predictors.

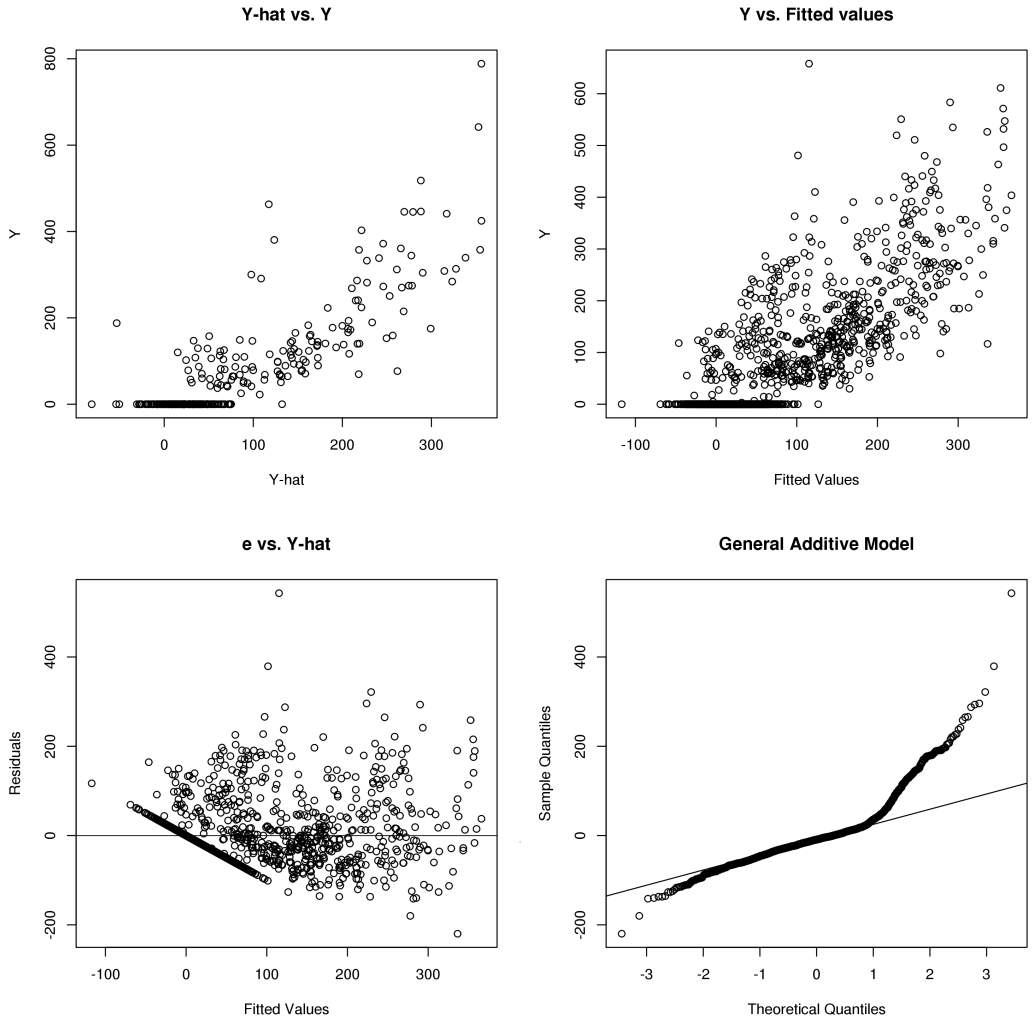

Figure 8: GAM Diagnostics and Residual Analysis

The QQplot for residuals indicate that the GAM model is unable to capture the behavior of the predictor variables. This has happened for the General Linear Model as well. This model is inadequate in fitting the independent variables with respect to the response.

To fit a GAM model, all the integer and numeric variables were smoothed and the unordered factors were kept as is. Those predictors which weren't significant either in the parametric ANOVA or the non-parametric ANOVA were dropped. Then we used *gam.scope* to implement various degrees of freedom for spline (s, smoother).

5.3 Bayesian Additive Regression Trees (BART)

We have fitted[13] the training data using the *"use_missing_data"* option in R. Following is our variable importance plot.

