Breadth of Life:
Trees as Bioindicators of Air Quality

Clara Chen Harvard ‘25
Elizabeth Zhong Harvard ‘25
Helen Xiao Harvard ‘25
Jessica Li Harvard ‘25

Abstract

This study investigates the relationship between urban tree characteristics and air quality indicators across 46 cities. We employ various regression models to analyze the predictive power of tree-related variables on different Air Quality Index (AQI) values.

Our dataset is comprised of the number of trees, number of unique species, proportion of naturally occurring versus introduced trees, and tree density, alongside overall AQI and specific values for CO, Ozone, NO2, and PM2.5.

Our analysis includes a comprehensive suite of regression models: a full linear model as the baseline, a full interaction model, stepwise models with interaction terms (forward, backward, and both), LASSO, Ridge, decision trees, and random forest. All models log-transform the response variable to address the right-skewed distribution of AQI values. This research aims to elucidate the extent to which urban tree characteristics can serve as bioindicators of air quality, contributing valuable insights to urban planning and environmental management.

1 Introduction

Urban trees are increasingly recognized for their environmental and public health benefits, particularly in mitigating air pollution. The capacity of trees to act as bioindicators of air quality forms the cornerstone of this study. Trees interact with air pollutants in complex ways, potentially influencing the concentration and distribution of pollutants in urban areas.

1.1 Motivation

Urban trees serve as vital components of urban ecosystems, offering a multitude of environmental and public health benefits. Among these, their role in air quality improvement is particularly significant, especially in the context of increasing urbanization and associated air pollution challenges. This research delves into the complex interactions between urban trees and air pollutants, aiming to quantify and understand these relationships better. Understanding the relationship between urban trees and air quality is crucial for urban planners, environmentalists, and policymakers. This study aims to provide empirical evidence to guide effective urban forestry practices and environmental policies to improve air quality and public health.

Studying the relationship between urban trees and air quality is important for many reasons. Trees can absorb pollutants like carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), and particulate matter (PM2.5), which are common in urban environments. By studying these interactions, we can better understand how different species and densities of trees contribute to reducing pollution levels and mitigating climate change. Additionally, insights from this study could guide urban planning, public health policy, and environmental policy by informing decisions about where and what types of trees to plant for optimizing the benefits for air quality. Finally, we hope that our statistical study can lay the groundwork for further research in environmental science, particularly in exploring the synergies between urban greening and pollution control strategies.

1.2 Objective

Our primary objective is to assess the inferential relationship between various tree characteristics in urban environments and air quality, as measured by different AQI values. This approach allows us to explore the potential of trees as natural mitigators and indicators of urban air pollution. Due to the small size of our dataset, our goal is to use regression models to help us assess whether or not there are associations between predictor and response variables rather than trying to build models that make accurate predictions.

2 Data and EDA

This section describes the data used in our study, discussing both predictor (tree data) and response variables (air quality indicators) as well as the rationale behind

their selection. We will also include in this section the data cleaning and transformation we did prior to conducting our analysis.

2.1 Data Overview

We required two datasets for our project: a dataset of tree data and a dataset of air quality data that occurred in similar cities as the tree dataset. We ultimately sourced a tree dataset from Dryad, an international open-access repository of research data, and an air pollution dataset sourced from Kaggle, an online community of data scientists and machine learning practitioners under Google LLC.

2.2 Data Selection

The air pollution dataset contained 23464 cities from all around the globe, with 2872 cities from the United States. For each city, there were recorded values for AQI, CO AQI, Ozone AQI, NO2 AQI, and PM2.5 AQI. The tree dataset only contained 63 cities, all in the US, so we simply chose to use the data where cities were shared between the tree data and the air pollution data. The tree dataset contained much fewer cities, but many more columns and values per city. We eventually narrowed down to only using the columns `scientific_name`, `native`, `longitude_coordinate`, and `latitude_coordinate` from the tree data, which we used to calculate the number of trees observed, the number of unique species of trees, the proportion of native trees (as opposed to introduced trees), and the density of trees in the observed area per city. We felt that these values were both feasible to derive from the data and might be most useful for us in gauging the relationship between trees and air pollution in cities.

2.3 Metrics for Air Pollution

To measure levels of air pollution in cities, we look into common pollutants, including nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), and particulate matter (PM2.5).

Nitrogen Dioxide, a gas naturally introduced into the air also forms from cars, trucks and buses emissions, power plants and off-road equipment. Extensive exposure can worsen respiratory conditions, like asthma, or contribute to development of asthma and respiratory infections.

Ozone is created by chemical reactions between oxides of nitrogen and volatile organic compounds (VOC), and differs from the "good" ozone located in the upper atmosphere, which protects the earth from UV rays. Ground level ozone, in particular, can lead to respiratory health problems, such as reducing lung function and worsening bronchitis, emphysema, and asthma. The Ozone molecule affects also negatively impacts vegetation and ecosystems by damaging sensitive vegetation during the growing season.

Carbon Monoxide, like nitrogen dioxide, is released into the air by cars, trucks and other vehicles or machineries that burn fossil fuel. Breathing in CO can severely reduce the amount of oxygen transported in the blood stream to critical organs like the heart and brain.

Atmospheric **Particulate Matter**, unlike the previous pollutants, are mixtures of solid and liquid matter in the air. They are classified as group 1 carcinogen by the International Agency for Research on Cancer (IARC). PM2.5 specifically refers to those particles with a diameter of 2.5 micrometers or less.

The **overall AQI** is calculated by taking the maximum of the other four air quality values. This is typically PM2.5 for most cities with the exception of a few that have Ozone as the maximum instead.

2.4 Variable Summary and Hypothesized Importance

The first variable we chose to include in our final dataset is the number of trees since the number of trees is an integral factor in mitigating air pollution. In order to calculate the number of trees in each city, we found the number of rows in each city's dataset. A greater number of trees in an area means a larger collective surface area that can absorb and filter air pollutants such as carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), and particulate matter (PM2.5) and so, we hypothesized that number of trees and AQI are negatively correlated.

The city datasets included a column specifying the species of each tree and so, we theorized that the number of unique tree species in a particular city could have an interesting correlation with that city's AQI value. Different tree species have varying abilities to absorb and process different pollutants. For example, some species are more effective at absorbing specific gasses like CO2, NO2, or O3, while others are better at intercepting particulate matter. A diverse array of trees ensures a broader range of pollutants can be mitigated, but at the same time, a diverse tree ecosystem could also suggest that there is not a strong correlation with any of the AQI values since it does not specialize in reducing any one of the pollutants measured by the AQI.

Naturally occurring trees, being native to the area, are generally well-adapted to the local climate, soil conditions, and ecological interactions and therefore, are often more efficient in using local resources such as wa-

ter and nutrients compared to non-native species. This efficiency can translate into better growth and larger biomass, which is beneficial for air pollution absorption as larger trees with more leaves can capture more pollutants. Because of this, we wanted to explore whether the proportion of naturally occurring trees could provide interesting insights on correlation with air quality. Our original city datasets included data about whether each tree was native to that region or introduced, so we calculated the proportion by dividing the number of naturally occurring trees by the total number trees with non-NA values for that column.

