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

Team Harmanik | Manan Shah | Hardik Shah | Nikhil Soni

04-24-2018

Table of Contents

1. Background
2. Current Tools
3. Motivation
4. Data Collection and Cleaning
5. Response
6. Data
7. Data Visualization
8. Models and comparisons
9. Variable Importance Plots
10. Model diagnostics for the final model
11. Inference
12. Conclusion(s)
13. Future scope of work

Table of contents

1. Background
2. Current Tools
3. Motivation
4. Data Collection and Cleaning
5. Response
6. Data
7. Data Visualization
8. Models and comparisons
9. Variable Importance Plots
10. Model diagnostics for the final model
11. Inference
12. Conclusion(s)
13. Future scope of work

1. Background

- The occurrences of smog in the National Capital Region of India has gone up, the concentration of PM 2.5 are through the roof.
- Deteriorating air quality has far reaching effects on health such as multiple sclerosis and lung cancer
- 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.

2. Current Tools

- Most of the tools available today, use geospatial variables such as **Aerosol Optical Depth** combined with environmental variables such as temperature, humidity and solar radiation.
- These tools primarily use models such as multiple linear regression which utilizes an implication of a parametric function.

Table of contents

3. Motivation

- We have endeavored to apply non-parametric and non-linear approaches in capturing the data
- In addition to using machine learning techniques, we have included some additional predictor variables such as:
    - Green Cover (vegetation surrounding the city)
- Build a tool which can be used in other regions of India, where installing a weather station to monitor PM2.5 levels might not be feasible.

Table of contents

4. Data collection and cleaning

- We required spatio-temporal data (HDF format), Aerosol Optical Depth, which was collected using historical data from MODIS Aerosol Product.
  - We used a lot of computing power to just parse the dataset and make it in a consumable dataset.
- Climatic variables were scraped from publicly available resources such as Accuweather and Weather Underground
- We wanted to check for the effect of vegetation on the air quality, hence we sourced the data from New Delhi Forest Department, Government of Delhi, India.