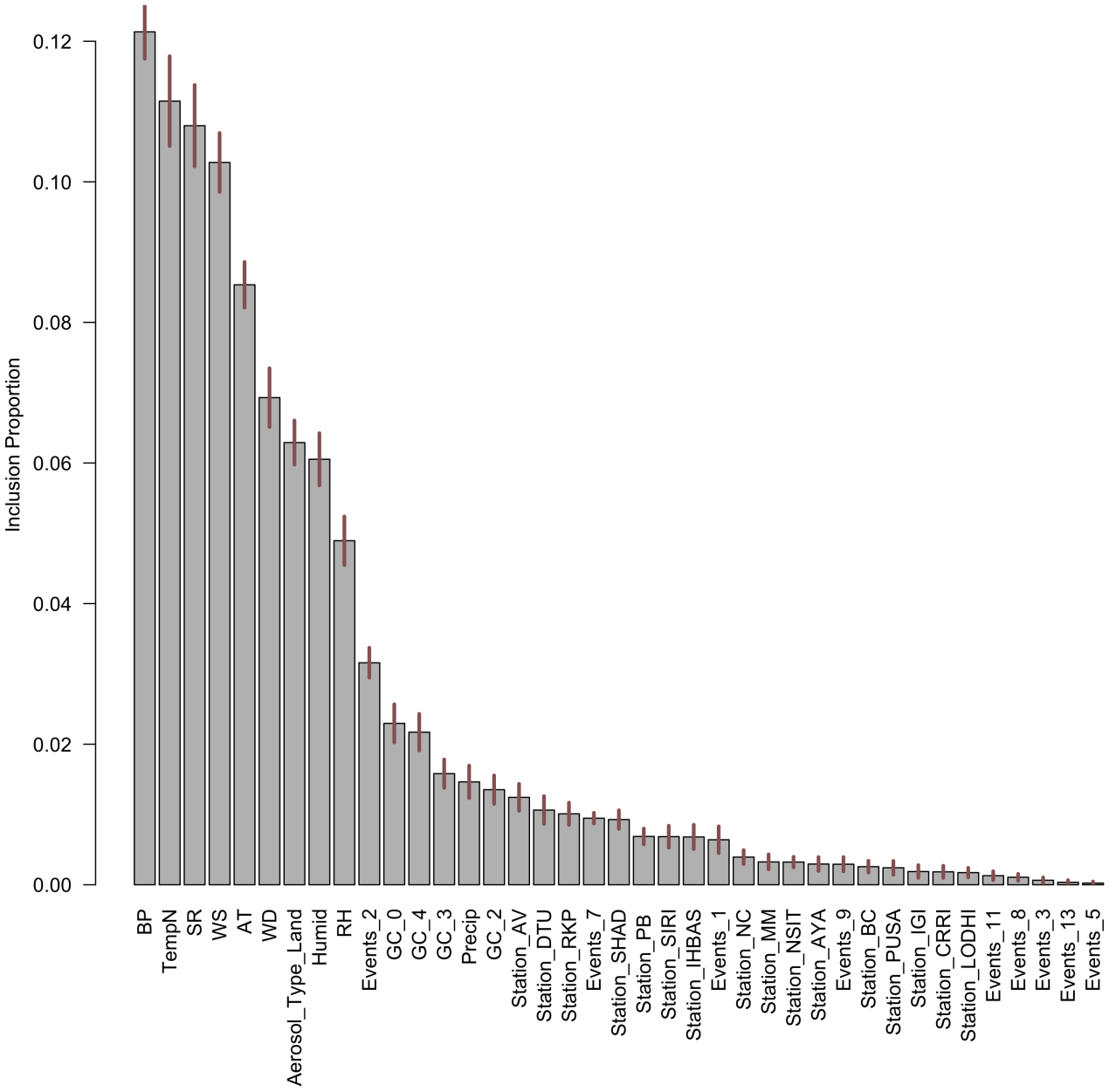

Figure 9: BART Variable Importance Plot

The variable importance plot gives some interesting insights into the working of the model. Among the top 10 predictors, Events_2 is one. Events_2 is Fog, and air quality in Delhi takes a nosedive in the winter months especially with foggy conditions which leads to smog. Also, temperature and temperature inversion and humidity are known to affect the levels of PM2.5 positively[4].

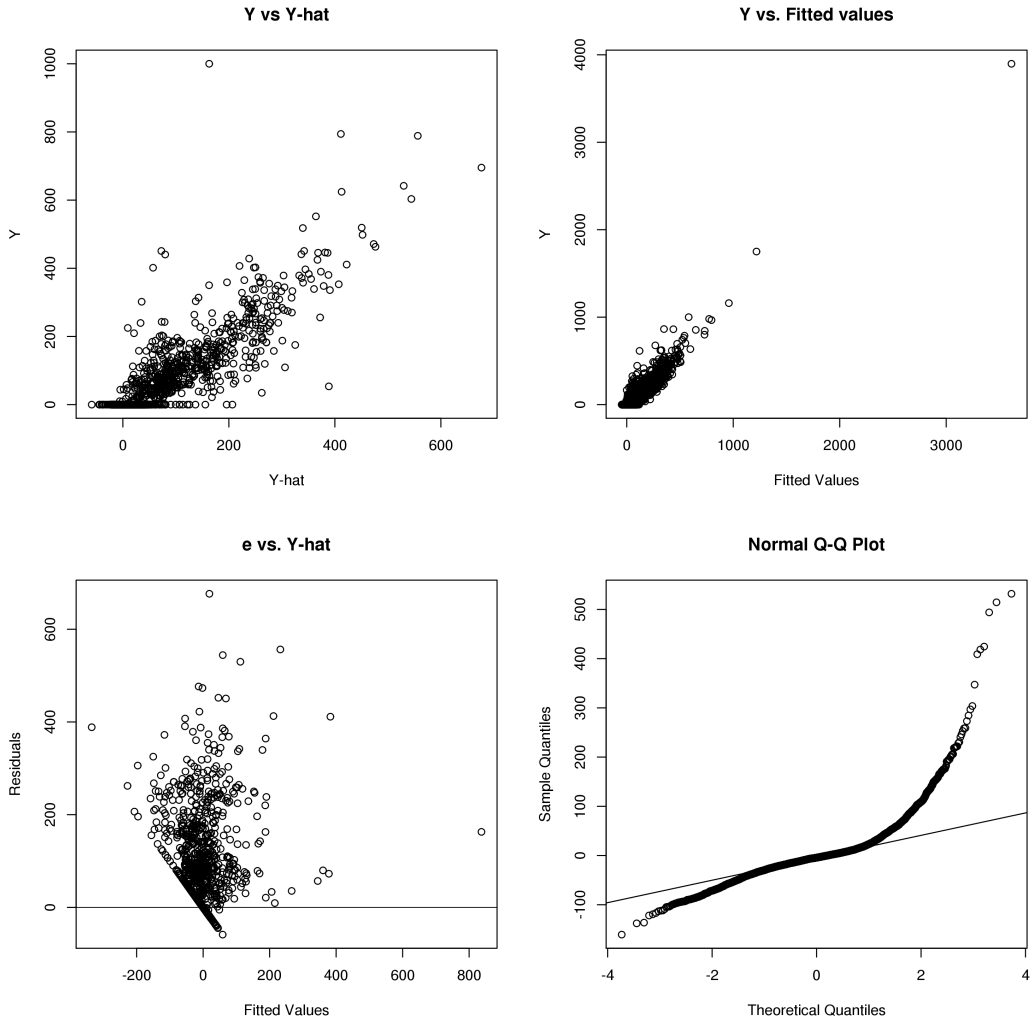

Figure 10: BART Diagnostics and Residual Analysis

Initially, we hypothesized that GC (Green Cover) may have an effect on PM2.5 however according to the variable importance plot, that does not seem to be the case.

The in-sample cross validated RMSE and the out of sample RMSE (refer table3, pg.20) are considerably good. Therefore, this model in general will give decent predictions. Looking at the Y vs. Fitted values, the fitted values with respect to the actual values seem good, since the outlier has been correctly predicted. However, the values are concentrated in the lower quadrant. Also, the majority of residuals are positive with respect to the fitted values, which tells us that we are over-estimating PM2.5 concentration. Lastly, from the QQplot, it is visible that the normality assumption is being violated. The QQplot has very prominent tails and is not able to capture the behavior of the residuals properly.

5.4 Classification & Regression Trees (CART)

Tree based[14] segment the response into regions by splitting on variable values and then predicts the value of an observation, using the mean or mode of the points which lie in that region in which it belongs.The splitting rules and values can be summarized in a tree form. The regions are predicted based on the minimum root mean square value obtained for all points lying in that region, and since this approach cannot be feasible for many points, we use what is called the top down, greedy approach which minimizes the cost function, which is called as recursive binary splitting. This is a numerical procedure where all the values are lined up and different split points are tried and tested using a cost function. The split with the best cost (lowest cost because we minimize cost) is selected. All input variables and all possible split points are evaluated and chosen in a greedy manner (e.g. the very best split point is chosen each time). The option of pruning the tree to improve the over-fitted model and hence reduces bias. The fastest and simplest pruning method is to work through each leaf node in the tree and evaluate the effect of removing it using a hold-out test set. For our model, we have chosen to prune the tree by using the prune function and setting the cp limit to 0.13. This pruned model was also cross validated across all ten folds of the training data set and hence we can compare the performance of the pruned and unpruned model on the same cross validated set to evaluate its out of sample performance.