Similar to the number of trees, the density of trees in a city would also have a significant impact on air quality as a higher density implies a greater leaf surface area. Further, the density of trees accounts for the area from which the trees are sampled and can provide a more insightful analysis into the impact of trees on air quality. Overall, we hypothesize that density and the AQI values will be negatively correlated.

A consideration into our four predictors is that the original tree datasets do not necessarily sample all trees present in a city, meaning that our values for number of trees, number of unique tree species, proportion of naturally occurring trees, and tree density might not accurately reflect a city's actual density. Additionally, the number of trees and density of trees might be correlated, especially given that we use the number of trees in our calculation of density. We theorize that this will affect our results.

2.5 Data Cleaning

In our data cleaning efforts, we first selected only cities, where there were no NA values in the specified 4 columns (some datasets contained only NA values in certain columns). After isolating these cities, we ended up with 46 cities.

Next, the tree dataset actually came as 63 CSV files, one for each city. In each CSV file was the recorded data of all trees in that city. We do not care about the specific trees within an area or the data of the specific trees. Instead, we cared about summary statistics for a given city such as the "density" of trees and the proportion of naturally or introduced trees.

To do this, we created a function that looks at each city dataset and adds the city name, number of trees sampled, number of unique tree species, proportion of naturally occurring trees, observed area, and density of the observed area to a singular dataframe.

In our density calculation, we had some difficulty converting latitude and longitude to area. We tried a variety of packages and calculations to convert between coordinates and area, but some produced a lot of NA values. Our final cleaning procedure involves converting the coordinates into distance from the minimum latitude/longitude coordinates in a city, then using a function which finds the area of the polygon formed by the convex hull of the points. Some of these area values didn't seem to line up with what we would expect based on the known size of the cities. After further inspection of the raw data for a few cities, it seems that the calculated area is highly dependent on how spread out the trees are. Stockton, for example, had very high area values and has trees that are grouped into two very spread out clusters while a city like Tampa, which had a more reasonable area value, has trees evenly spread out in the range of longitude and latitude values (see Appendix Figures 5 and 6).

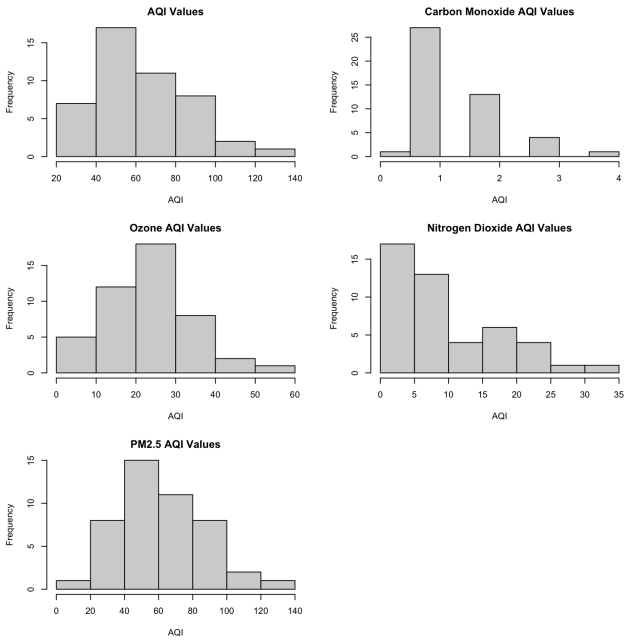

Figure 1: Response variable distributions

2.6 Data Exploration

2.6.1 Response Variable Distributions

Figure 1 shows histograms of each of the response variables considered, including overall AQI, carbon monoxide (CO), ozone, nitrogen dioxide (NO2), and fine particulate matter (PM 2.5) for the cities included in our final combined dataset. The histograms for carbon monoxide and nitrogen dioxide are very right skewed. Since linear regression models assume normality, we will be log transforming the response variable for the models predicting CO and NO2.

Additionally, while most air quality values span a large range and take on many distinct values, the car-

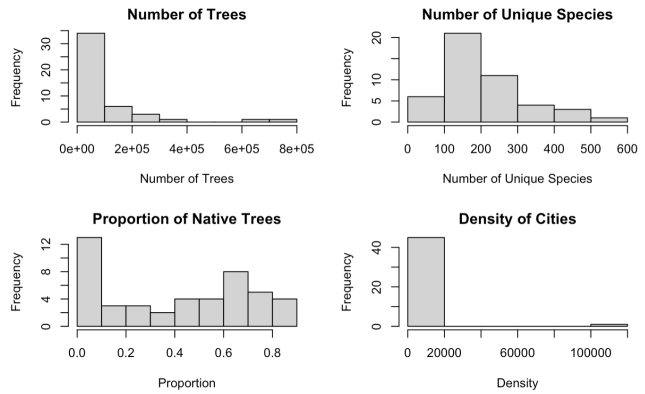

Figure 2: Predictor variable distributions

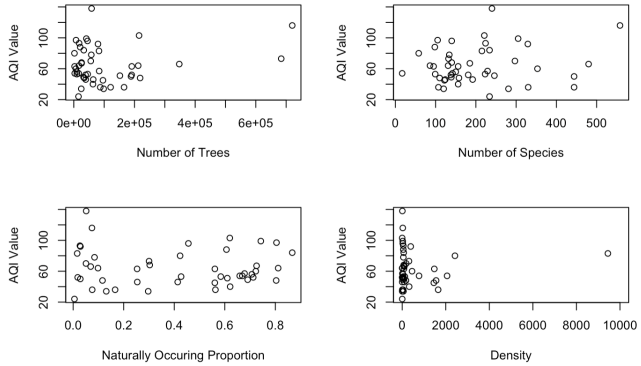

Figure 3: Scatter plots of tree predictors against AQI

bon monoxide values only seem to take on integers from 0 to 4. Carbon monoxide rarely occurs in large quantities outdoors, which may explain why the scale is much smaller for CO than the other air quality variables.

2.6.2 Basic Predictor Variable Statistics

Figure 2 shows histograms of each of the predictors variables – including number of trees, number of unique species, proportion of naturally occurring trees, and density. The number of trees and the density variables are right skewed. In particular, for density all the observations have density below 20000 with the exception of only a few outliers. This may be due to the complications in calculating density mentioned in the data cleaning section. The number of unique species and the proportion of native species are also slightly right skewed, but not as severely. Although the skew of these variables indicates that we may want to transform them, in order to maintain the interpretability of our models, we will not be transforming any of the predictors especially since some of our response variables will be transformed.

2.6.3 Exploratory Data Analysis Plots

Before we start to fit regression models, we want to get a sense of the relationship between the tree and air quality variables. Figure 3 shows each of the predictor tree variables plotted against AQI values. Additional figures showing the relationship with the other response variables can be found in the appendix (Figures 7-10). There seems to be some weak correlations between the variables, but no strong relationships between any individual tree variables and pollution at a first glance. This motivates the need to test more complex models to see if there are certain combinations of tree predictors that can predict air quality better than one predictor alone. The different regression techniques might also highlight some correlations or interactions that cannot be seen from the scatter plots alone.