Table of contents

5. Response

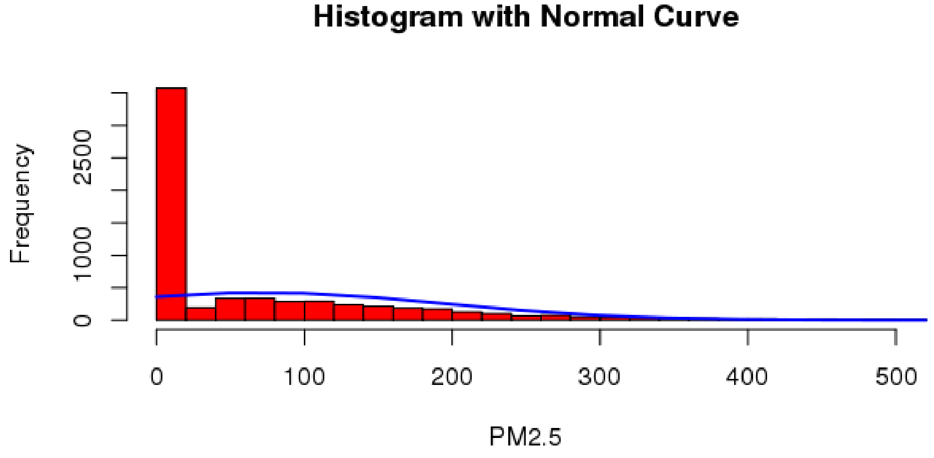
$$PM\ 2.5\ (\mu g^{-3})$$

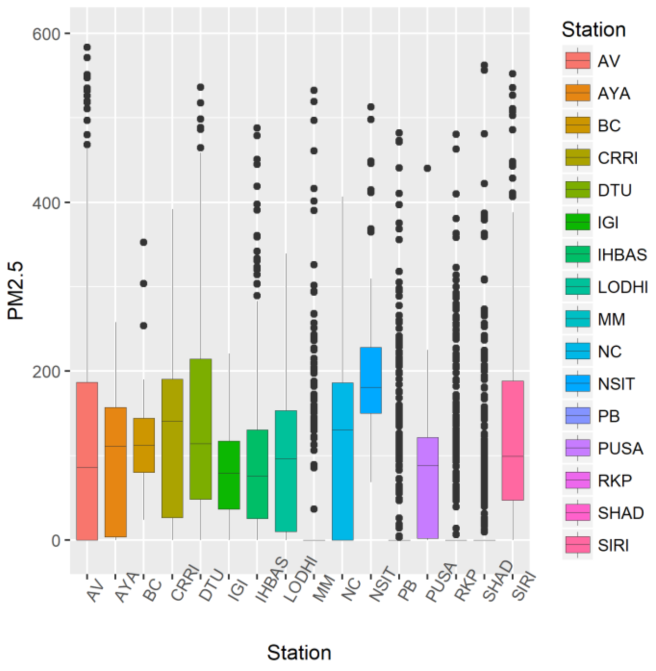
6. Data

Acronym	Description	Source
Date	Date in year, month, date	CPCB
Station	Abbr. Station Name	CPCB
WS	Wind speed in m/s	CPCB
WD	Wind direction in degrees	CPCB
AT	Ambient Temperature in C	CPCB
RH	Relative Humidity in %	CPCB
SR	Solar Radiation in W/m^-2	CPCB
BP	Barometric pressure in mmHg	CPCB
Aerosol_Type_Land	Aerosol Optical Depth	NASA MODIS
TempN	Temperature in C	AccuWeather
Humid	Humidity in %	AccuWeather
Precip	Precipitation	AccuWeather
Events	Natural Phenomena	AccuWeather
GC	Green cover near station	Delhi.gov