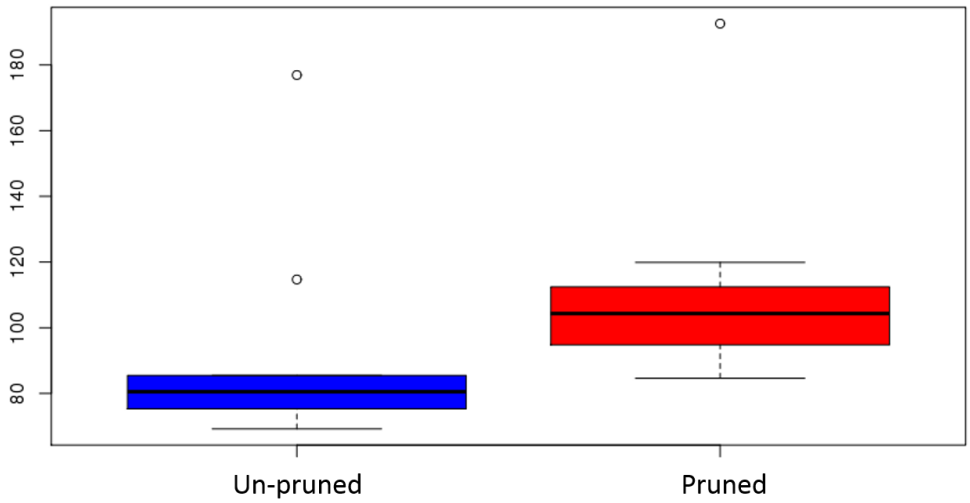

Figure 11: Box plot of Un-pruned vs Pruned CART RMSE values

Based on these values, we can safely say that since the unpruned model performs with much lower average RMSE (refer table3, pg.20) values, we would go with that model. We go on to predict the final out of sample RMSE (refer table3, pg.20) values using this model.

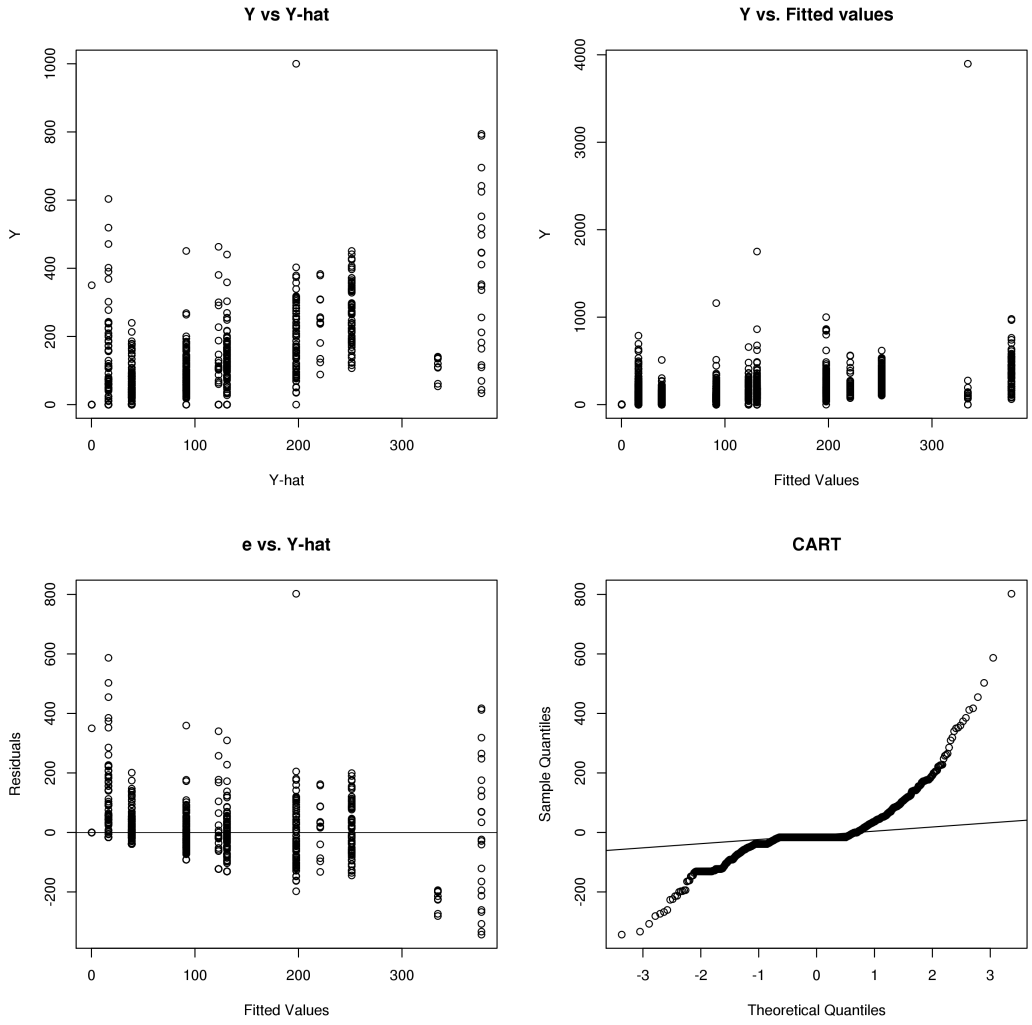

Figure 12: CART Diagnostics and Residual Analysis

The diagnostics and residual analysis reveals severe problems. The predictions and fitted values are not linear with respect to the response. Also, the fitted and predicted values are grouped on specific x-values. This behavior is analogous to plotting factor variables with respect to a response, our response variable is continuous. The error variance is heteroscedastic and the residuals are not normally distributed.

A disadvantage of using this model is that all terms are assumed to interact. Every variable in the tree is forced to interact with every variable further up the tree, whereas some variables do not interact in our dataset. This is extremely inefficient if there are variables that have no or weak interactions. Another drawback is that this model tends to over-fit as it is a low bias high variance model and hence is not good on out of sample predictions for our data set. The next option is to use decision trees with bootstrapping which would reduce the variance.

5.5 Random Forest

Random forest[15] is an ensemble method for predicting response by growing multiple decision trees and combining their performance and randomly selecting subsets of predictor variables at each node to improve predictive power. RandomForest uses a so-called bagging approach. The idea is based on the classic bias-variance trade off. RandomForest tries to achieve this by doing a so-called bootstraps/sub-sampling. The prediction is the average of individual estimators so the low-bias property is successfully preserved. And further by Central Limit Theorem, the variance of this sample average has a variance equal to variance of individual estimator divided by square root of N. So now, it has both low-bias and low-variance properties, and this is why RandomForest often outperforms stand-alone estimator. Formally, a random forest is a predictor consisting of a collection of randomized base regression trees

$$r_n(x, \theta_m, D_n), M \geq 1$$

, where

$$\theta_1, ..., \theta_m$$

are i.i.d outputs of a randomizing variable. The other salient part of this model is the random subset selection. Due to this, effects of multicollinearity in the dataset are greatly reduced. This is because with random selection the relative importance of each feature can be distinctly calculated. The random forest algorithm beautifully balances the bias-variance and multicollinearity-Rsq tradeoffs and does the best variable selection while ensuring that the R-Squared value is not affected greatly. Also, the random forest model does not have assumptions about distribution of data like in case of regression models. In other words, it is a non-parametric model.

For our analysis, we fit the model using all the variables, and based on the variable importance plot and out of sample prediction accuracy, the final model is selected using the reduced model. The results from this model and the inferences made are explained in detail later in the final model section (refer sec.6, pg.20) as this was the best model for out of sample prediction on our data.