The density outlier is also very apparent in the scatter plots, which could potentially skew our analysis. However, given that our dataset is already very small, we did not want to remove this outlier as that would decrease our sample size, as well as our inferential and predictive power, even more.

3 Models

3.1 OLS Regression Models

For each of the OLS Regression Models below, we regressed each of the different air quality index values as the response variable against the various tree data predictor variables.

1. `baseline` is a linear model that takes in every predictor variable for trees: `NumTrees`, `NumUniqueSpecies`, `NaturalProportion`, and `Density`. These predictors were used in five models for each of the outcome variables `AQI.Value`, `CO.AQI.Value`, `Ozone.AQI.Value`, `NO2.AQI.Value`, `PM2.5.AQI.Value`. At its core, baseline is simply a standard linear model with multiple predictor variables. Due to a skew in the outcome variables `CO.AQI.Value` and `NO2.AQI.Value`, we substituted in the linear model formula `log(CO.AQI.Value)` for `CO.AQI.Value` and `log(NO2.AQI.Value)` for `NO2.AQI.Value`. The linear model formulas were of the form `lm(AQI.Value~NumTrees+NumUniqueSpecies+NaturalProportion+Density,data=data)`

2. `interaction` is a linear model taking in every predictor variable for trees as well as all two-way interactions between the predictor variables, and outcome variable as each air quality index value. Again, due to a skew in the outcome variables `CO.AQI.Value` and `NO2.AQI.Value`,

we substituted in the linear model formula `log(CO.AQI.Value)` for `CO.AQI.Value` and `log(NO2.AQI.Value)` for `NO2.AQI.Value`. The linear model formulas were of the form `lm(AQI.Value~(NumTrees+NumUniqueSpecies+NaturalProportion+Density)^2,data=data)`

3. `forward` is a step-wise linear model. The lower model is an intercept model of the form `lm(AQI.Value 1, data = data)` where we start the step-wise iteration from.

4. `backward` is a step-wise linear model. The upper model is the two-way `interaction` model, where we start the step-wise iteration from.

5. `both` is a step-wise linear model. The lower model is an intercept model of the form `lm(AQI.Value 1, data = data)` and the upper model is the two-way `interaction` model. We start the iteration at the `baseline` model.

3.2 Penalized Linear Models

1. `baseline_ridge` is a penalized linear model. At the lambda_min value, we cross validated the model reporting the train RSME and test RMSE. We will use these RMSE values computed for each outcome variable against the RMSE values for the LASSO model.

2. `baseline_LASSO` is a penalized linear model. At the lambda_min value, we cross validated the model reporting the train RSME and test RMSE. We will use these RMSE values computed for each outcome variable against the RMSE values for the Ridge model.

3.3 Tree-based Regression Models

In our exploratory data analysis, we did not see strong linear relationships between the predictor and response variables. A tree-based regression model may serve as a suitable alternative in this case, since it takes a non-parametric approach that does not assume linearity or other underlying patterns in the data. Because of the small size of our data, our primary objective in fitting these tree-based models is to determine the importance of our predictor variables in predicting air quality rather than creating models that would work well on predicting out-of-sample data. For each response variable (AQI, CO, Ozone, NO2, and PM2.5), we fit a decision tree and a random forest.

1. Decision trees: Initially, the decision tree parameters were found through cross validation, but this often led to trees with no splits. Therefore, each final decision tree had parameters `cp=0, maxdepth=5`. This allowed for maximal complexity to give us more information about how the tree predictors relate to each pollution variable.

2. Random forest: For each random forest, we used 3-fold cross validation to tune the `mtry` parameter. The optimal `mtry` based on RMSE was used for the final model.

4 Results

4.1 OLS Regression Models

After fitting our linear models to predict each of the five air pollution variables, we summarized our findings in Tables 1-5.[1] For each of the five AQI variables, we show the formula used for each linear model, the p-value, RMSE, and AIC. Notably, all of the p-values are not significant – as we hypothesized due to our tree dataset being incomplete.

4.1.1 Baseline Assumptions

For the linear models, we assume independence, constant variance, linearity, and normality. We checked these assumptions for our baseline model predicting `AQI.Value`. A similar process could be repeated for the other response variables. Given that each observation is a different city, independence seems plausible, although there may be some dependencies in the air quality values if there are cities nearby each other. Looking at the residuals vs. fitted values plot in the appendix (Figure 11), the residuals are quite symmetrically distributed around 0, so linearity seems plausible. Constant variance seems to be violated as there is much less variance for higher fitted values than lower ones. This may be due to our small sample size and the slight right skew of the response variable. The QQplot (See Appendix Figure 12) shows that the residual distribution follows a normal distribution quite closely, with the exception of some outliers on the tail, so overall the normality assumption seems to be met.

4.1.2 Findings

Looking at Table 1, we can see that the model with the lowest p-value for `AQI.Value` is found by both forward and bidirectional stepwise regression. These models use tree density and the number of trees as predictors with coefficient estimates of 3.240e-04 and 3.531e-05, and p-values of 0.119 and 0.138 respectively. Intuitively, this means that an additional tree per square kilometer

[1] For readability purposes, we abbreviated the predictor variables as follows: NT = NumTrees, NUS = NumUniqueSpecies, NP = NaturalProportion, D = Density.

AQI.Value				
Model	Formula	p-value	RMSE	AIC
Baseline	AQI ~ NQ + NUS + NP + D	0.3504	22.5052	429.007
Interaction	AQI ~ (NT + NUS + NP + D)^2	0.5244	21.1069	435.106
Stepwise (Forward)	AQI ~ D + NT	0.1248	22.6047	425.413
Stepwise (Backward)	AQI ~ NUS + NP + D + NUS:D + NP:D	0.3047	22.0657	429.192
Stepwise (Both)	AQI ~ NT + D	0.1248	22.6047	425.413

Table 1: Results for linear models predicting AQI value

CO.AQI.Value				
Model	Formula	p-value	RMSE	AIC
Baseline	log(CO) ~ NT + NUS + NP + D	0.6318	2.04158	208.205
Interaction	log(CO) ~ (NT + NUS + NP + D)^2	0.9396	1.99527	218.094
Stepwise (Forward)	log(CO) ~ 1	—	2.10508	203.023
Stepwise (Backward)	log(CO) ~ 1	—	2.10508	203.023
Stepwise (Both)	log(CO) ~ 1	—	2.10508	203.023

Table 2: Results for linear models predicting carbon monoxide value

Ozone.AQI.Value				
Model	Formula	p-value	RMSE	AIC
Baseline	Ozone ~ NT + NUS + NP + D	0.5248	10.4057	358.039
Interaction	Ozone ~ (NT + NUS + NP + D)^2	0.7049	9.8511	364.999
Stepwise (Forward)	Ozone ~ NP	0.1658	10.5741	353.516
Stepwise (Backward)	Ozone ~ NP	0.1658	10.5741	353.516
Stepwise (Both)	Ozone ~ NP	0.1658	10.5741	353.516