Table of contents

7. Data Visualization





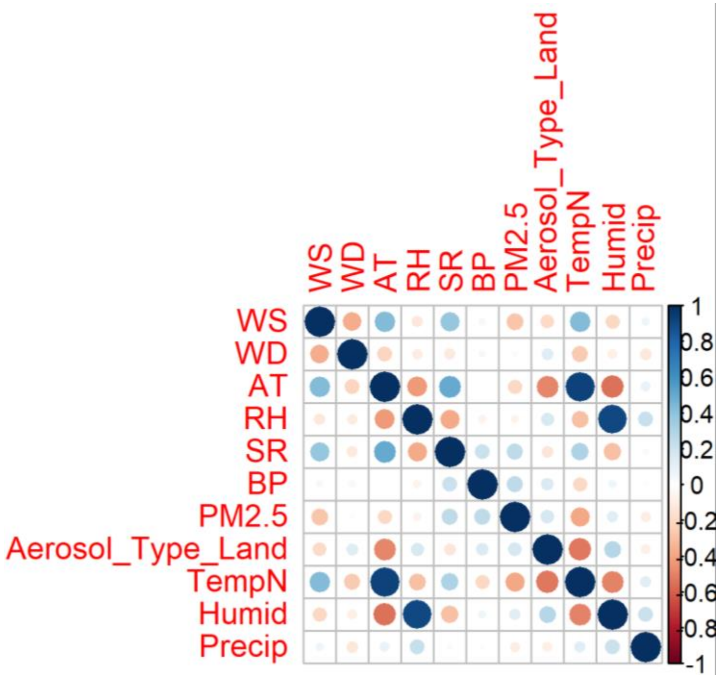


Table of contents

8. Models and comparisons

