# PREDICTING THE AIR QUALITY INDEX OF CITIES USING BIOMASS DATA

MGT 6203: DATA ANALYTICS IN BUSINESS
SUMMER 2023
PROFESSOR BIEN
TEAM 33


**Emily Dobar**
Education Research Data Analytics
B.A. in Mathematics-Economics
Agnes Scott College
edobar3@gatech.edu

**Ian Hollenberg**
Marketing Data Analytics
B.S. in Sports Management
New York University
ihollenberg3@gatech.edu

**Joshua Gauntt**
Data and Automation Analytics
B.S. in Statistics and Data Science
University of Texas, San Antonio
jgauntt3@gatech.edu

**Ross Kelly**
Bridge Engineering
M.S. in General Structural Engineering
Imperial College London
rkelly68@gatech.edu

**Nikolas Preece**
Seismic Data Processing
B.S. in Physics
University of Texas
npreece33@gatech.edu


## Contents

# 1 Overview

## 1.1 Background

Air quality is an important factor in the livability of a city because it influences the health and quality of life of its residents. The air quality index (AQI) is a scale commonly used to measure air quality. It was developed by the Environmental Protection Agency to provide a uniform way to report air quality conditions. It measures five pollutants: fine particles, ground-level Ozone, sulfur dioxide, nitrogen dioxide, and carbon monoxide. Vegetative biomass is also vital for all people, businesses, governments, and other such societal entities due to such things as forests contributing substantially to air quality and overall health, supply chain by way of lumber and paper, and other similar byproducts of vegetation.

In seeing the close connection between air quality and vegetative biomass, we believe that there is a significant predictive relationship between the two, which could be beneficial to everyone, but especially businesses and governments, in that the importance of air quality and being able to reliably predict it could guide decisions involving economic development, supply chain, and other major social and health-related issues that come up in the business and government sectors in an ever-growing environmentally-centered society.

## 1.2 Problem Statement and Research Questions

We intend to use global vegetative biomass data to predict the AQI of a subset of the world's cities.

- Can the AQI of a city be predicted by the amount of vegetative biomass in its vicinity?
  - Which biomass indicators are the strongest predictors of AQI?
  - Are different measures of air quality related to different biomass indicators?

## 1.3 Hypotheses

Our preliminary hypothesis is that the amount of biomass will be a significant predictive feature of value on the Air Quality Index. We also believe that similar features will be predictive of different air quality indicators: CO, Ozone, NO2, and PM2.5.

## 1.4 Business Justification

Green and sustainable initiatives are becoming more commonplace in the business world as the effects of global warming and diminishing air quality become increasingly important societal issues. Poor air quality directly influences health, and by extension, consumer activity and economic output, as consumers may seek products from competing businesses that support green initiatives or may move to locations with cleaner air. Businesses may use this information to determine movements or new locations for buildings, as well as to attract positive attention by way of green initiatives. Businesses care about shareholder opinions, and by decreasing carbon footprint by way of green practices or alternative market location, this could positively impact stock holding price, as well as boost shareholder belief and morale in terms of company share price. In particular, the agriculture, hospitality, and tourism industries could benefit the most in terms of business location and biomass, as air quality directly affects these industries' abilities to provide services/products to their target markets and sustain their businesses financially. Government agencies could also benefit in terms of economic planning, as this could support them in determining which industries to allow in different locations to maintain upbeat morale in their cities, as government leaders care about their public images and how their cities are perceived in public media and as compared to others.

## 1.5 Literature Review

Poor air quality not only negatively affects health and well-being but business productivity and financial success. As of October 2022, approximately "1.2 billion workdays are lost each year due to air pollution" [2], with this number growing exponentially as air quality continues to worsen over time. Lost days could be attributed to work absences due to health concerns or premature deaths related to poor air quality affecting worker health. Similarly, by 2018, the global economy was facing costs around $225 billion each year due to air quality's negative effects on worker productivity, employee sicks days in part due to health issues relating to air pollution, and even recruitment costs and struggles as places with poor air quality tend to be less desirable as places to live and work [5]. Encouragingly, studies are showing potential ways for cities to leverage vegetation as a way for predicting air quality, as trees and other vegetative biomass are able to directly affect air quality in a positive way [4], as validated in a study in Cambridge where air quality was successfully modeled using weather and vegetation factors [1], which could be beneficial for cities, and ultimately businesses, in planning vegetation, economic development, and in planning green initiatives which are becoming more and more necessary in a more environmentally-focused society that is beginning to push governments and industries towards greener initiatives and smaller carbon footprints. Not only will it be more cost-effective in the future for businesses to start early on these greener trends before costly regulations take place, but it will help businesses maintain positive reputations and consumer following, which will ultimately yield regular and even increases in revenue.

# 2 Data

## 2.1 Sources

Data source 1 comes from Kaggle and contains measures of air quality for approximately 14,000 cities around the world [6]. It has variables for an overall air quality measure, as well as measures for Carbon Monoxide, Ozone, and Nitrogen Dioxide. Each variable has a numerical value and a categorical value. See Figure 1

Figure 1: Air Quality Dataset

| Country | City | AQI Value | AQI Category | CO AQI Value | CO AQI Category | Ozone AQI Value | Ozone AQI Category | NO2 AQI Value | NO2 AQI Category | PM2.5 AQI Value | PM2.5 AQI Category | lat | lng |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Russian Federation | Praskoveya | 51 | Moderate | 1 | Good | 36 | Good | 0 | Good | 51 | Moderate | 44.7444 | 44.2031 |
| Brazil | Presidente Dutra | 41 | Good | 1 | Good | 5 | Good | 1 | Good | 41 | Good | -5.29 | -44.49 |
| Brazil | Presidente Dutra | 41 | Good | 1 | Good | 5 | Good | 1 | Good | 41 | Good | -11.2958 | -41.9869 |
| Italy | Priolo Gargallo | 66 | Moderate | 1 | Good | 39 | Good | 2 | Good | 66 | Moderate | 37.1667 | 15.1833 |
| Poland | Przasnysz | 34 | Good | 1 | Good | 34 | Good | 0 | Good | 20 | Good | 53.0167 | 20.8833 |

Data source 2 comes from the Earth Science Data System which provides open access to NASA's collection of Earth science data [3]. The dataset includes several measures of vegetative biomass across the Earth. It covers the entire Earth's surface, binned to 1-degree by 1-degree squares or latitude and longitude and includes measurements taken over time. The total number of rows is 1,310,400. See Figures 2 and

Figure 2: Global Biomass Dataset

| lat | lon | AGB_ha | AboveGroundBiomass | BrBCEF | ConBCEF | ConiferShare | ForestStockChange | Forest_carbonstock | R | Total_Biomass | Unc_Area | Unc_GS | area | stock | stockperarea | time_bnds |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -54.5 | -71.5 | 136 | 10670082 | 0.63 | 0.11 | 0.11 | 1.64 | 6318823 | 0.26 | 13444303 | 2.2 | 3.2 | 78572 | 14508100 | 185 | 60 |
| -54.5 | -70.5 | 136 | 13551226 | 0.63 | 0.11 | 0.11 | 2.08 | 8025036 | 0.26 | 17074545 | 2.2 | 3.2 | 99788 | 18425600 | 185 | 61 |
| -54.5 | -69.5 | 136 | 21641097 | 0.63 | 0.11 | 0.11 | 3.32 | 12815858 | 0.26 | 27267782 | 2.2 | 3.2 | 159360 | 29425400 | 185 | 60 |
| -54.5 | -68.5 | 122 | 24029542 | 0.63 | 0.18 | 0.17 | 3.66 | 14230295 | 0.26 | 30277223 | 3.2 | 3.2 | 196848 | 29771700 | 151 | 61 |
| -54.5 | -67.5 | 99 | 16198869 | 0.71 | 0.20 | 0.17 | 1.77 | 9592970 | 0.26 | 20410575 | 3.2 | 3.2 | 164338 | 17741500 | 108 | 60 |

## 2.2 Description of Features

Figure 3 defines each of the 19 variables contained within the dataset we used for analysis, as well as noting what type of variable each is and the range of values for each.