5.6 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS)[16] is a non-parametric regression method that models the non-linearities in the data using 'piecewise continuous linear model' or *hinge functions*. Thus, the non-linear data can be modeled using these *hinge functions* with 'kinks' or inflection points in the line of the best fit. MARS builds models which are weighted sum of the *Basis functions*. Mathematically they can be represented as:

$$f(x) = \sum_{i=1}^{k} c_i B_i(x)$$

As such each *Basis functions* $B_i(x)$ can take one of three values:

1. a constant value 1 or the intercept term.
2. a *hinge function*. A *hinge function* is of the form: $\mathbf{max}(0, x - constant)$ or $\mathbf{max}(0, constant - x)$. A MARS model selects the variable and the values of those variables based on the hinge functions automatically.
3. two or more *hinge functions*.[17]

The non-linearity of the model can be better fitted if we increase the number of hinges to represent small changes in the data. The fitting routine in a MARS model involves searching of all the possible combinations of the variables (including the intercept) and over all inflection points with respect to each variable. This fitting method is known as *forward pass*. In general, the *forward pass* method leads to over-fitting of the model. It is also possible to implement the *backward pass* where all the possible combinations of variables are trimmed over unneeded inflection points. This is similar to variable pruning in which we start with the

least influential variable and begin discarding the least important variables until we end up with a suitable model.

For our project, we build two models - an unpruned and a pruned MARS model and compare both the models after cross-validating both the models. The *earth* package automatically performs the variable selection and pruning techniques.

5.6.1 Unpruned MARS model

First, we build an unpruned MARS model using the *earth* package. We can achieve this by adding an argument **pmethod** = "none". We can see from fig. 13 that MARS has build a model using nine covariates and their interactions.

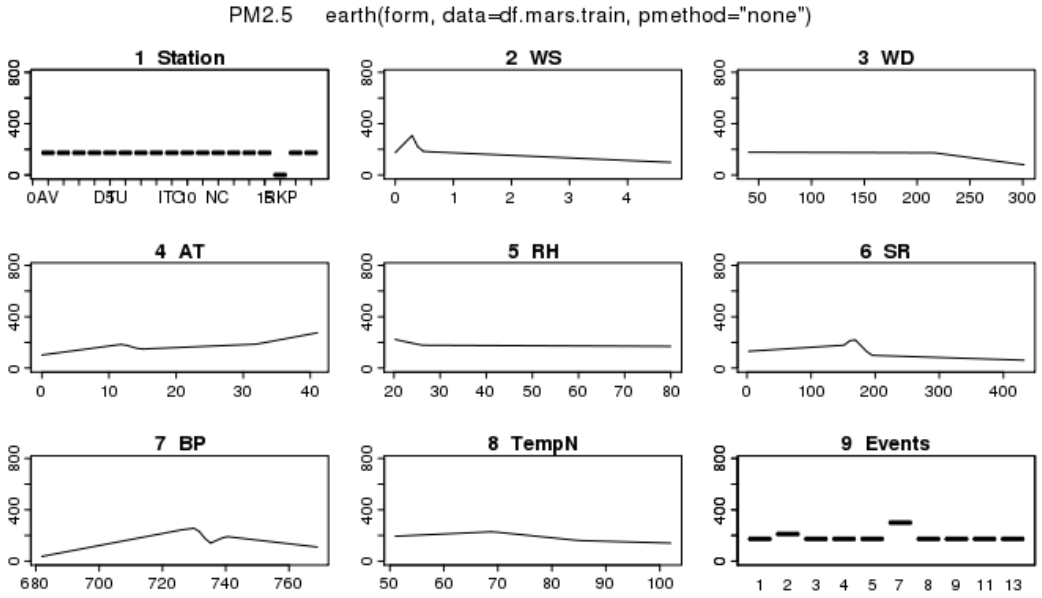

Figure 13: Variable plots for Unpruned MARS model

From the plot, we can observe that predictors "WS", "WD" and "TempN" have a somewhat decreasing relationship with PM2.5. "AT" has a clear increasing relationship with PM2.5. "BP" increases with PM2.5 and there is an inflection point at around 730 mmHg after which it decreases with increasing values of PM2.5. Also, we can observe that the model is over-fitting the data for predictors "WS", "SR" and "BP". There is no visible effect in "SR" and "RH" corresponding to PM2.5.

5.6.2 Pruned MARS model

Now, we build a pruned MARS model using the default pruning method available in the *earth* package.

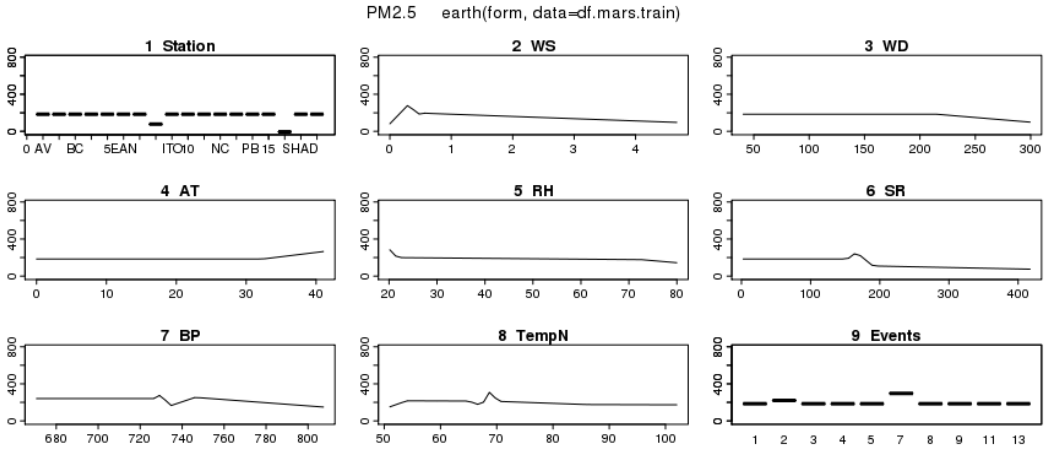

Figure 14: Variable plots for Pruned MARS model

From fig.14, we can observe that the issue of over-fitting in predictor "WS" is somewhat addressed in the pruned model. There isn't much deviation in the relationships from the unpruned model. Also, the pruned model used 26 of the 32 total terms to build the model as compared to all the terms used by the earlier model.

Next, we compare both the MARS model by cross-validating the models in a 10-fold loop and plotting a boxplot of the respective root mean squared errors at each of the folds. The results are obtained as:

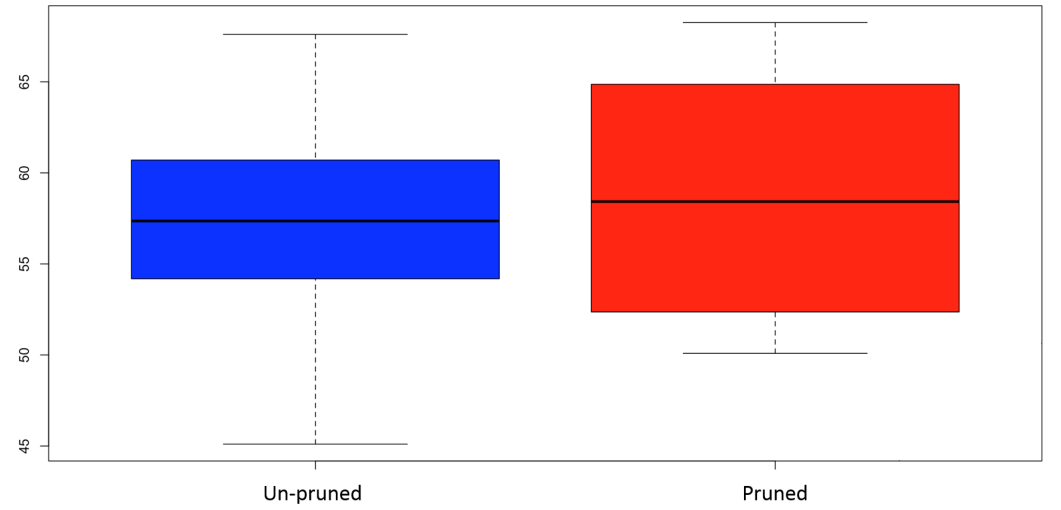

Figure 15: Comparison of out-of-sample RMSE of MARS models

From fig.15, we can conclude that the mean out-of-sample root mean squared error for the Pruned MARS model is slightly higher than the unpruned MARS model. Also, the pruned model has more noise compared to the other model. Although the unpruned model slightly over-fits the data , it performs better on out-of-sample data and has less variability of errors. So, we use the unpruned model to predict PM2.5 on the test data.

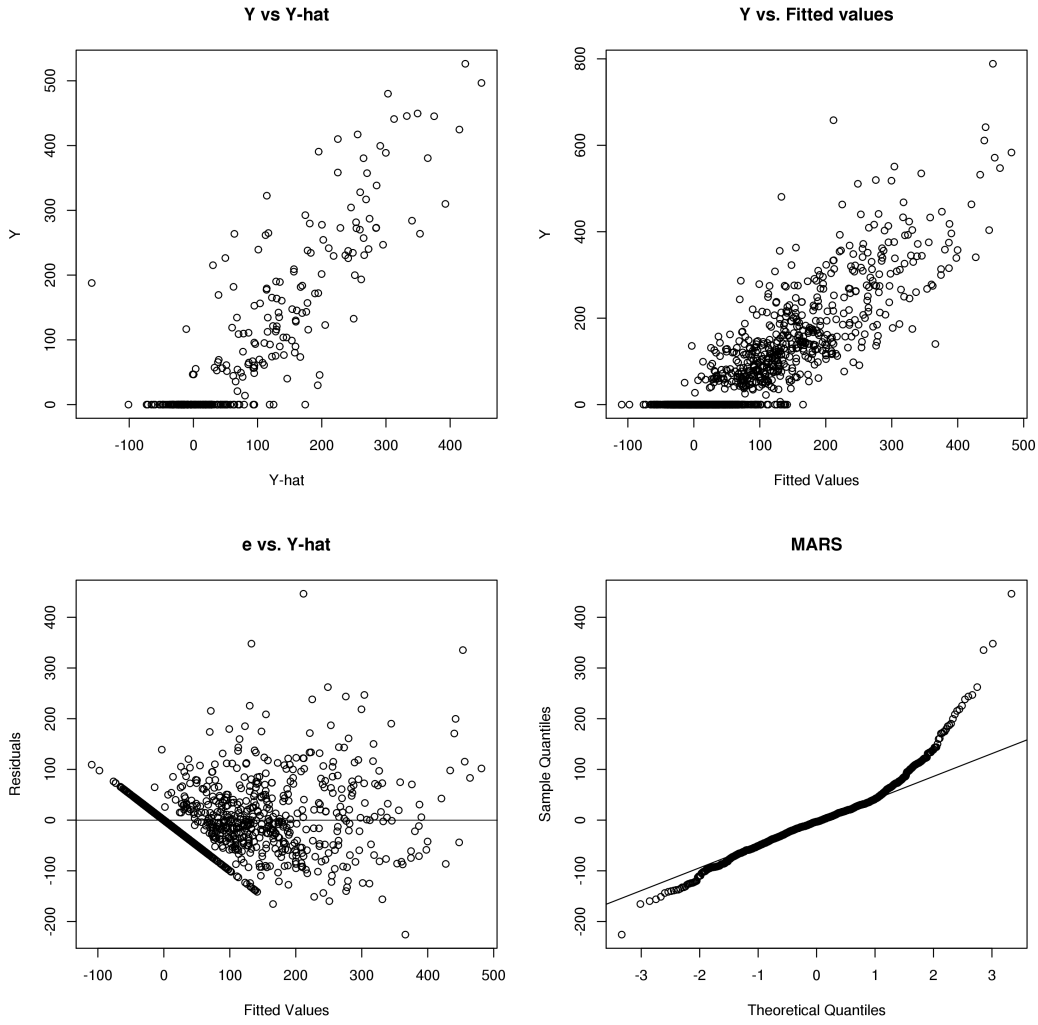

Figure 16: MARS Diagnostics and Residual Analysis

Residuals vs. fitted values indicate a megaphone pattern, thus heteroscedasticity. Also, the QQ-plot although largely normal, there is an evident tail effect. The fitted and predicted values indicate good predictions.

A advantage of the MARS model is that it allows a lot of flexibility due to the *hinge functions*. Also, MARS model seems to give good results without scaling the data. Pruning the model here does not seem to have a rather significant effect on over-fitting of the training data. Pruning should have reduced the problem of over-fitting, however for the increased cost of pruning the performance on out-of-sample is unsatisfactory.

5.7 Support Vector Machine (SVM)

SVMs[18] were developed by for binary classification and the approach behind SVM can be described as: The basic idea behind SVM is to find the optimal separating hyperplane between any two classes by maximizing the *margin* such that the distance between the two classes' closest points is maximized. The points lying on the *margin* separating the two classes are known as *support vectors* or *sparse vectors* . If the *support vectors* lie on the margin then then the margin is known as 'soft margin' , while if there are *support vectors* inside the separating hyperplane the margin is called 'hard margin'. The margin can be represented in the form of optimization equation as:

$$max \left\{ y_i(\beta^T x_i + \beta_o) / |\beta| \right\}$$

where $y_i(\beta^T x_i + \beta_o)$ is the orthogonal distance of the *support vectors* on the margin and $|\beta|$ is the norm vector. Thus, we can formulate this as a quadratic optimization problem which can be solved using Lagrangian multipliers and finding the K.K.T conditions. Since, the orthogonal distance is always greater than 0, we can maximize the function by minimizing the norm vector. The quadratic optimization problem now can be formulated as:

$$min \left\{ 1/2 \left|\beta\right|^2 \right\}$$

given that

$$y_i(\beta^T x_i + \beta_o) \geq 1 \forall x_i$$

Solving this quadratic optimization problem gives us the optimal hyperplane. In case of non-linear planes, the data points are projected in a higher dimensional space using a *kernel function* where the transformed data points can be linearly separated.

The regression capabilities of the SVMs are demonstrated using the dataset build in the project. Again, we split the data into a train and a test set and use 10-fold cross-validation to check the prediction accuracy of the SVM model.

We set the kernel parameters cost = 100 and gamma = 0.0001 to build our SVM model.