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 of contents

9. Variable Importance Plot

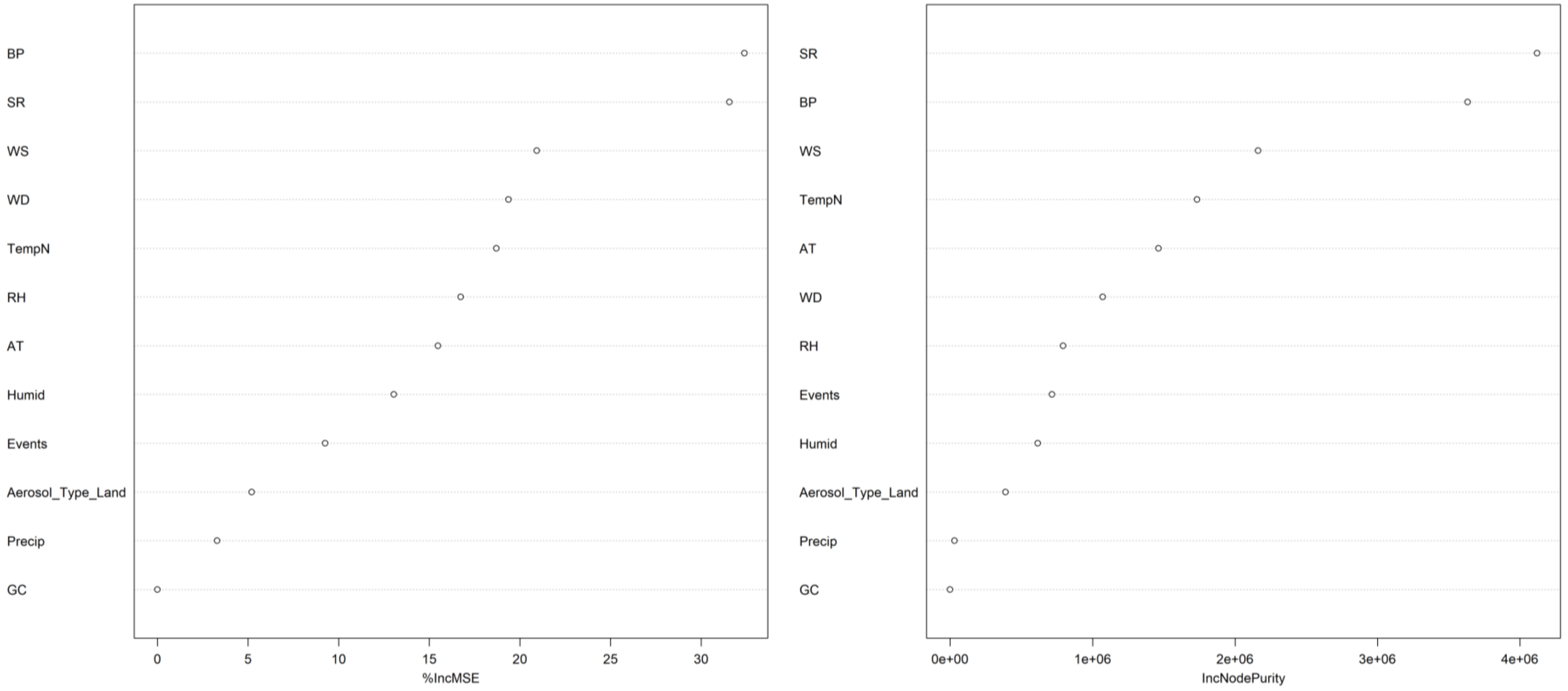