Figure 3: Data Summary

| Item | Type | Description | Range |
|---|---|---|---|
| Country | String | Name of country | NA |
| City | String | Name of city | NA |
| AQI Value | Float | Air-quality index value | 7 to 500 |
| lat_aqi | Float | Latitude of City | -54.80 to 70.77 |
| lon_aqi | Float | Longitude of City | -156.51 to 178.02 |
| AGB_ha | Float | Above ground biomass per forest area | 0 to 265.6 |
| AboveGroundBiomass | Float | Aboveground forest tree biomass | 0 to 259,798,694 |
| BrBCEF | Float | IPCC default biomass conversion and expansion factors (BCEFs) for Broadleaved | 0 to 3.75 |
| ConBCEF | Float | IPCC default biomass conversion and expansion factors (BCEFs) for Conifers | 0 to 3.06048 |
| ConiferShare | Float | Fraction of forest cover that is conifers | 0 to 1.0 |
| ForestStockChange | Float | Change in forest tree carbon stock between intervals | -18.0959 to 43.8784 |
| Forest_carbonstock | Float | Forest tree carbon stock | 0 to 151,410,679 |
| R | Float | Root-shoot ratio | 0 to .43 |
| Total_Biomass | Float | Total forest tree biomass | 0 to 322,150,381 |
| Unc_Area | Float | Uncertainty for forest area | 1.200 to 5.200 |
| Unc_GS | Float | Uncertainty for growing stock | 1.2 to 5.2 |
| area | Float | Forest Area | 4.9 to 2035590.0 |
| stock | Float | Forest growing stock | 528 to 371,140,992 |
| stockperarea | Float | Forest growing stock per hectare | 3.65 to 350.037 |

### 2.3 Pre-Processing

Data source 2 contains biomass measurements for years 1950 to 2010 inclusive. Data source 1 is not dated. For the purpose of the project, it was assumed that data source 2 is compatible with the data points in data source 1 for the most recent year, 2010. Data points for other years were discarded from data source 2.

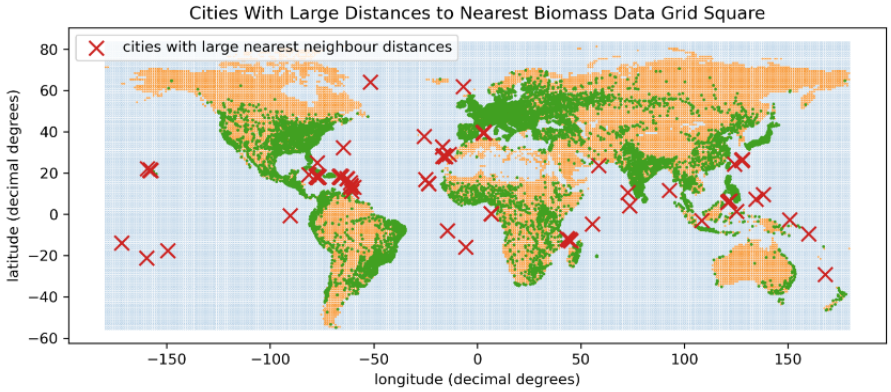
For each grid point, data source 2 contains a variable 'nv', with value 0 or 1, which is not explained in the accompanying explanatory notes. It was found that for all grid points, all the biomass variables were the same for both values of 'nv'. The data points corresponding to nv=1 were discarded from data source 2.

It was found that data source 1 has no missing/null values and data source 2 has a very large number of null values. The spatial distribution of the data source 2 data points revealed that the null data points represent points on the Earth that are ocean, the Arctic, the Antarctic, and some of the world's deserts. The non-null data points appear to be the remaining habitable landmass - refer to Figure 4.

Data sources 1 and 2 were merged as follows: for each city in data source 1, the biomass features from the closest non-null grid point in data source 2 were assigned using the K-nearest-neighbors algorithm with K=1 and the Haversine distance as the similarity measure. The biomass measurements in data source 2 were aggregated into 1-degree cells to consider the variation in the area when moving from the equator towards the poles. Each cell therefore represents an area of approximately 100km2 which was considered a reasonable area of biomass to assign to a single city. Therefore, K=1 was chosen.

The spatial distribution of these outliers (Figure 4) shows they are ocean islands, far from land and from the nearest non-null data point in data source 2. Although such cities clearly have non-zero surrounding biomass, they are situated geographically in locations not covered by the resolution of the data source 2 measurements. For this reason, they were omitted from the merged data set (n=115 omitted).

Figure 4: Cities with Large Nearest Neighbor Distances




The data was split into three sets at random for use in analysis:

I Training set (70% of total) – to be used for training models

II Validation set (20% of total) – to be used for comparing different models

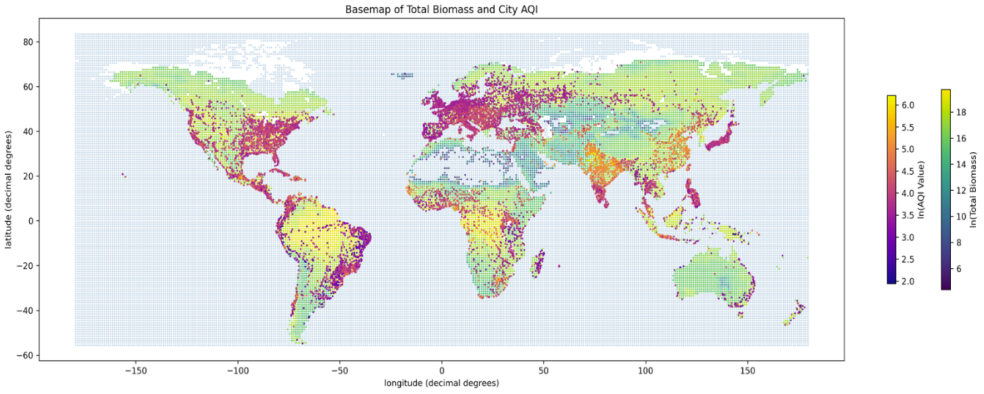
III Test set (10% of total) – to be used for reporting the performance of the final selected model

## 3 Methods

### 3.1 Exploratory Data Analysis

Figure 5 shows a heatmap for the natural log of 'Total Biomass' and 'AQI Value' across the globe. The purpose of this plot was to get a sense if cities with low AQI tend to be near geographical regions with high total biomass density. The plot does not show clearly if cities close to areas of high total biomass tend to have lower AQI. It is apparent that locations on the Earth with very high and very low total biomass tend to have few cities in their midst.

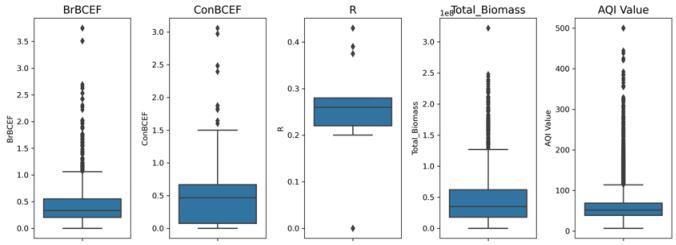
Figure 5: Base Map of Total Biomass and City AQI




### 3.1.1 Outlier Investigation

The presence of outliers was investigated visually and numerically by using a box plot and interquartile range (IQR) score respectively, for each variable. The outlier trend was found to be the same across all variables, so a subset of variables is presented in Figure 6, with the full set of plots included in the Appendix. Further investigation showed that outliers implied are just extreme values in the Earth's biomass distribution that occur naturally.

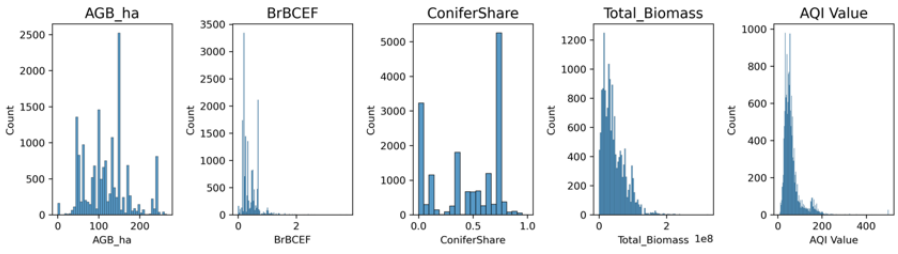
Figure 6: Outlier Detection by Box Plots




### 3.1.2 Univariate

A characteristic subset of the distributions of the continuous variables are shown as histograms in Figure 7. The full gallery for all variables is shown in the Appendix. Some of the variables are right skewed with long tails. Long tails would be expected given the distributions observed in the box plots.