Table 3: Results for linear models predicting ozone value

NO2.AQI.Value				
Model	Formula	p-value	RMSE	AIC
Baseline	log(NO2) ~ NT + NUS + NP + D	0.7154	2.3967	222.962
Interaction	log(NO2) ~ (NT + NUS + NP + D)^2	0.9549	2.3398	232.748
Stepwise (Forward)	log(NO2) ~ 1	—	2.4578	217.275
Stepwise (Backward)	log(NO2) ~ 1	—	2.4578	217.275
Stepwise (Both)	log(NO2) ~ 1	—	2.4578	217.275

Table 4: Results for linear models predicting nitrogen dioxide value

PM2.5.AQI.Value				
Model	Formula	p-value	RMSE	AIC
Baseline	PM2.5 ~ NT + NUS + NP + D	0.3315	23.5087	433.021
Interaction	PM2.5 ~ (NT + NUS + NP + D)^2	0.5025	22.0235	439.016
Stepwise (Forward)	PM2.5 ~ D + NT	0.1228	23.6489	429.568
Stepwise (Backward)	PM2.5 ~ NUS + NP + D + NUS:D + NP:D	0.2954	23.0664	433.273
Stepwise (Both)	PM2.5 ~ NT + D	0.1228	23.6489	429.568

Table 5: Results for linear models predicting PM2.5 value

will increase AQI by 3.240e-04 holding all else constant and one additional tree will increase AQI by 3.531e-05 holding all else constant.

For `CO.AQI.Value`, both our baseline and full in-

teraction models had very high p-values: 0.6318 and 0.9396 respectively. Interestingly, all three of the stepwise models chose the intercept model as the best model due to none of our predictors being very good at predicting carbon monoxide air pollution. As previously stated, this is likely due to our dataset being insufficient.

Looking at Table 3, we can see that the model with the lowest p-value for `Ozone.AQI.Value` is found by all three of the stepwise regression approaches. This model found that only using the proportion of naturally occurring trees to predict `Ozone.AQI.Value` leads to the best fitted model. The coefficient for proportion indicates that increasing the proportion of naturally occurring trees in a city from 0% to 100% will increase the ozone AQI by 7.883. This is counterintuitive to our original hypothesis as we believed that more naturally occurring trees would lead to less air pollution due to trees being better adapted to their environment. This result might be explained by cities with more introduced trees putting more effort into air pollution reduction programs.

Similar to `CO.AQI.Value`, both our baseline and full interaction models for `NO2.AQI.Value` had very high p-values – 0.7154 and 0.9549 respectively. Once again, all three of the stepwise regression approaches chose the intercept model as the best model, implying that none of our predictors are significant in predicting nitrogen dioxide pollution.

Lastly, for `PM2.5.AQI.Value`, we can see that the model with the lowest p-value is found by both forward and bi-directional stepwise regression. This model also found that density and the number of trees are the best predictors for `PM2.5.AQI.Value` with coefficients being 3.370e-04 with a p-value of 0.121 and 3.750e-05 with a p-value of 0.133 respectively. Intuitively, this suggests that an additional tree per square kilometer increases `PM2.5.AQI.Value` by 3.370e-04 holding all else constant and that an additional tree will increase `PM2.5.AQI.Value` by 3.750e-05 holding all else constant.

4.1.3 Best OLS Models

Across all five of the air pollution values, the full interaction model had the lowest RMSE. This can be explained by the full interaction model including all of our predictors as well as all of the interactions and potentially being overfit due to its complexity. Additionally, both the forward and bidirectional stepwise approaches always resulted in the same model which had the lowest AIC for each of the response variables. The consistency of model selection in both approaches strengthens the reliability of this model as it seems to have a good trade-off between model simplicity and the ability to explain the variation in the response variables.

4.2 Penalized Linear Models

4.2.1 Findings

A summary of the RMSEs and optimal λ values in LASSO and Ridge regression can be found in tables 6 and 7. Given the values, we note that the optimal lambda values for the LASSO model are generally lower than those for the Ridge model. This indicates that LASSO regularization (shrinking coefficients towards zero) is applied more lightly, which suggests that the tree predictors may be indeed relevant to predicting AQI values.

In the LASSO model, the RMSE for NO2 and CO on the test set is very close to the training set, suggesting that the model generalizes well for these pollutants. However, when choosing between the models, if the goal is to minimize test error, the LASSO model is more compelling for AQI, Ozone, and NO2 since it has a lower test RMSE in comparison to the Ridge model for these pollutants.

In contrast, for the Ridge model, the discrepancy between training and test RMSE is quite pronounced for AQI and PM2.5. This could be a sign of overfitting as the model may be capturing noise along with the actual signal in the training datas—$|||et, whichdoesnotgeneralizetothetestdataset$.

Appendix Figures 13 and 14 are examples of the coefficient values as we vary λ for PM2.5 as the response variable. The plots for other response variables look very similar as well. We can see that the natural proportion is the variable that is the last to be shrunk for both Ridge and LASSO, indicating that out of the predictors we have, natural proportion is the best for predicting air quality.

Overall, the results indicate that the LASSO model better balances model complexity and performance for several pollutants, as indicated by the optimal and RMSE values.

4.3 Tree-Based Modeling

A summary table for the RMSEs of the decision trees and random forests (as well as the tuned `mtry` parameter for each forest) can be found in Table 8. `mtry` refers to the number of variables the random forest considers at each split. Since there are only 4 total predictors, the models with log-transformed responses consider all predictors at every split while the models with original-scale responses only consider 2 variables at each split.

The RMSE values for the decision trees are all very similar to the linear models. However, the random forest RMSE values are much lower. This makes sense because random forests utilize bootstrapping, and given our small sample size, being able to artificially create more data to

Ridge Model			
Response	Optimal λ	Test RMSE	Train RMSE
AQI	0.498	19.996	23.873
log(CO)	0.0267	5.533	1.241
Ozone	0.225	10.810	10.795
log(NO2)	0.296	8.143	9.025
PM2.5	0.517	21.752	24.807

Table 6: Results for Ridge model

LASSO Model			
Response	Optimal λ	Test RMSE	Train RMSE
AQI	0.027	31.798	21.650
log(CO)	0.001	1.182	2.781
Ozone	0.010	5.908	11.581
log(NO2)	0.016	10.358	9.730
PM2.5	0.028	34.573	22.281

Table 7: Results for LASSO model

Decision Tree and Random Forest Models			
Response	Decision Tree RMSE	Random Forest RMSE	mtry
AQI	21.3659	11.8762	2
log(CO)	1.9622	0.7674	4
Ozone	9.2306	5.4519	2
log(NO2)	2.1913	0.9565	4
PM2.5	22.0012	12.6266	2

Table 8: Results for decision trees and random forest models

fit models on is very beneficial. Fitting many models and averaging across them also gives the random forest much more flexibility to capture higher complexity. However, the random forest model is also very likely overfitting, which we have no way of verifying without an out-of-sample dataset.