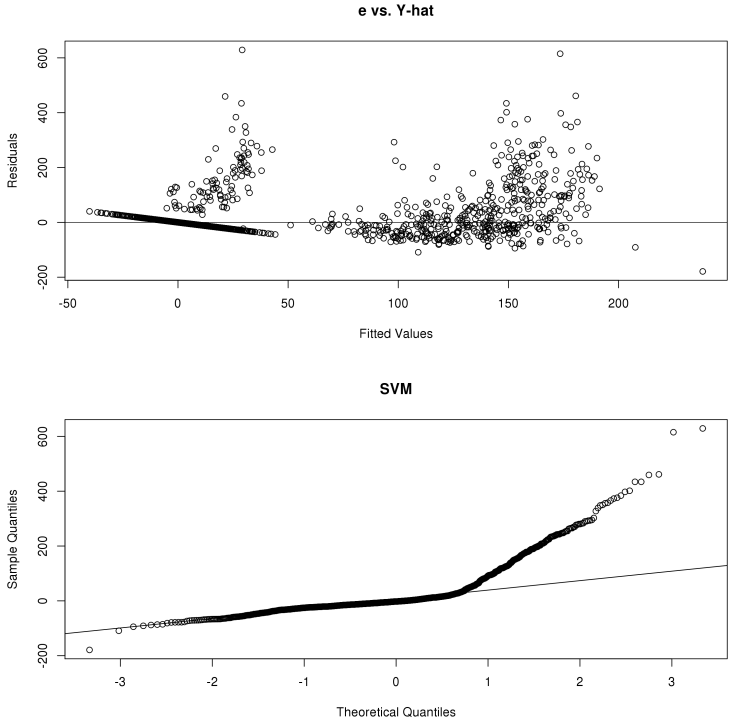

Figure 17: SVM Diagnostics and Residual Analysis

SVM builds a very robust model since the quadratic optimization problem gives us the global minima or maxima. But some of the drawbacks of SVM evident here are: SVMs scale rather badly with the non-linearities in the data due to quadratic optimization algorithm and kernel transformation. Also, the correct choice of the kernel parameters is crucial here to obtain good predictions[18]. The residuals severely violates the assumption of homoscedasticity. The normality plot is satisfactory, although we should be concerned about the tail.

6 Final Model

Based on the comparison of out-of-sample RMSE among the models, along with taking into account the interpretability of the model, we have decided that **Random Forest** is the best model. For reference, following is the performance of the models and comparison of their in-sample and out-of-sample RMSE values.

Models	***In sample RMSE***	***Out of sample RMSE***
GLM	126	116
GAM	111	68
Unpruned CART	91.289	79.2616
Random Forest	21.24	47.58
BART	85	75
Unpruned MARS	58.15	60.61
SVM	89	91.75

Table 3: RMSE Comparison Table

From the table above, we see that the RMSE values for the random forest model is the least for out of sample data and hence, we conclude that this model works best for our data and predicting any new data point.

For some of the models, RMSE $In-Sample > Out-Of-Sample$. This is due to the random selection of training rows. So there might be a selection of rows for which the model is highly under-fit and hence the resulting RMSE for that particular fold is high. This particular RMSE may have skewed the overall expected value of error.

6.1 Variable Importance

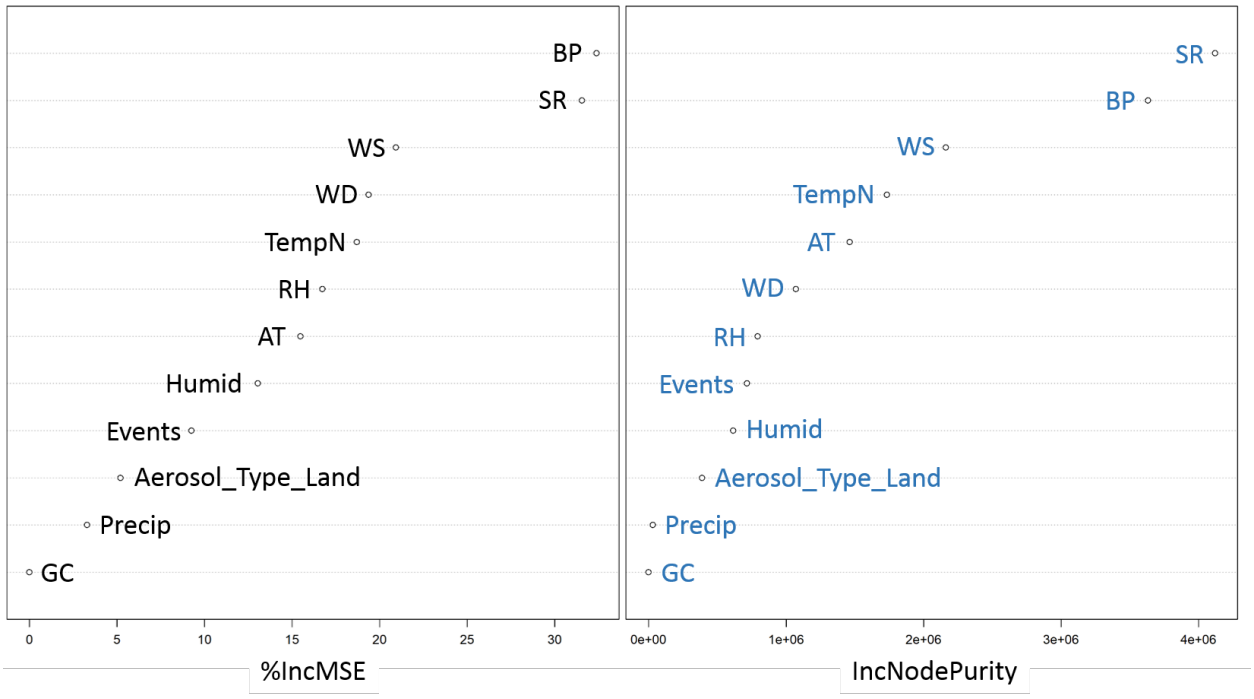

Figure 18: Variable importance plot for Final Model

6.1.1 MSE Importance Plot

The variable importance plot indicates the relative importance of the 12 variables used in our final model.The plot is indicative of the out-of-sample RMSE increase for a particular variable if it is selected.

The following variables affect the out-of-sample RMSE the most:

1. BP (Barometric Pressure)
2. SR (Solar Radiance)

6.1.2 Node Purity Plot

According to the node inclusion purity plot (Fig.18,pg.20). The following predictor variables are the most important in fitting the model properly:

1. SR (Solar Radiance)
2. BP (Barometric Pressure)

Node Inclusion Purity indicates the improvement in the fit of the model with particular predictor in the model. It considers the best split on that variable as voted upon by multiple trees to achieve the best RMSE values for prediction. We have considered $\%IncMSE$ as metric for variable importance.