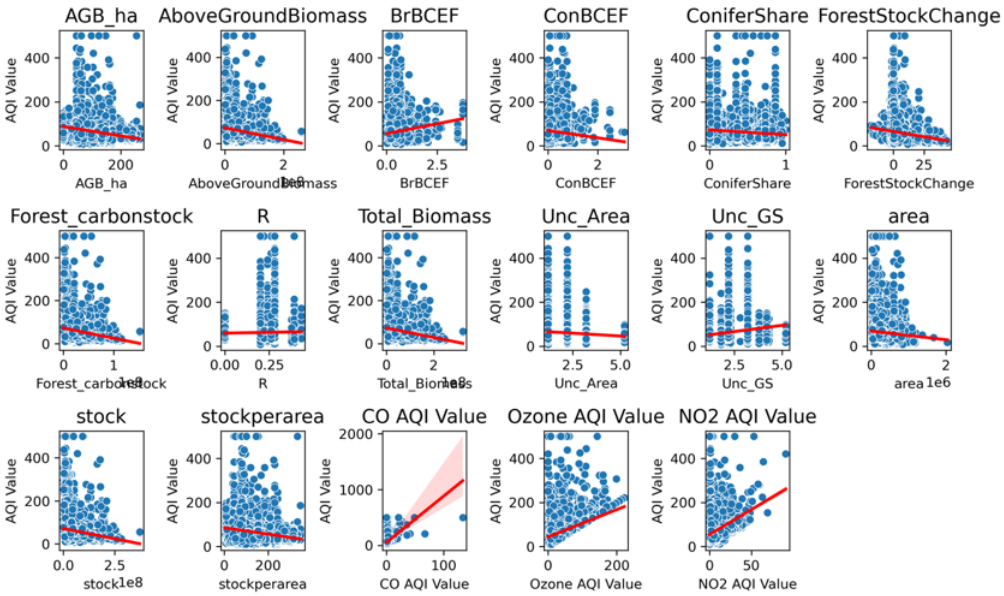
Figure 7: Distributions of a Characteristic Subset of the Continuous Variables



### 3.1.3 Bivariate

The relationships between each feature and the 'AQI Value' variable are shown in Figure 8. Different measures of AQI have a strong linear relationship as might be expected. There are no clear linear relationships between any of the other predictors and the 'AQI Value' variable.

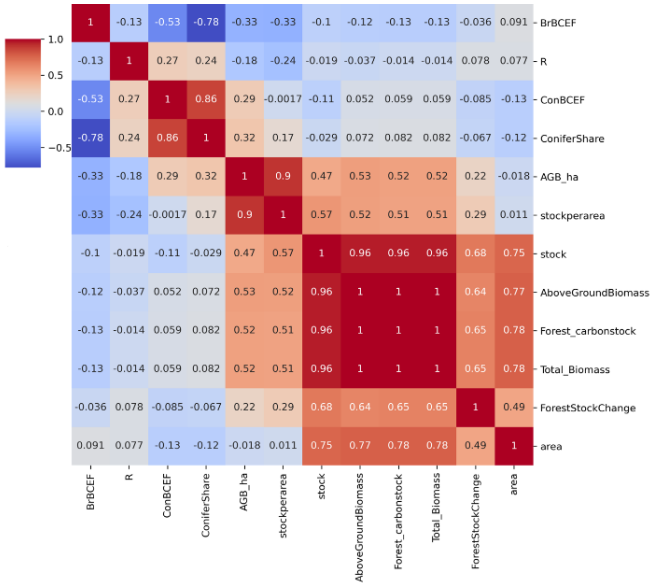
Figure 8: Bivariate Relationships Between Features and Dependent Variable 'AQI Value'




### 3.1.4 Multivariate

The correlation matrix for the biomass features is shown in 9. It is apparent that some variables are perfectly correlated (coefficient = 1). These variables are linear multiple of one another so it will be desirable to eliminate all but one prior to analysis. There are other correlation coefficients close to -1 and +1 indicating the presence of multicollinearity among the features. It will be necessary to deploy feature selection models such as LASSO and Ridge regression that can handle multicollinearity.

Figure 9: Biomass Feature Correlation Matrix



## 3.2 Regression Analysis

### 3.2.1 Initial Variable Selection

There are 12 features in the dataset, discounting the features that define each point in name and location ('Latitude', 'Longitude', 'City Name', and 'Country'), and discounting two features that describe uncertainty in the other measurements ('UNC_gs' and 'UNC_area'). Those features were discarded from the full dataset used for analysis. The correlation coefficients calculated during Exploratory Data Analysis showed many features have a high degree of collinearity. This was investigated further with a set of VIF calculations on the dataset. The VIF calculations showed that several factors had large VIF values. Several rounds of calculations were run, successively eliminating variables until all variables had a VIF below 4. See Figure 10.

Figure 10: VIF Values

```
     AGB_ha       BrBCEF       ConBCEF  ConiferShare            R Total_Biomass         area        stock  stockperarea
29.123647     3.485135    11.513794     10.762960     1.288291     43.480355     6.870534    43.634998     34.817718
     AGB_ha       BrBCEF       ConBCEF            R Total_Biomass         area        stock
 3.743432     1.720269     2.930854      1.236138     30.091533     6.553010    22.264437
     AGB_ha       BrBCEF       ConBCEF            R Total_Biomass         area
 3.738477     1.505862     1.614331      1.180216      8.967670     6.536881
     BrBCEF       ConBCEF            R Total_Biomass         area
 1.504726     1.563203     1.134745      2.950497      2.982856
```

The regression analysis models were run on two sets of features – the full set, and the reduced set as determined by the VIF calculations. This reduced set of features is: BrBCEF', 'ConBCEF, 'R', 'Total_Biomass', and 'area'.

### 3.2.2 Multiple Linear

Initially, we performed multiple linear regression on the full set of features in the dataset. Multiple linear regression is a predictive statistical model that uses more than one independent variable to predict a dependent variable. Linear regression models assume linearity, homoskedasticity, independence of errors, and independence within the independent variables. These assumptions were met during both the exploratory data analysis and the variable selection processes.

For comparison, we ran multiple linear regression models on both the full dataset and the trimmed dataset as determined by the variable selection process. The full model yielded 10 statistically significant coefficients out of the 12 in the model, while all coefficients in the trimmed model were statistically significant. See Figure 11. It should be noted that the variable Forest_carbonstock was removed due to singularities with Total_biomass, which caused coefficients to render undefined.

We also examined the $R^2$ values and mean-squared errors (MSE). The full model performed better in terms of $R^2$ and MSE. However, even with the worse metrics, we can assume the trimmed model is more accurate due to the multicollinearity included within the full model.

Figure 11: Multiple Linear Regression AQI Coefficients

| Feature | Full | Trimmed |
|---|---|---|
| Intercept | 94.86 | 63.00 |
| AGB_ha | -0.50 | -- |
| AboveGroundBiomass | 0.00 | -- |
| BrBCEF | -13.88 | 7.40 |
| ConBCEF | 4.17 | -10.96 |
| ConiferShare | -17.54 | -- |
| ForestStockChange | -0.31 | -- |
| R | 143.80 | 35.72 |
| Total_Biomass | 0.00 | 0.00 |
| area | 0.00 | 0.00 |
| stock | 0.00 | -- |
| stockperarea | 0.05 | -- |

Red indicates statistical significance at 90% confidence.

From here, the multiple linear regression model was used as a baseline to compare the other regression techniques and performances against.

### 3.2.3 Stepwise

Stepwise regression is a form of multiple linear regression that performs automatic feature selection. Forward stepwise regression starts with no features and incrementally adds features that improve the quality of the model, as measured by the Akaike Information Criterion (AIC). Backward stepwise regression starts with a model containing all features and

incrementally removes features with the least contribution to the model quality as measured by AIC. Forward-backward stepwise regression is a combination of the two.