Table of contents

10. Model diagnostics for the final model

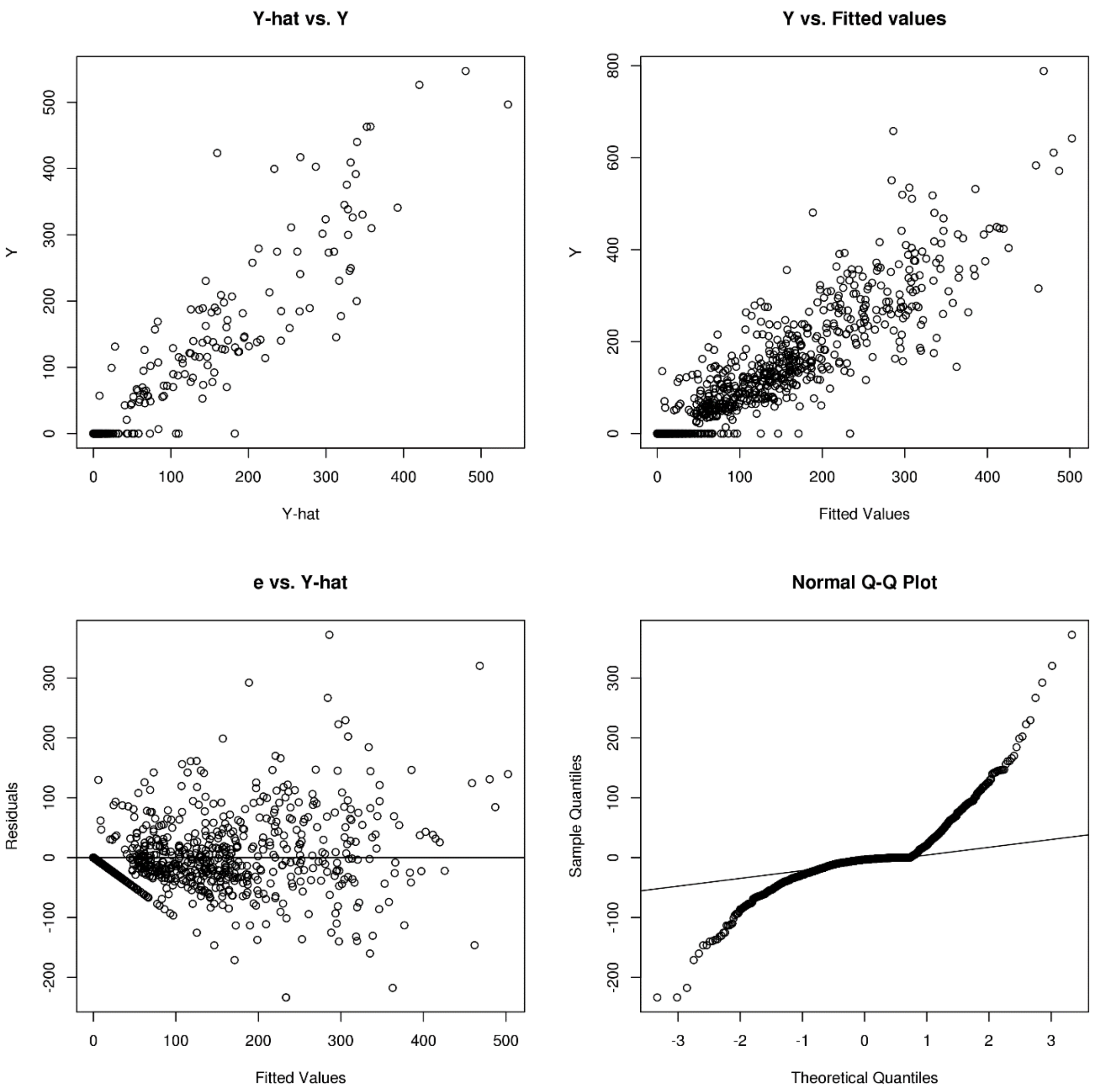
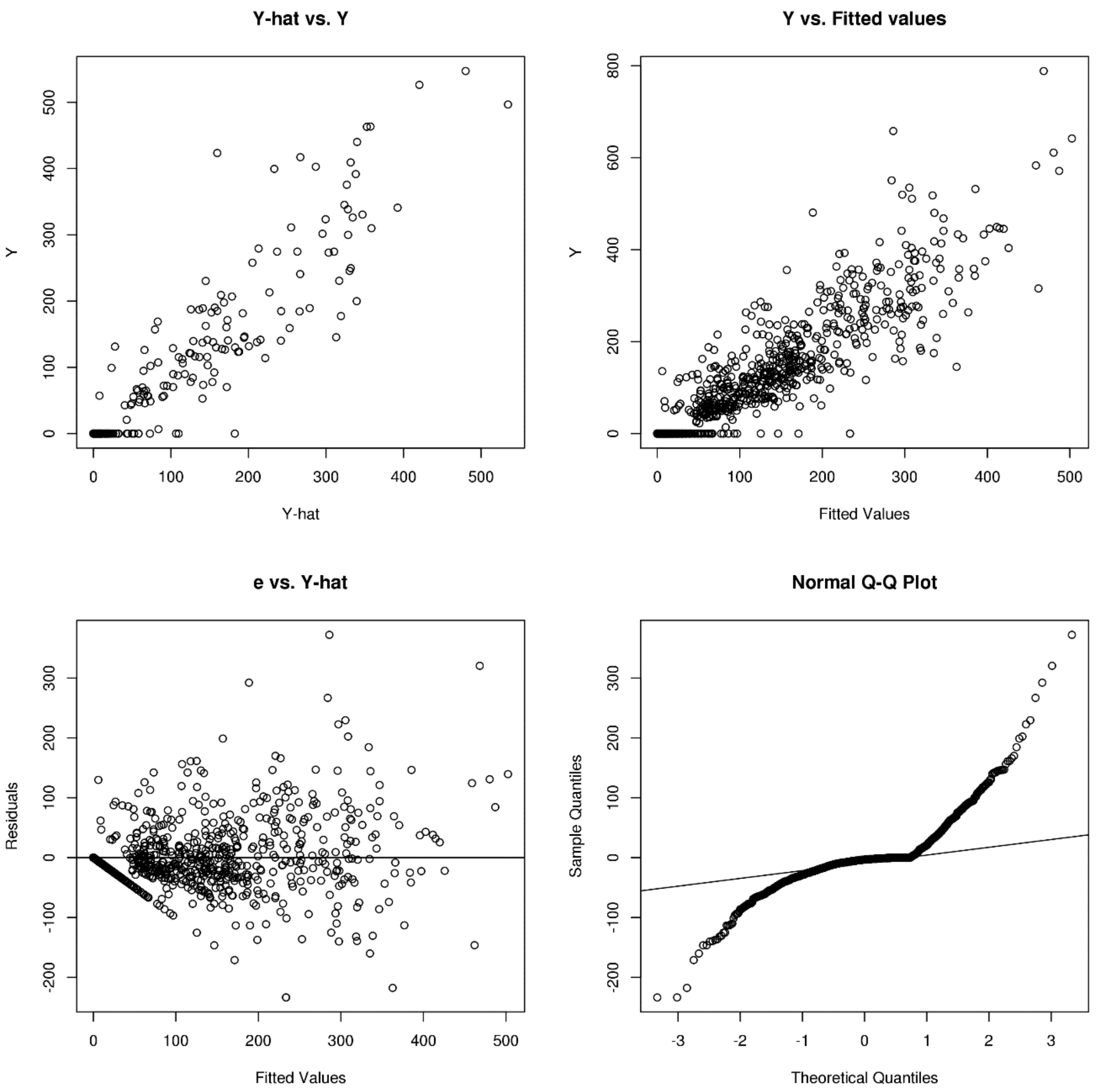