6.2 Diagnostics

6.2.1 Residual Analysis

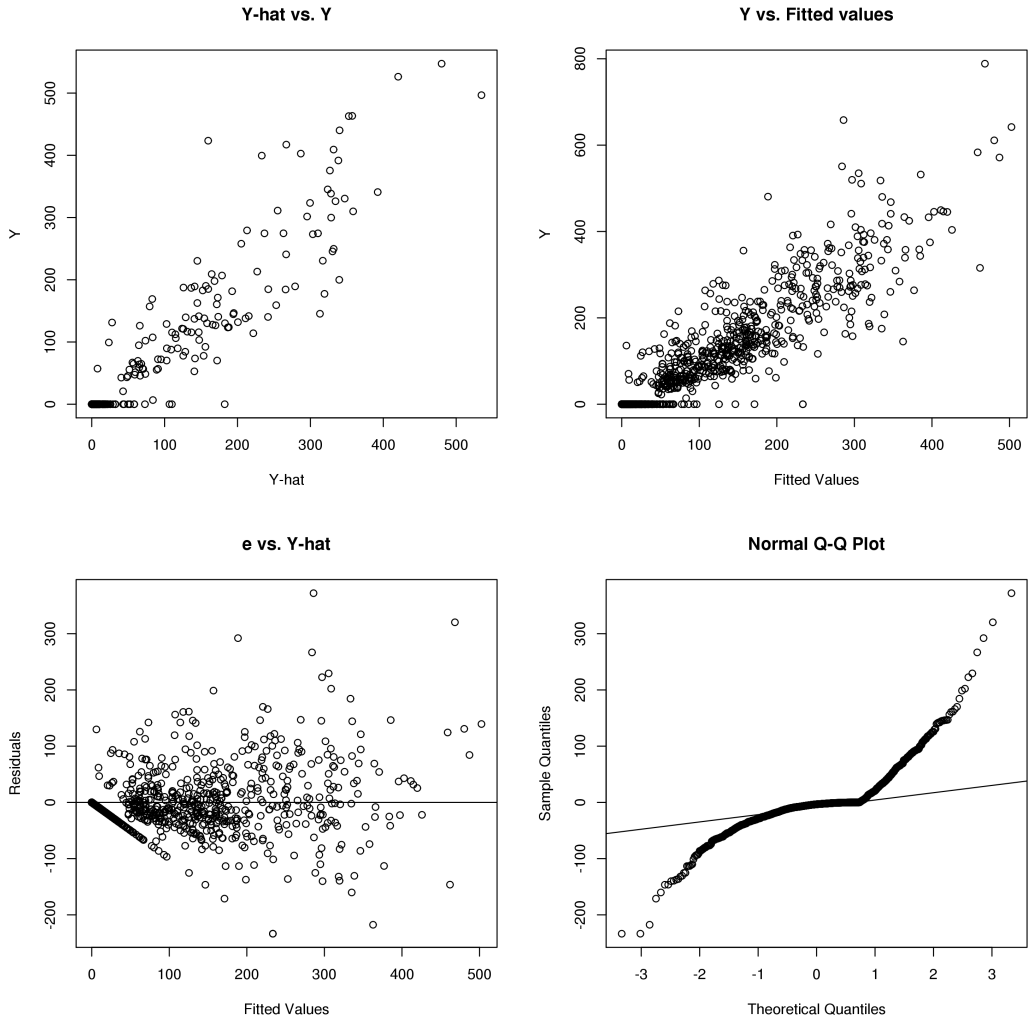

Figure 19: Final Model Diagnostics and Residual Analysis

1. **Y-Hat vs. Y**: This plot indicates the out-of-sample performance of the model. The predictions look good, in the sense that the points follow a linear trend (when examined visually).

2. **Y vs. Fitted values**: This plot indicates the in-sample performance of the model. The model fitted the in-sample points properly. Meaning the fitted values follow a linear trend (visually) with respect to the actual values.

3. **e vs. Y-hat**: This plot has a megaphone pattern. Meaning the residuals are heteroscedastic. The residuals are concentrated near zero which means there are no outliers or influential values. This behavior may have been originated from autocorrelations among the individual predictor values.

4. **Normality Plot**: The assumption of normality is violated. This is a direct result of heteroscedasticity.

6.2.2 Partial Dependence Plots

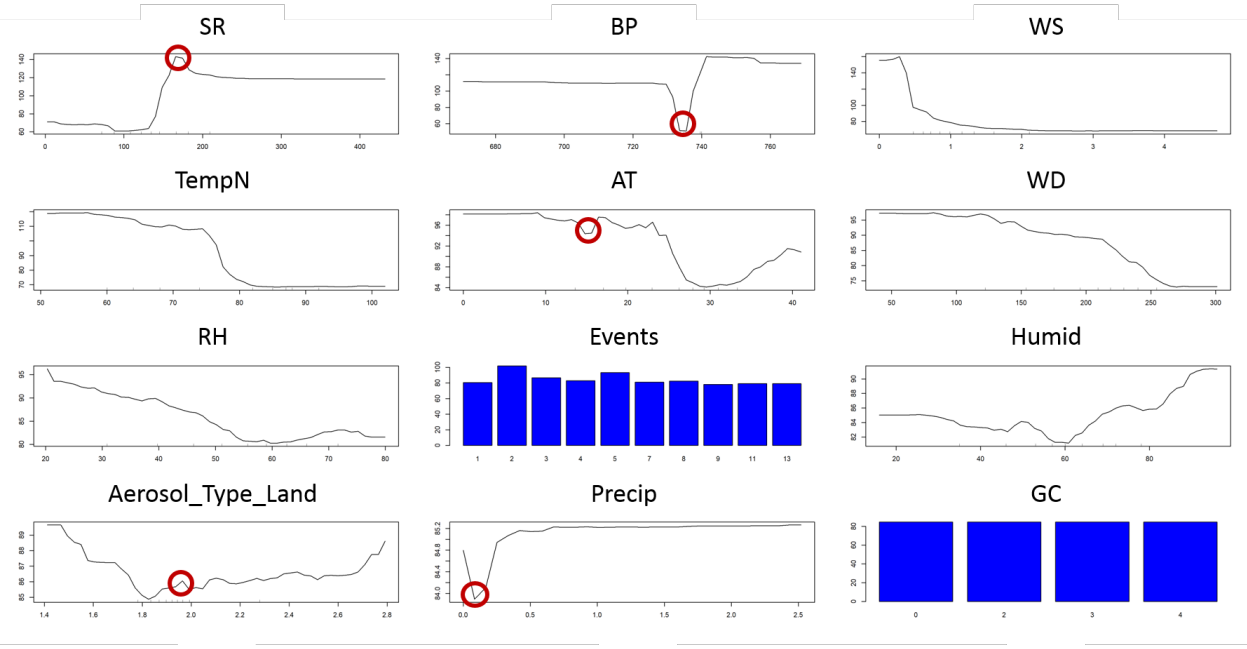

Figure 20: Partial Dependence Plots for Final Model

Note: Red circle indicates over-fitting.

The partial dependence plot indicates the over-fitting of certain predictor variables as listed:

1. SR (Solar Radiance)
2. BP (Barometric Pressure)
3. AT (Ambient Temperature)
4. Aerosol_Type_Land
5. Precip (Precipitation)

The possible reasons for over-fitting are as follows:

1. Dataset is sparse, although the date range is from January 2015 to January 2018, the values for a lot of days and stations are missing. Hence, the dataset is not continuous.
2. Random forest is an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing
3. For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

Partial Dependence Inferences

Negative Influence	Positive Influence	Unusual
SR	RH	BP
Precip	AT	AOT
Humid.	TempN	GC
	WD	
	WS	

Table 4: Response behavior w.r.t Predictor Variables

Unusual Behavior:

1. BP (Barometric Pressure): The PM2.5 concentration increases for higher barometric pressure. Also, PM2.5 decreases for lower pressures. This is due to winds flow from the region of higher pressure to lower pressure thus increases the turbulence in the atmosphere. This turbulence hinders the mixing of precursors which form PM2.5. A kink in the partial dependence plot indicates an anomaly, which needs to be addressed in future.