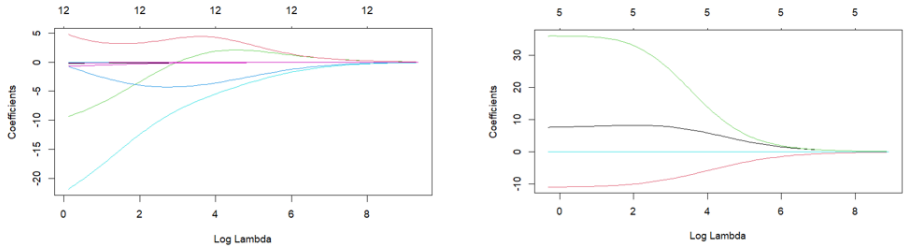
For the initial stepwise regression, 'AQI Value' was used as the dependent variable, and all features were input into the backward regression. Forward, backward, and forward-backward regressions were run.

For the forward stepwise regression, all variables except 'stockperarea' were found to improve the model. For the backwards regression model, 'stockperarea' and 'ConiferShare' were both eliminated. For the forward-backward regression, only 'stockperarea' was not used in the final model. The one thing these three models agreed on is that 'stockperarea' is not an important factor. The adjusted R-squared values for each model were not particularly good and are reported in Figures 17 and 18. An additional set of regressions were run for the trimmed set of variables. All of the regressions with the trimmed datasets performed worse, both in terms of MSE and R-squared.

### 3.2.4 Ridge

Ridge Regression is a regularization method used in multiple linear regression that focuses on tuning the model and shrinking the coefficients towards zero as a method of reducing the bias and complexity of models. Coefficients cannot be shrunk to exactly zero so ridge regression cannot perform direct feature selection. Ridge regression initially was performed on the full training set, with AQI_Value as the dependent variable. To find the optimal lambda, k-fold cross-validation was performed to find the lambda value where MSE is minimized (see Appendix), and once the optimal lambda was found, it was used to perform ridge regression on the full validation set. Similarly, this was performed on the feature-selected model. Figure 12 visualize the shrinkage of the coefficients in each model. The coefficients with the steepest curves tend to be less meaningful than those that level out. Overall, as in the multiple and stepwise regression models, the trimmed model's $R^2$ is lower than that of the full model, though the full model contains multicollinearity and may have inflated predictive power.

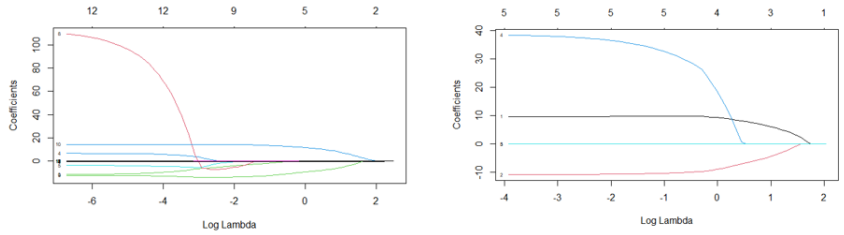
Figure 12: Ridge Regression Coefficients vs Log($\lambda$)




### 3.2.5 LASSO

Similar to Ridge, LASSO Regression is a regularization model used to simplify the model, but LASSO is also a method of variable selection, as LASSO forces select coefficients to zero, whereas Ridge shrinks coefficients towards zero but never exactly to zero. LASSO was performed on both the full training set of variables and the trimmed set of variables, and after performing k-fold cross-validation to find the optimal lambda (see Appendix), the models were run through again using the respective validation sets, with the resulting coefficient trace plot below (see Figure 13). Comparing the trace plots with those in Ridge seems to suggest that LASSO performs slightly better than Ridge, as the coefficients, save for one, have marginally shallower curves in the plot than in Ridge. The full-feature model yielded an R2 = 0.1732 and the feature-selected model a lower R2 value of 0.0448, which is similar to the results for Ridge.

Figure 13: LASSO Regression Coefficients vs Log($\lambda$)



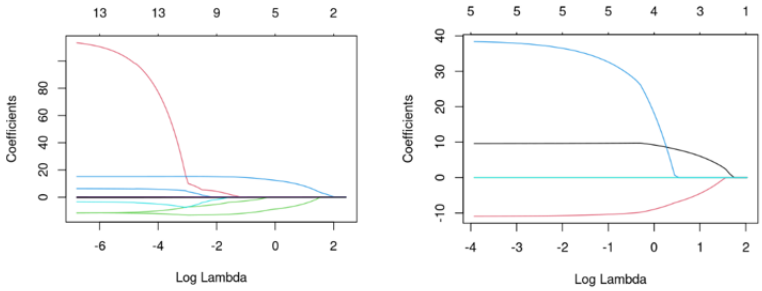
### 3.2.6 Elastic Net

Elastic net is a regularization method that combines the abilities of Ridge and LASSO and also uses k-fold cross-validation to find an optimal level of lambda (see Appendix).

The best results for the full-feature elastic net with 'AQI_Value' as the dependent variable came when our alpha value was equal to 1 (i.e., LASSO regression). As alpha decreased, R2 continued to decrease from the 0.1732 value as seen in the LASSO regression (see left chart) to the 0.1530 value as seen in the Ridge regression.

The results for the feature-selected model using variables 'BrBCEF', 'ConBCEF', 'Total_Biomass', 'R', 'area' as independent variables and 'AQI_Value' as the dependent variable were nearly identical across the range of alpha values, with all R2 values equal to 0.0489. This is significantly smaller than the R2 value of the full-feature model, but as previously mentioned, the full-feature model may be inflated by multicollinearity. In looking at Figure 14, the coefficient angles more-closely resemble those with LASSO.

Figure 14: Elastic Net Regression Coefficients vs Log($\lambda$)




### 3.2.7 Random Forest

Random forest regression is a non-parametric supervised learning method. A random forest is created using an ensemble of decision trees, which are trained on bootstrapped samples from the training data. Each tree learns decision rules from the independent variables in the training data. The final model predicts values for a continuous dependent variable by averaging the predictions of the individual trees. The ensemble approach helps to prevent overfitting which improves the robustness of the model (i.e., its ability to perform well on unseen data).

The random forest regressor is a flexible model which does not rely on the same assumptions as linear regression, namely: residuals need not follow a normal distribution (non-parametric), heteroscedasticity in the dependent variable is permissible, and a non-linear relationship between the independent variables and the dependent variables is permissible. The random forest regressor's predictive power is also robust to multicollinearity in the features.

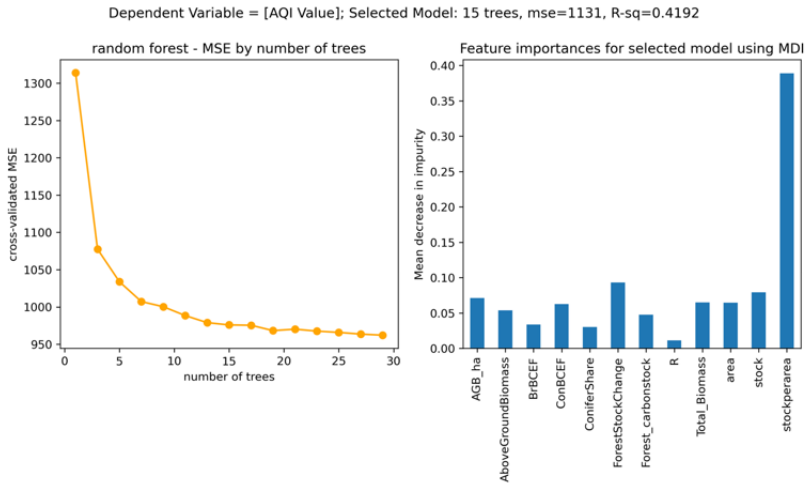
The main hyperparameter to be tuned was the number of decision trees to use to create the 'forest'. An appropriate number of trees was selected by inspection of the plot of MSE against number of trees using an 'elbow method' approach. As the number of trees increases, the MSE decreases. The magnitude of the change is large at first but begins to taper off as the number of trees increases. The point at which the reduction in MSE becomes marginal is the 'elbow', and the number of trees corresponding to the elbow point was the final number used.