4.3.1 Decision Tree Structures

For each of the five air quality response variables, a decision tree was fit with the same transformations that were applied for the linear models, if applicable.

The decision tree for predicting AQI can be seen in Figure 4. The top split was on the proportion of naturally occurring trees. Cities with less than 0.075 proportion of naturally occurring trees were predicted to have an AQI of 78, which is higher than the predictions for cities with natural proportion above the 0.075 threshold. If a city has natural proportion above 0.075 and more than 60,000 trees, they are predicted to have the lowest AQI of 52. These results intuitively make sense in that having more trees, particularly naturally occurring trees would be beneficial to better air quality.

Decision trees for the remaining four variables can be seen in the appendix (Figure 15). The first split for PM2.5 matches the first split for AQI. Given that AQI is determined by the maximum value of the other four air quality indicators and PM2.5 is almost always the largest value, it makes sense that the PM2.5 tree would look very similar to AQI.

The first split for the tree predicting Ozone is natural proportion with cities having proportion below 0.11 predicted to have lower Ozone values. This result is counter-

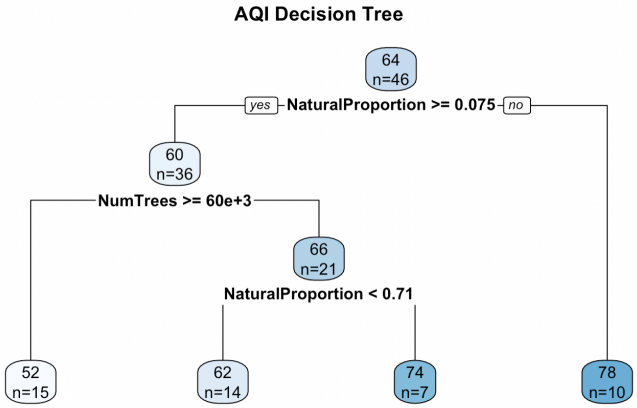

Figure 4: Decision tree predicting AQI

intuitive as we would expect places with more naturally occurring to have better air quality, but this matches the result from the OLS models where ozone pollution and natural proportion had a positive coefficient. One possible explanation for this is that there may be some confounding variables. For example, maybe cities with ozone pollution problems are introducing non-natural trees in an effort to reduce air quality.

The first split for both log-transformed CO and log-transformed NO2 were the number of trees being greater than 16,000, and being above this threshold was associated with higher predicted CO and NO2 values. This result is also counterintuitive as we would expect more trees to result in less air pollution. It's possible that trees do not have a significant impact on the altering CO or NO2 levels or, similarly to the ozone pollution, cities are

with high CO and NO2 levels are actively planting more trees in an effort to reduce pollution.

It is worth noting that when using cross validation to tune the decision tree complexity, the resulting tree for almost all response variables had no splits, with the exception of PM2.5 which had only the first split on the natural proportion. This indicates that in a predictive setting, these trees would be overfitting, and the associations from the splits in these trees are likely not very strong.

4.3.2 Variable Importance

The variable importance plots for both the decision trees (Figure 16) and the random forests (Figure 17) can be found in the appendix. Although we found in the linear models that none of the predictors had a significant association with the response, these variable importance plots can give us a sense of how these predictors would rank for predicting a given air quality value regardless of whether it's significant or not.

For AQI, Ozone, and PM2.5, natural proportion was a clear winner for having the most importance among the four predictor variables in the decision trees. In OLS section, we also found that the best model through stepwise selection for predicting ozone was a model with natural proportion as the sole predictor, so this aligns with our previous result. However, for AQI and PM2.5, the stepwise models previously found density and number of trees to be the more significant predictors. The discrepancy between the linear model and the decision tree might be because the natural proportion is not linearly related to AQI and PM2.5, so the linear-based models did not capture this relationship. Looking at the EDA scatterplots (Figure 3, Appendix Figures 7-10), there does seem to be a weak pattern in the naturally occurring proportion plots that is non-linear.

For the CO and NO2 decision trees, the number of trees was the most important variable, which aligns with the first splits of the decision trees for these variables.

For the random forest models, the top variable was natural proportion for all response variables. This further suggests that air quality may be strongly associated with the proportion of naturally occurring trees non-linearly. For CO and NO2, the importance of natural proportion is much more obvious than for the other air quality response variables. This is likely due to the fact that `mtry` for CO and NO2 was 4, so natural proportion was always a variable that could be chosen if it was the most predictive. On the other hand, the other models had `mtry` = 2, so there were certain splits where natural proportion was no considered, leading it to have deflated variable importance.

5 Discussion and Conclusions

5.1 Overall Findings

Across all our approaches, we generally found that our predictors were not very powerful for predicting air quality. The p-values of our linear regression models were very high (none were significant), and the decision trees, when using cross-validated tuning parameters, resulted in models with no splits.

However, there were some predictors that arise as potentially being associated with air quality. In particular, proportion of naturally occurring trees was shown to be a good predictor of ozone pollution in the stepwise models, the decision tree, and the random forest. In the Ridge and LASSO models, natural proportion was the last predictor to be regularized for all of the response variables, which also shows that it may be a good predictor. Additionally, density, number of trees, and natural proportion may be potential predictors of AQI and PM2.5 based on the top decision tree splits and the stepwise models.

5.2 Challenges Faced

In assessing the relationship between several tree characteristics in urban environments and air quality measured by AQI values, the goal is to explore the role of trees as natural mitigators and indicators of urban air pollution. However, several challenges and considerations arise.

In particular, our small dataset may not represent the wide variety of tree species, urban environments, and pollution levels. This limits the ability to generalize findings to broader contexts. This was a concern when conducting cross validation for fitting the Ridge and LASSO models. Given that the response variable was measured using AQI, the small dataset made it more probable to have uniform values in a single fold, which was also a challenge when constructing the decision tree and random forests. There was greater risk of overfitting, especially as the number of predictors increased when introducing interaction terms. The small sample size of the dataset also contributed to the large p-values, as it reduced the statistical power of the analysis and increased the risk of Type II errors.

5.3 Future Considerations

To solve this data issue, conducting nationwide tree surveys can provide more comprehensive data. Having more coverage across the US, as well as filling in the gaps for specific species, ages, health status, and locations of trees would increase the statistical power of this analysis. However, there are many obstacles to overcome to conduct this data collection. Comprehensive surveys require significant time, labor, and financial resources.

Moreover, gaining access to all urban areas, especially private properties, can be difficult, as well as ensuring consistent data collection methods across different regions for data quality.

Having access to a more comprehensive and robust dataset opens avenues to new paths of exploration. Specifically, some species might be more efficient due to their physical characteristics, like leaf structure or transpiration rates. Given that we found from this study that the proportion of native species plays a role in predicting air quality, investigating individual tree species' effectiveness in combating air pollution may yield interesting and significant results.