2. Aerosol_Type_Land (AOT): AOT indicates the loading of aerosol matter in a swath of land. Thus it has been shown through studies [4], that AOT positively affects PM2.5. According to the plot, PM2.5 decreases, initially, with increase in AOT and increases thereafter. This behavior is unusual.

3. GC (Green Cover): There seems to be no interpretation of GC, which is unusual.

7 Conclusion

In the national capital region of India, New Delhi, high levels of PM2.5 are usually accompanied by low temperatures and foggy conditions. According to the partial dependence plots (fig.20,pg.23) the response **PM2.5** increases with a drop in **TempN**. Also, the concentration of **PM2.5** is higher when **Event 2 (Events' partial dependence plot)** occurs. Event 2 is fog. Thus, this result statistically shows the association of low temperatures and fog with elevated PM2.5 levels.

Lower **WS** (wind-speeds) are conducive to elevated levels of PM2.5. Since, New Delhi is landlocked, the wind speeds are usually lower than that of the coastal regions. Increase in wind-speed increases the agitation in the atmosphere thereby reducing the probability of precursor's interactions.

Also, the concentration of **PM2.5** increases with lower levels of **RH** (Relative Humidity). This is due to fact that higher RH means some form of precipitation, either droplets or rain. Also, humidity affects the AOT-PM2.5 relationship owing to the hygroscopic growth of aerosols.[4]. High relative humidity enhances the growth of secondary pollutants and can change the optical properties of these particles as well as change the size distribution. This production of secondary particles increases the concentration levels of PM2.5[19]. Higher **SR** (Solar Radiance) leads to higher levels of PM2.5. This is due to the fact that higher solar radiance leads to photochemical reactions amongst the pollutants and thereby increases the chemical concentration of the precursors.[4]

In essence, we can understand why the predictive accuracy of the **random forest** model is the best for our dataset. The theory behind this performance is the random sub-sampling of data variables amongst various trees and assigning random probabilities to these trees.

1. We can see that random forest is very easy to train and requires almost no input data preparation
2. It provides an implicit feature selection on its own and gives a good indicator of feature importance and hence has a low bias
3. Since, we perform bagging of random uncorrelated bootstrap samples, we manage to reduce the high variance from the regression trees.
4. Versatility and simplicity of random forest is another feature of why it is preferred over other models
5. Random forest can be grown in parallel very easily and cannot be done for other bagging and boosted models
6. Random Forest maintains accuracy even in case of sparse data sets
7. The drawbacks are the model is computationally intensive and also it is sometimes very hard to interpret.

In conclusion, the findings of our model are consistent with the everyday observations of high PM2.5 associated with weather and environmental changes. For instance, the concentration of PM2.5 shoots up drastically in winter months especially in the presence of fog. Our model has statistically shown that this is indeed the case. We expect that these findings can be extended to other regions of India where we face similar issues of high surface level concentrations of PM2.5. Other areas particularly include landlocked small towns with high industrial activity.

8 Future Scope

1. Identify the predictor(s) for which the variance is not properly captured (reason for heteroscedasticity). This will solve the problem for normality as well.
2. Search for other avenues to look for quality controlled data.
3. Apply models to more number of stations to increase the training input.
4. More research can be done to check the effect of green cover (vegetation) on AQI. We have included GC (Green Cover) in our study. However, we suspect that we did not have proper data for the same. Therefore, we would like to explore the effects of green cover in future.
5. Solve the problem of auto-correlations

Appendix

Stations

Sr. No.	Station Name	Abbreviation
1	Anand Vihar	AV
2	Aya Nagar	AYA
3	Burari Crossing	BC
4	Central Road Research Institute	CRRI
5	Delhi Technical University	DTU
6	Indira Gandhi International Airport	IGI
7	Institute of Human Behavioural and Allied Sciences	IHBAS
8	Lodhi Road	LODHI
9	Mandir Marg	MM
10	North Campus	NC
11	Netaji Subhas Institute of Technology	NSIT
12	Punjabi Bagh	PB
13	Pusa	PUSA
14	R.K. Puram	RKP
15	Shadipur	SHAD
16	Siri Fort	SIRI
17	East Arjun Nagar	EAN
18	Income Tax Office	ITO

Table 5: List of 18 weather stations in Delhi

Events' Description

Factor	Description
1	No Activity
2	Fog
3	Fog,Rain
4	Fog, Rain, Hail, Thunderstorm
5	Fog, Rain, Thunderstorm
6	Fog, Thunderstorm
7	Fog, Tornado
8	Hail, Thunderstorm
9	Rain
10	Rain, Hail, Thunderstorm
11	Rain, Thunderstorm
12	Snow
13	Thunderstorm

Table 6: Events' factors and their description

References

[1] Delhi wakes up to hazardous pollution levels, reduced visibility due to smog. *Indian Express*, 2017.

[2] Sweta Goswami. Delhi's worst smog yet wakes up govt., emergency measures announced. *Hindustan Times*, 2017.

[3] Delhi Pollution: Government issues health advisory as smog chokes city. *Hindustan Times*, 2017.

[4] Pawan Gupta and Sundar A. Christopher. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *Journal of Geophysical Research: Atmospheres*, 114(D14):n/a–n/a, 2009. D14205.

[5] Delhi wakes up to dense smog; train services suspended due to low visibility. *Times Now Bureau*, 2017.

[6] Pune Indian Institute of Tropical Meteorology. SAFAR-India, 2017.

[7] Robert C Levy, Lorraine A Remer, and Leigh A Munchak. A surface reflectance scheme for retrieving aerosol optical depth over urban surfaces in modis dark target retrieval algorithm. *Atmospheric Measurement Techniques*, 9(7):3293, 2016.

[8] Cheng-Hsiung Huang and Chih-Yuen Tai. Relative humidity effect on PM2.5readings recorded by collocated beta attenuation monitors. *Environmental Engineering Science*, 25(7):1079–1090, sep 2008.

[9] Government of Delhi India. Forest department.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[11] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[12] Trevor Hastie. *gam: Generalized Additive Models*, 2018. R package version 1.15.

[13] Adam Kapelner and Justin Bleich. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016.

[14] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2018. R package version 4.1-13.

[15] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[16] Stephen Milborrow. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. *earth: Multivariate Adaptive Regression Splines*, 2018. R package version 4.6.2.

[17] Multivariate adaptive regression splines, Apr 2018.

[18] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. R package version 1.6-8.

[19] Jun Wang and Scot T Martin. Satellite characterization of urban aerosols: Importance of including hygroscopicity and mixing state in the retrieval algorithms. *Journal of Geophysical Research: Atmospheres*, 112(D17), 2007.