Feature importance was measured using 'mean decrease in impurity' (MDI). Impurity is a measure of the variance in the prediction of the dependent variable. The MDI for a feature is a measure of the reduction in impurity achieved each time that feature is used to split the data in a tree, aggregated across all trees in the forest. The higher the MDI, the more that feature reduces impurity, and therefore the more important that feature was for prediction. Several random forest models were built using the following general procedure:

1. Select a set of independent variables, and a dependent variable.
2. Using the training data, train models with number of trees from one to 30 in increments of two.
3. For each model, calculate the MSE using k-fold cross-validation on the training data.
4. Choose an appropriate number of trees for the 'selected model' using the elbow method.
5. Calculate the MSE and R-squared value for the selected model using the validation data.

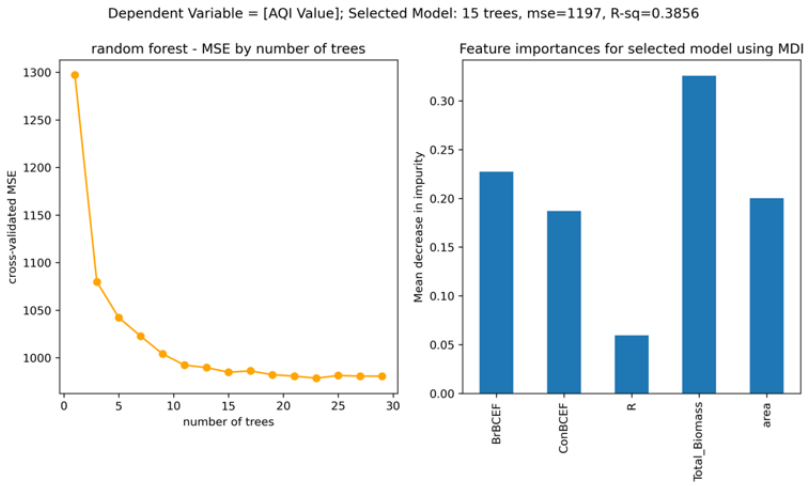
Model 1 was built using all features and 'AQI Value' as the dependent variable – results are shown in Figure 15. For the selected model, the MSE was 1131 and the R-squared value was 0.4192, meaning this model predicted 41.92% of the variance in the 'AQI Value' variable. The most important feature was 'stockperarea'.

Figure 15: Random Forest Model 2




Model 2 was built using the subset of features derived from the VIF analysis; results are shown in Figure 16. For the selected model, the MSE was 1197 and the R-squared value was 0.3856. The most important feature was 'Total_Biomass'. A slight reduction in variance explained was observed after eliminating seven features. However, Model 2 is more reliable for inferring the relative importance of the predictors because the unimportant colinear features were already removed.

Figure 16: Random Forest model 2




## 4 Results

Models were created for different dependent variables to test the research questions outlined in the Overview section. Each model was trained on the training data using the set of features derived from VIF analysis. The R-squared and MSE values were calculated on the validation data – refer to Figures 17 and 18. The random forest is a non-linear model; all others are linear.

The linear models have R-squared values between 0.0016 and 0.0596, which are considered very low. These models explain less than 6% of the variance in the range of the AQI measure variables. The random forest model performs better, with R-squared values between 0.1063 and 0.4524 (explaining between 10.63% and 45.24%) of the variance. The most successful model was the random forest.

Figure 17: Model $R^2$ by Dependent Variable

| Regression Model | Model R-squared by Dependent Variable | | | | |
|---|---|---|---|---|---|
| | AQI | CO AQI | Ozone AQI | NO2 AQI | PM2.5 AQI |
| Multiple Linear | 0.0448 | 0.0027 | 0.0501 | 0.0563 | 0.0435 |
| Forward Stepwise | 0.0444 | 0.0023 | 0.0497 | 0.0559 | 0.0431 |
| Backward Stepwise | 0.0444 | 0.0023 | 0.0497 | 0.0559 | 0.0431 |
| Forward-Backward Stepwise | 0.0444 | 0.0023 | 0.0497 | 0.0559 | 0.0431 |
| Ridge | 0.0480 | 0.0016 | 0.0545 | 0.0596 | 0.0475 |
| LASSO | 0.0489 | 0.0030 | 0.0581 | 0.0634 | 0.0442 |
| Elastic Net | 0.0489 | 0.0117 | 0.0513 | 0.0542 | 0.0449 |
| Random Forest | 0.3856 | 0.1063 | 0.4524 | 0.3153 | 0.3775 |

Red indicates highest $R^2$ value for each dependent variable.

Figure 18: Model MSE by Dependent Variable

| Regression Model | Model MSE by Dependent Variable | | | | |
|---|---|---|---|---|---|
| | AQI | CO AQI | Ozone AQI | NO2 AQI | PM2.5 AQI |
| Multiple Linear | 1751 | 5 | 497 | 32 | 1763 |
| Forward Stepwise | 1796 | 2 | 518 | 36 | 1807 |
| Backward Stepwise | 1796 | 2 | 518 | 36 | 1807 |
| Forward-Backward Stepwise | 1796 | 2 | 518 | 36 | 1807 |
| Ridge | 1852 | 8 | 480 | 33 | 1859 |
| LASSO | 1852 | 7 | 487 | 30 | 1843 |
| Elastic Net | 1853 | 2 | 538 | 34 | 1722 |
| Random Forest | 1197 | 7 | 278 | 24 | 1215 |

Red indicates lowest MSE value for each dependent variable.

Using the test data set to predict on 'AQI', the random forest R-squared value and MSE, were 0.3753 and 1217 respectively. The mean absolute error (MAE) evaluated on the test set was 19.55, meaning on average the model's predictions were 19.55 units away from the true values. From Figure 6, the mean of 'AQI Value' is approximately 50, so the average deviation as measured by MAE is substantial.

Primary research question: can the AQI of a city be predicted by the amount of vegetative biomass in its vicinity?

The results show that the AQI measures cannot be modelled as a linear combination of the local biomass features. The non-linear random forest has reasonable predictive power, but the biomass indicators alone are insufficient to offer a satisfactory explanation of the variance in the AQI measures. The model could likely be improved by the inclusion of other non-biological features such as: population density, industrial output, and policy and regulation.

Supplementary research question: which biomass indicators are the strongest predictors of AQI?

The VIF analysis showed that 'BrCEF', 'ConCEF', 'R', 'Total_Biomass', and 'area' are the most important input features. For the random forest model, 'Total_Biomass' was the most important feature for predicting all 'AQI' measures except 'Ozone AQI', for which 'BrCEF' was the most important feature. Refer to the Appendix for bar plots of the MDI for the different models.

Supplementary research question: are different measures of air quality related to different biomass indicators?

The relative importance of features was similar for 'AQI', 'CO AQI', 'NO2 AQI', and 'PM2.5 AQI', but different for 'Ozone AQI'. This suggests that 'Ozone AQI' is driven by different biomass factors. Refer to the Appendix for bar plots of the MDI for the different models .

Looking forward, businesses could benefit from the selected random forest regression model as it is more fine-tuned in the future, especially as further research allows for other impactful variables to be included. In doing so, even as-is, with approximately 37.53% of the variability in the data, this model could be insightful in businesses planning movements and policies to increase no only their public image but their overall profits and growth.