The potential of trees as natural mitigators of urban air pollution is a promising area of study. However, realizing this potential requires overcoming the challenges of small datasets through comprehensive data collection, species-specific analysis, and a multidisciplinary approach that includes public engagement and policy implications, and has significant implications for urban planning and public health.

5.4 Concluding Remarks

This study sought to explore the relationship between urban trees and air quality, as indicated by AQI values, and while the results were not as robust as we wanted, they provide a crucial starting point for further research. The findings tentatively suggest that the proportion of natural trees may have a mitigating effect on certain types of air pollution, such as ozone. However, the limitations of our dataset and the challenges faced underscore the need for more extensive data to draw firmer conclusions. Despite the encountered limitations — especially the small sample size and its impact on the robustness of statistical inferences — our research serves as a foundational step towards a more nuanced understanding of how urban greenery influences air quality. As cities continue to grow and environmental concerns become more pressing, understanding and leveraging the role of urban trees could contribute to healthier, more sustainable urban environments.

6 Appendix

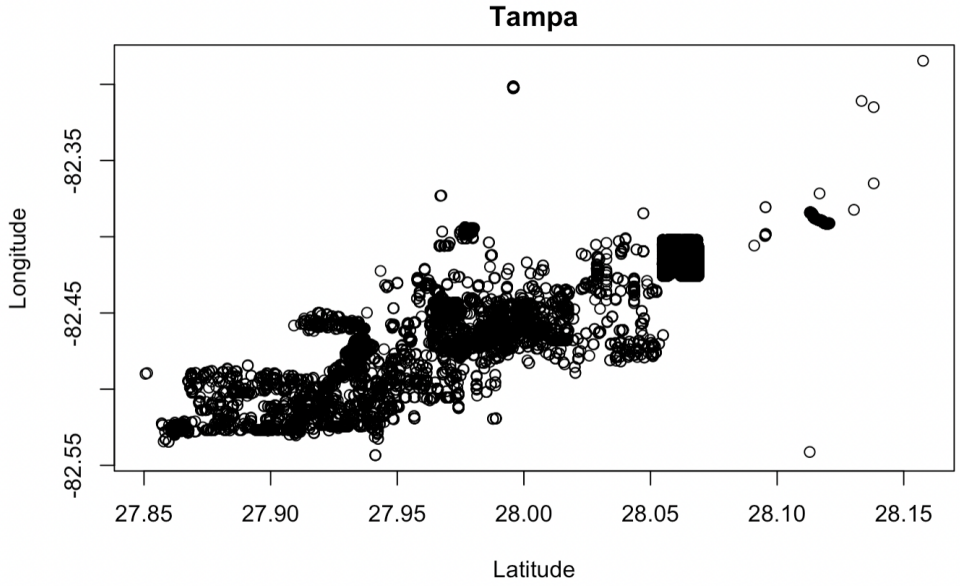

Figure 5: Distribution of trees across Tampa

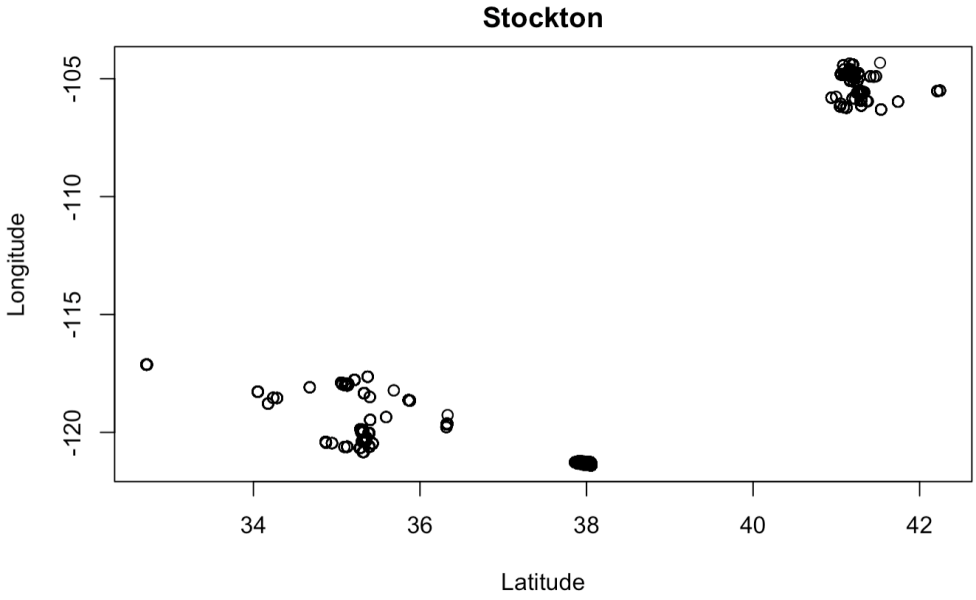