Table of contents

10. Model diagnostics for the final model



Issues with residuals?

1. Heteroscadasticity
2. Non-normal behavior of residuals

Table of contents

11. Inference

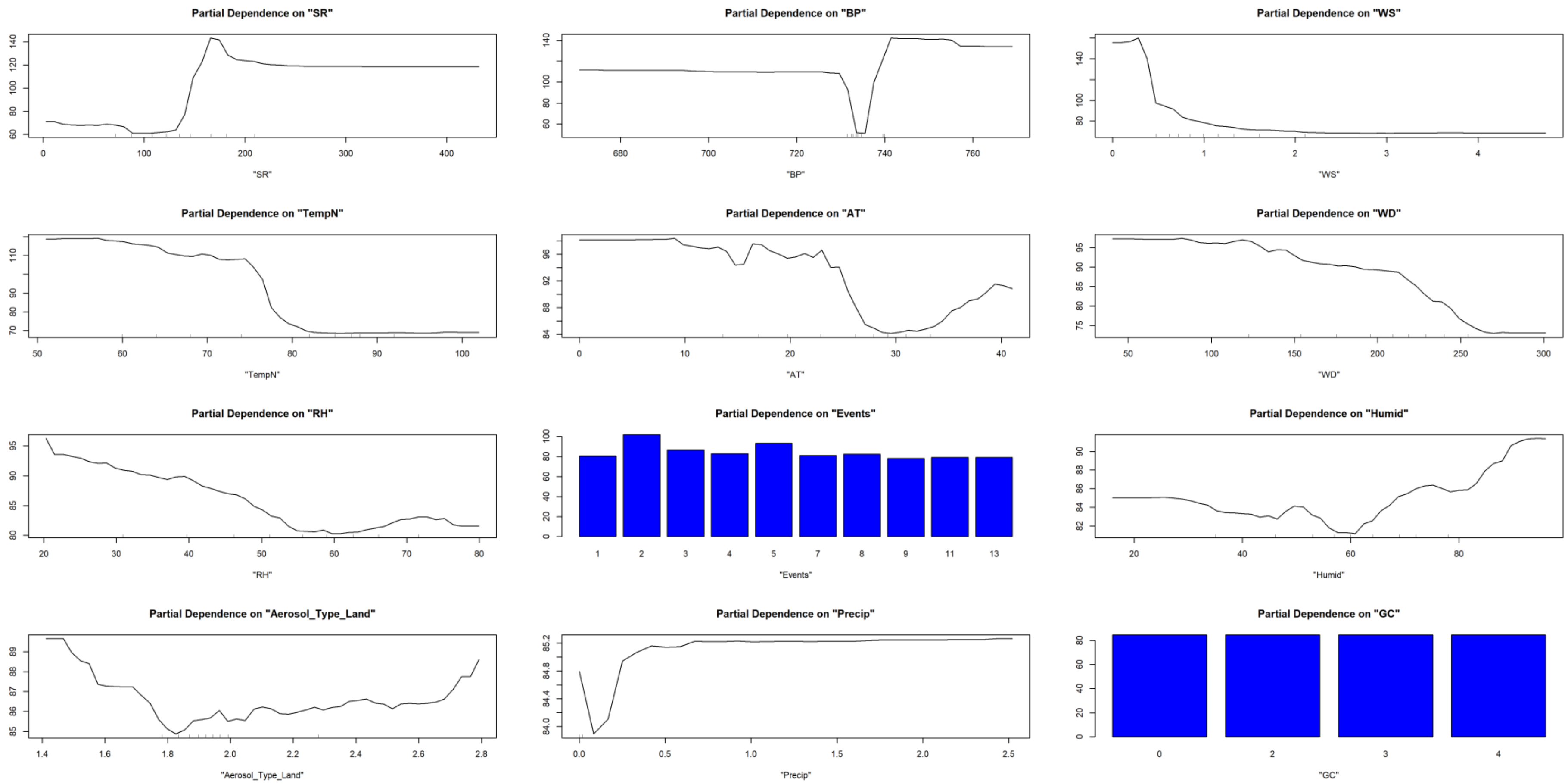
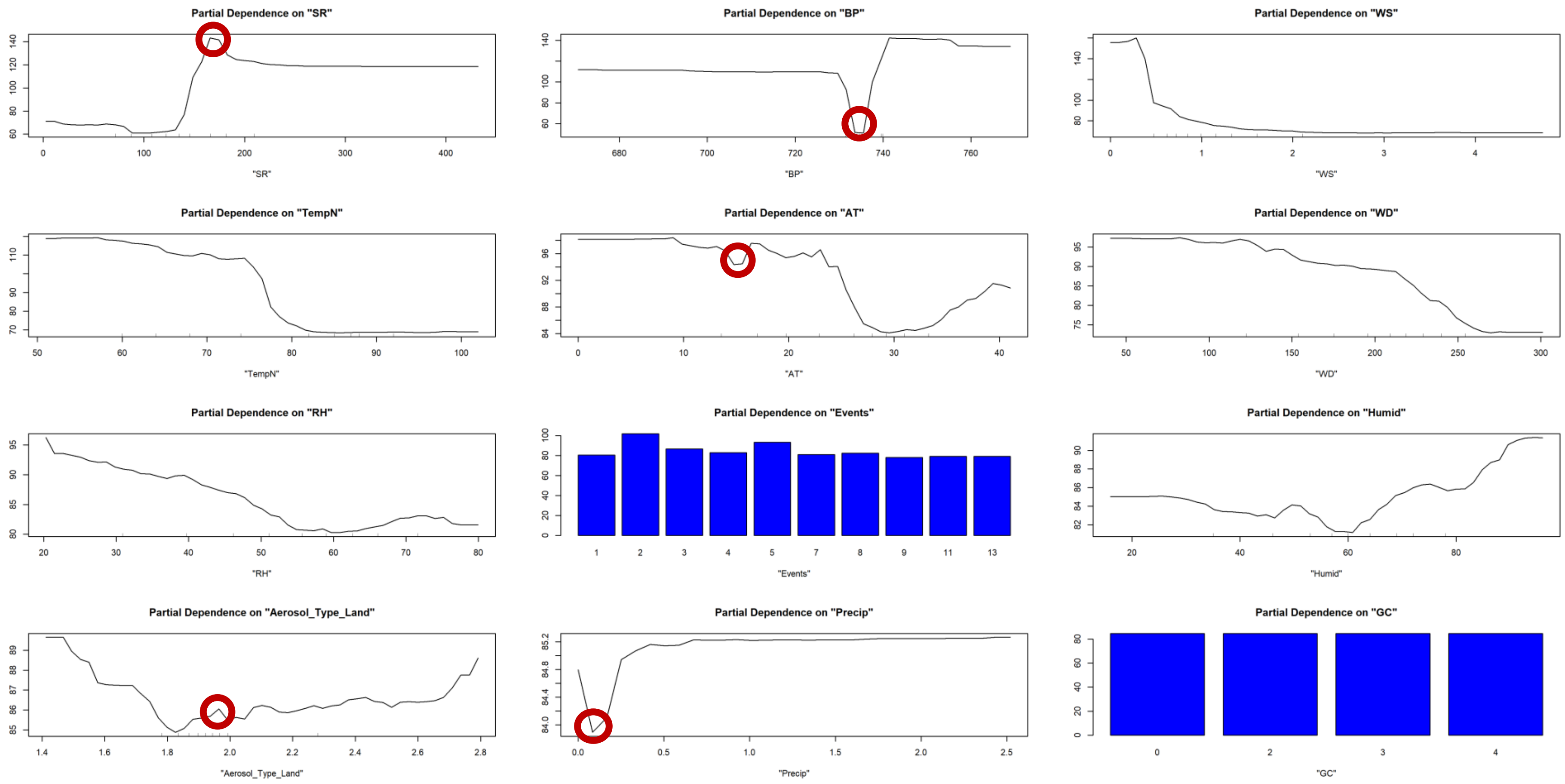


Table of contents

11. Inference



Predictor variables over-fit

Table of contents

12. Conclusion(s)

- The model has good predictive accuracy
- Although it is overfitting for the following variables:
  - SR
  - BP
  - AT
  - Aerosol_Type_Land
  - Precip
- Following can be the reasons :
  - The dataset is sparse
  - Random forest is an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing
  - 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.

Table of contents

12. Conclusion(s)

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

*Events: Fog (Event 2) increases the likelihood of PM2.5 increase

Table of contents

13. Further Scope of Work

- Identify the predictor(s) for which the variance is not properly captured (reason for heteroscedasticity). This will solve the problem for normality as well.
- Search for other avenues to look for quality controlled data.
- Apply models to more number of stations to increase the training input.
- More research can be done to check the effect of green cover (vegetation) on AQI.

Table of contents