## References

[1] Babu Saheer, L., Bhasy, A., Maktabdar, M., & Zarrin, J. (2022). *Data-driven framework for understanding and predicting air quality in urban areas*. Frontiers in Big Data, 5. https://doi.org/10.3389/fdata.2022.822573

[2] Daly, C. (2022, November 10). *How does air pollution affect businesses?*. Clean Air Fund. https://www.cleanairfund.org/news-item/how-does-air-pollution-affect-businesses

[3] Hengeveld, G.M., K. Gunia, M. Didion, S. Zudin, A.P.P.M. Clerkx, and M.J. Schelhaas. (2015). *Global 1-degree Maps of Forest Area, Carbon Stocks, and Biomass, 1950-2010*. ORNL DAAC, Oak Ridge, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1296

[4] Nowak, D. J., Hirabayashi, S., Bodine, A., & Greenfield, E. (2014). *Tree and forest effects on air quality and human health in the United States*. Environmental Pollution, 193, 119–129. https://doi.org/10.1016/j.envpol.2014.05.028

[5] Nowlan, A. (2018, November 5). *3 reasons why air pollution should be a top priority for businesses*. EDF+Business. https://business.edf.org/insights/3-reasons-why-air-pollution-should-be-a-top-priority-for-businesses/

[6] Ramachandran, A. (2023, May 7). *World Air Quality index by city and coordinates*. Kaggle. https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates

# 5 Appendix

Below are figures to support the Exploratory Data Analysis:

Figure 19: Boxplots for Numerical Variables

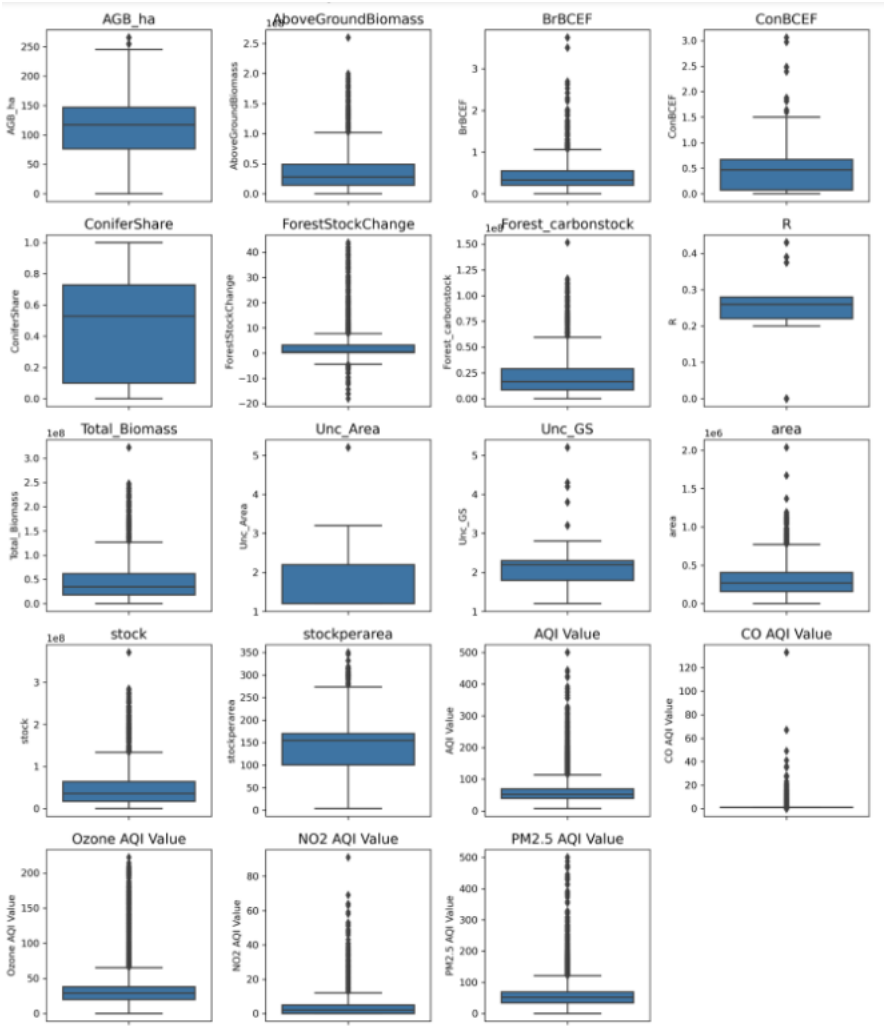
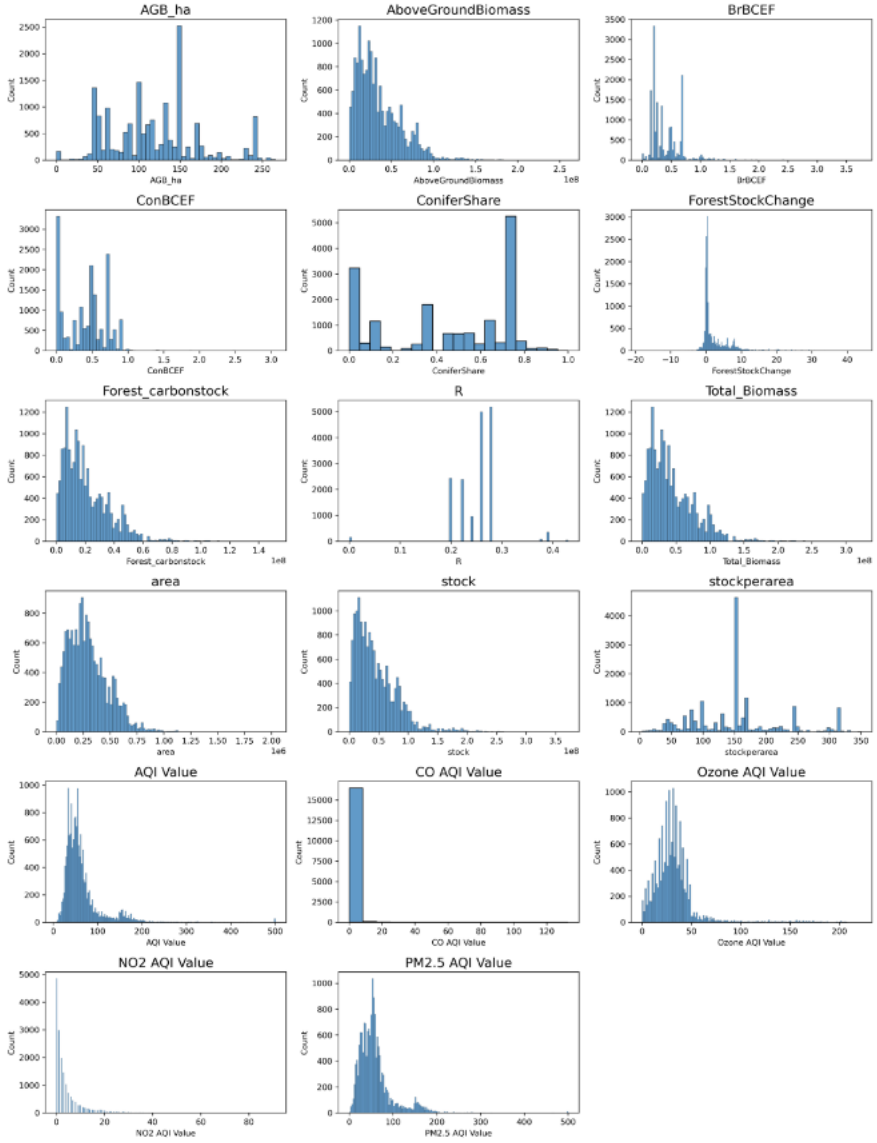


Figure 20: Distributions of Continuous Variables



Below are figures to demonstrate cross-validation for Ridge regression:

Figure 21: Ridge MSE for AQI

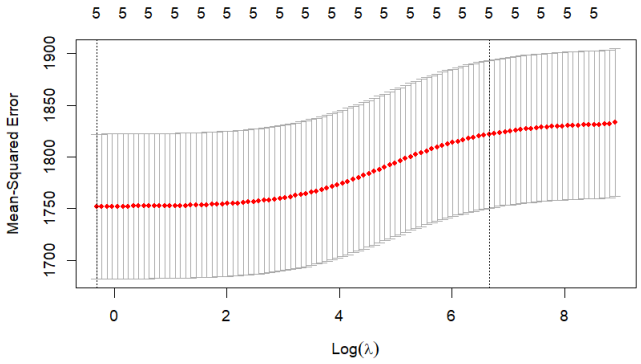



Figure 22: Ridge MSE for CO

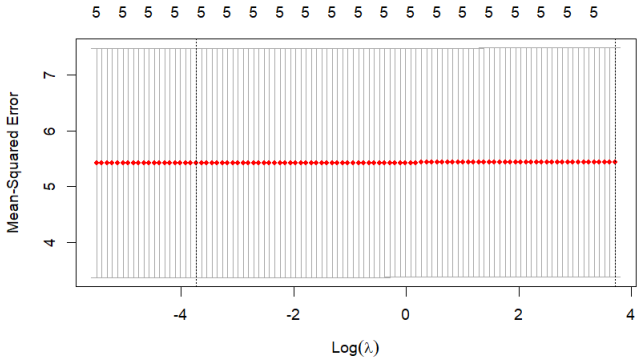



Figure 23: Ridge MSE for Ozone

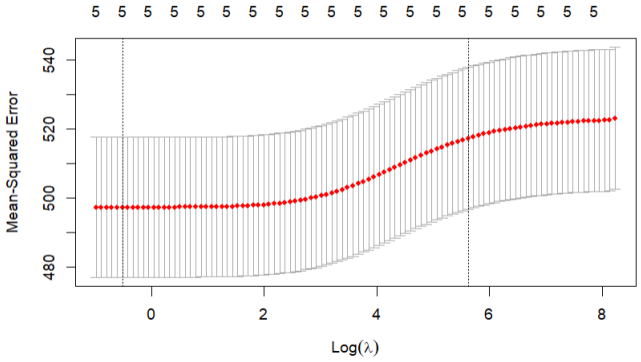


Figure 24: Ridge MSE for NO2

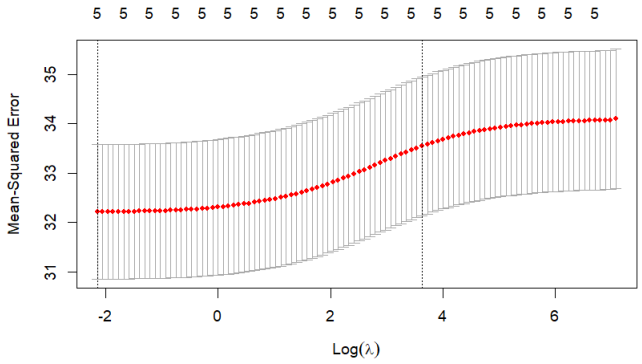
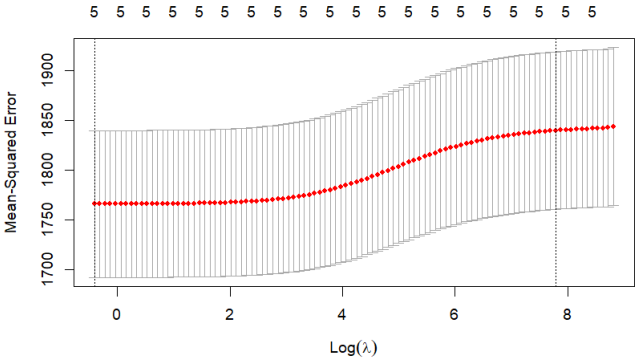



Figure 25: Ridge MSE for PM2.5




Below are figures to demonstrate cross-validation for LASSO regression:

Figure 26: LASSO MSE for AQI

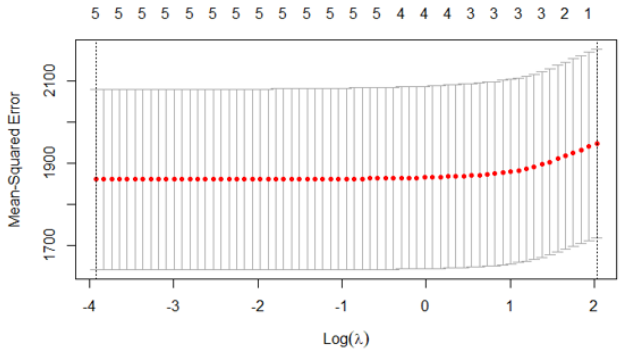


Figure 27: LASSO MSE for CO

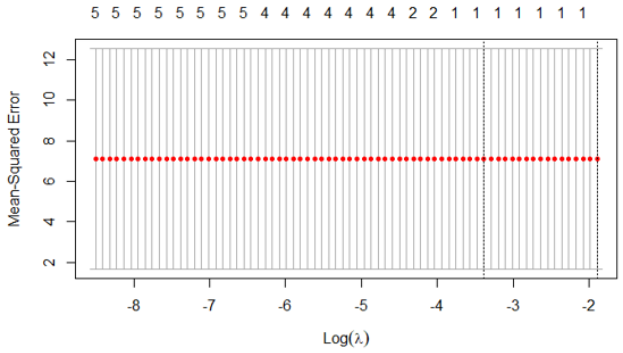



Figure 28: LASSO MSE for Ozone

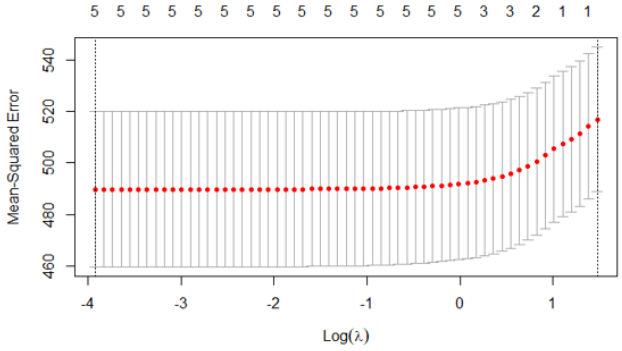



Figure 29: LASSO MSE for NO2

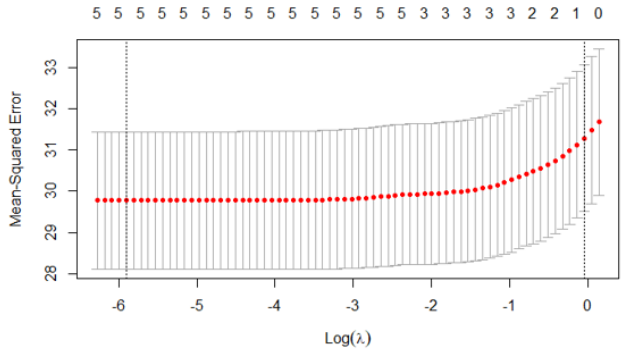
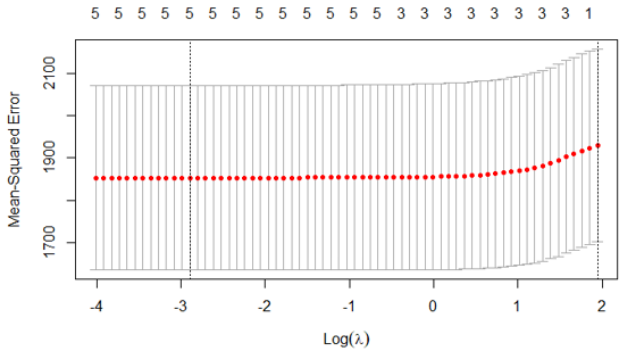


Figure 30: LASSO MSE for PM2.5




Below are figures to demonstrate cross-validation for Elastic Net regression:

Figure 31: Elastic Net MSE for AQI

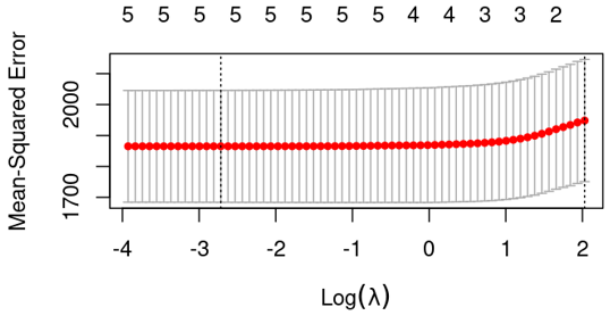



Figure 32: Elastic Net MSE for CO

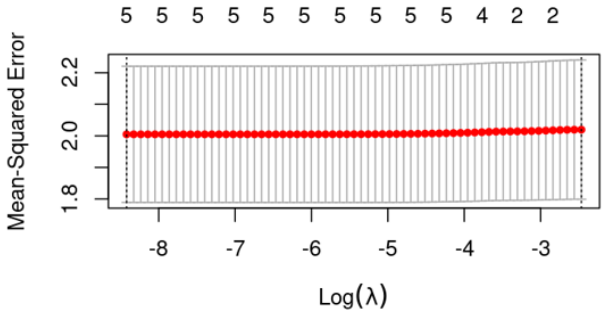


Figure 33: Elastic Net MSE for Ozone

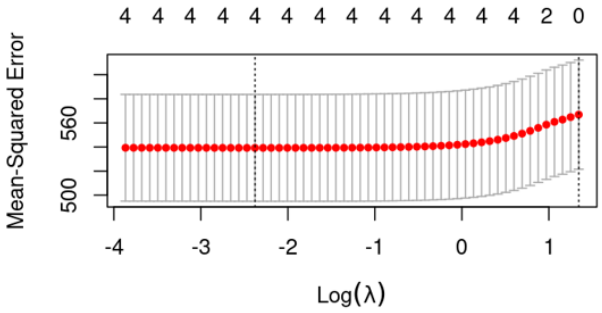



Figure 34: Elastic Net MSE for NO2

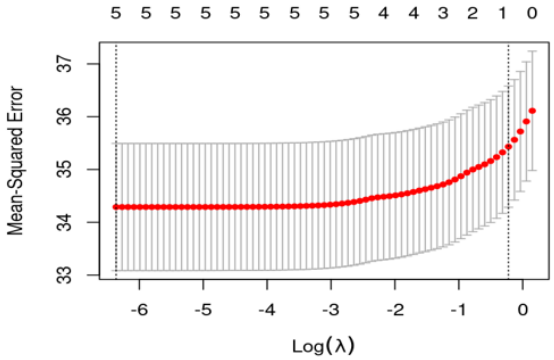
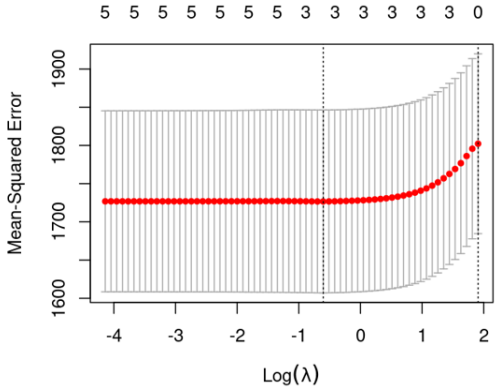



Figure 35: Elastic Net MSE for PM2.5



Below are figures to support the Random Forest Regression:

Figure 36: Random Forest Model for CO

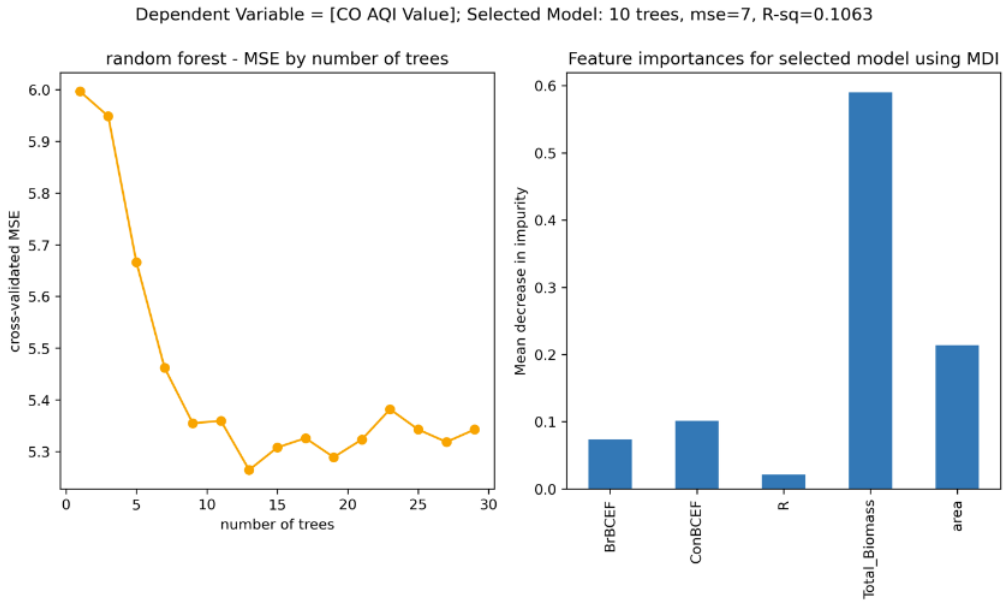



Figure 37: Random Forest Model for Ozone

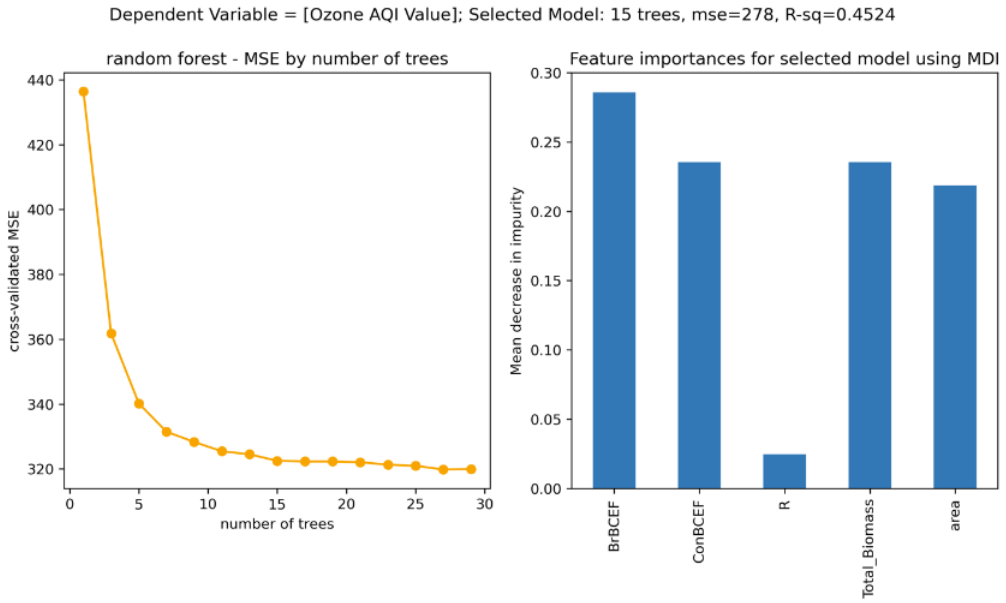


Figure 38: Random Forest Model for NO2

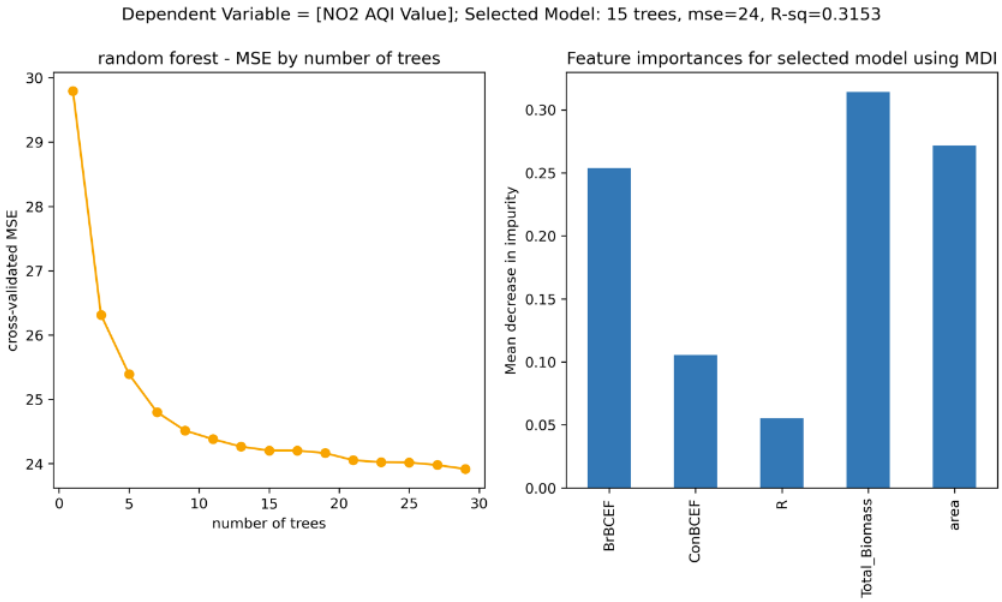



Figure 39: Random Forest Model for PM2.5