Figure 6: Distribution of trees across Stockton

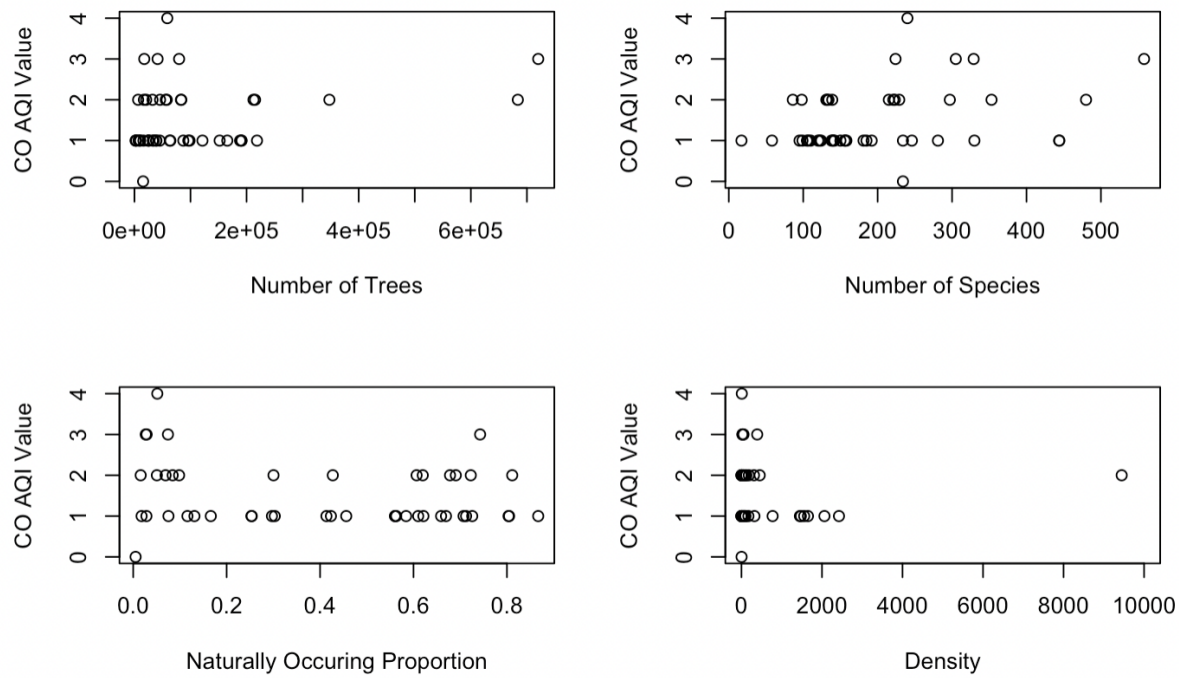

Figure 7: Scatter plot of tree variables against carbon monoxide value

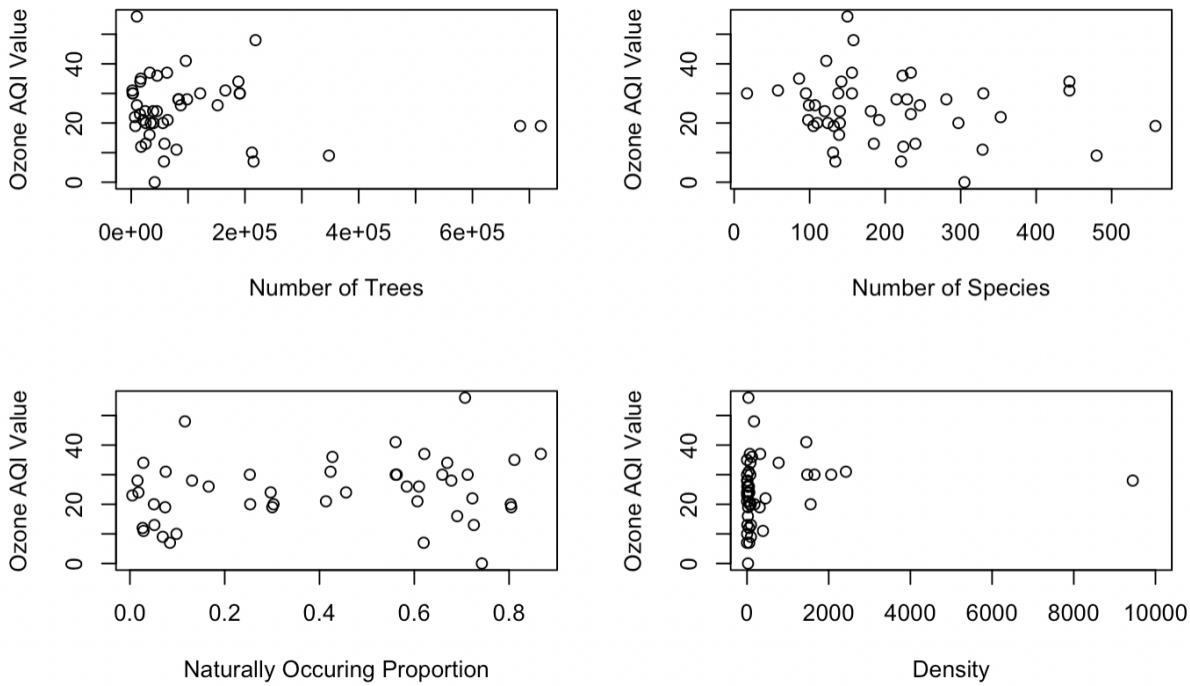

Figure 8: Scatter plot of tree variables against ozone value

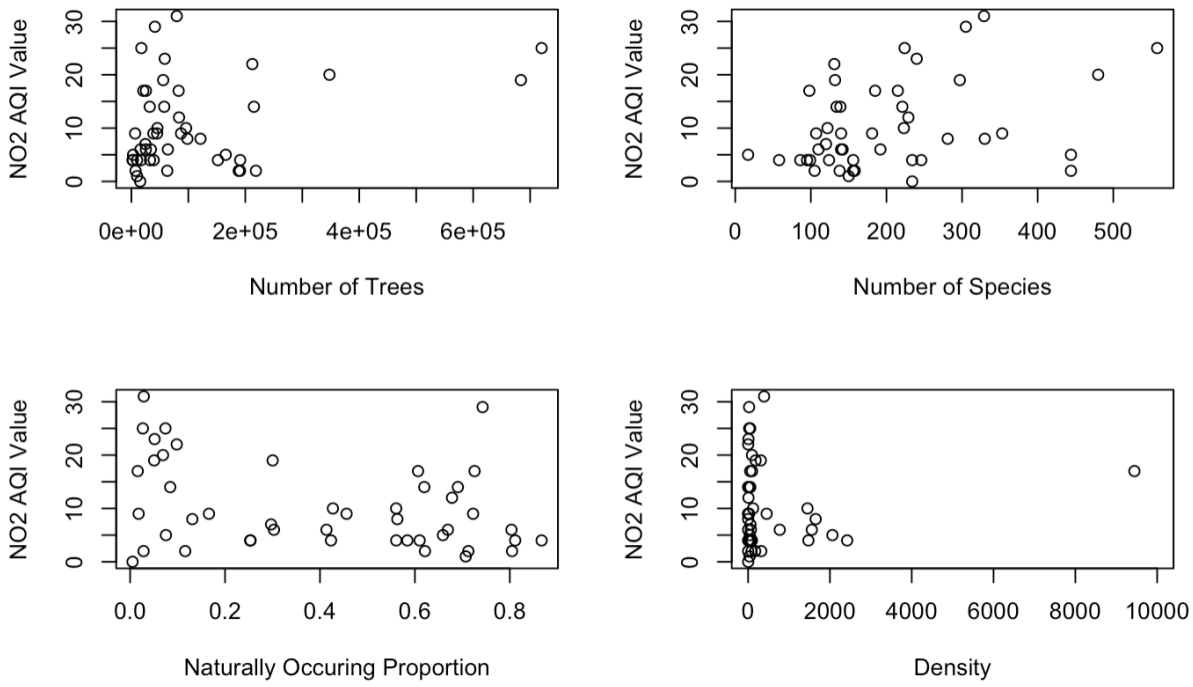

Figure 9: Scatter plot of tree variables against nitrogen dioxide value

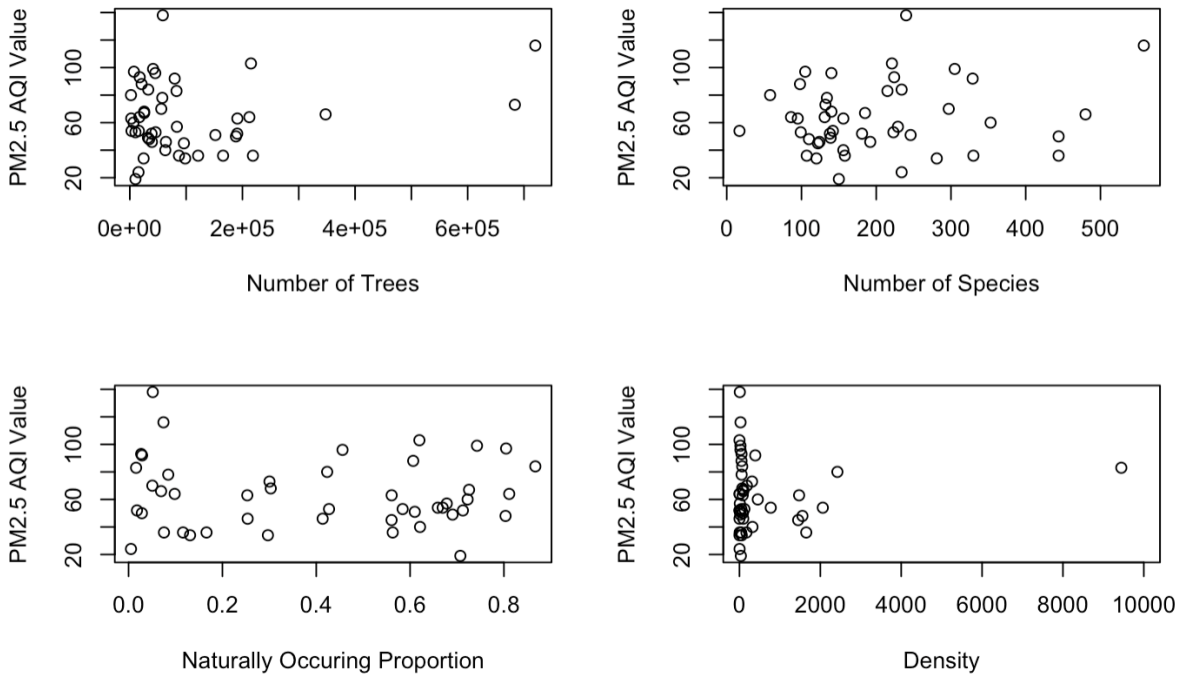

Figure 10: Scatter plot of tree variables against PM2.5 value

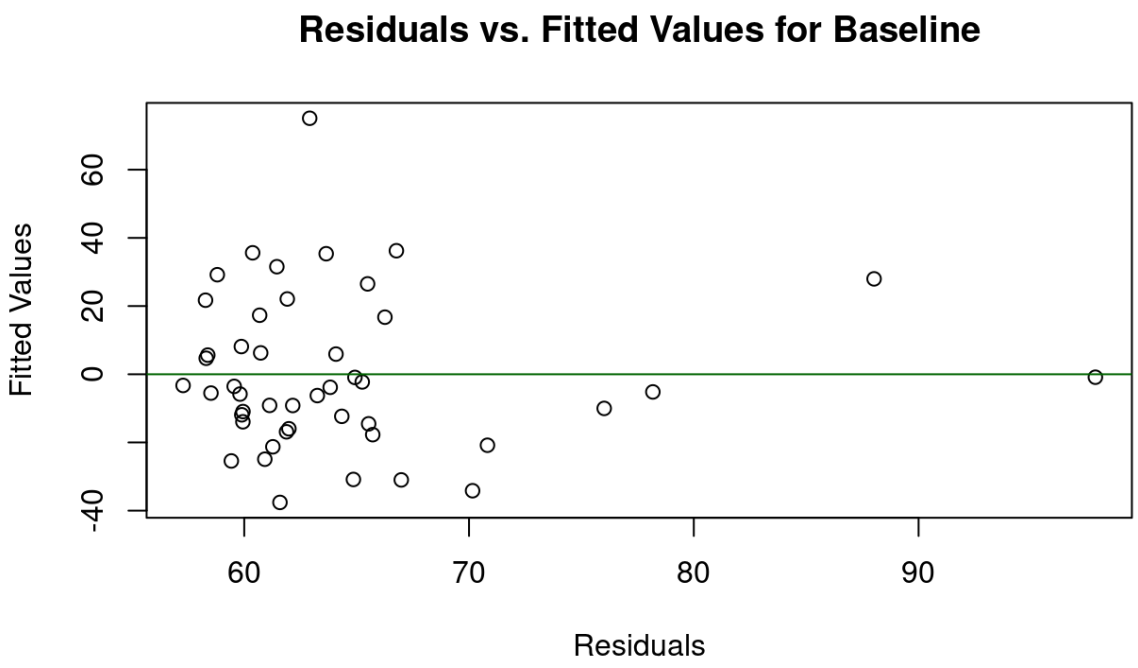

Figure 11: Fitted values vs. residual plot for baseline model predicting AQI to check constant variance and linearity assumptions

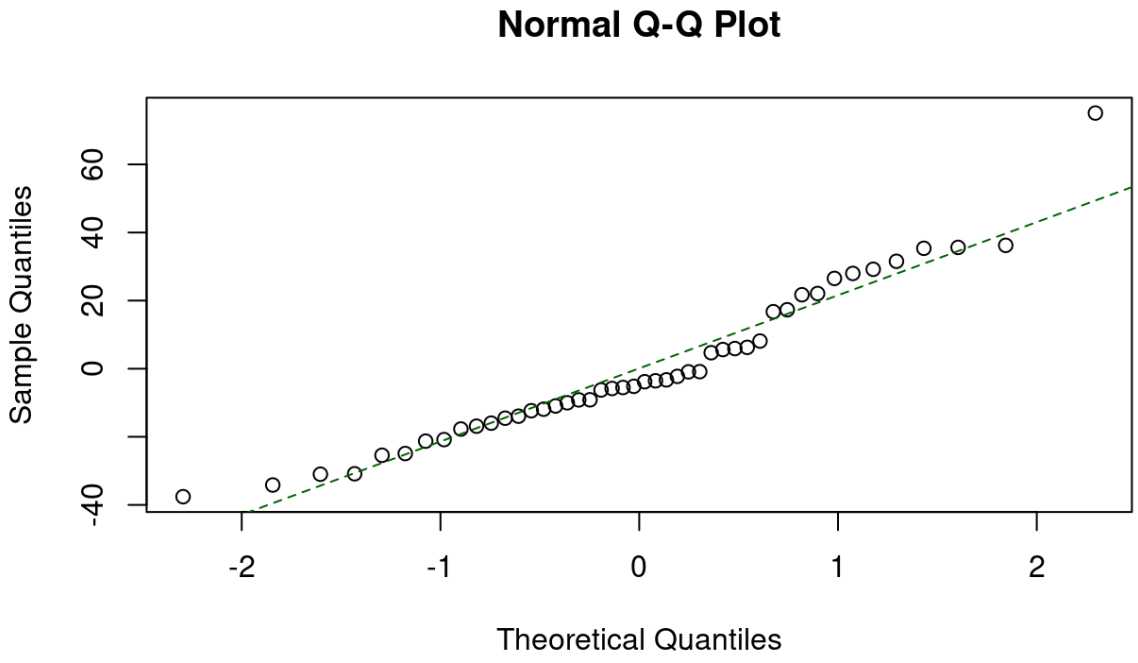

Figure 12: Normal QQ plot for residuals of baseline model predicting AQI

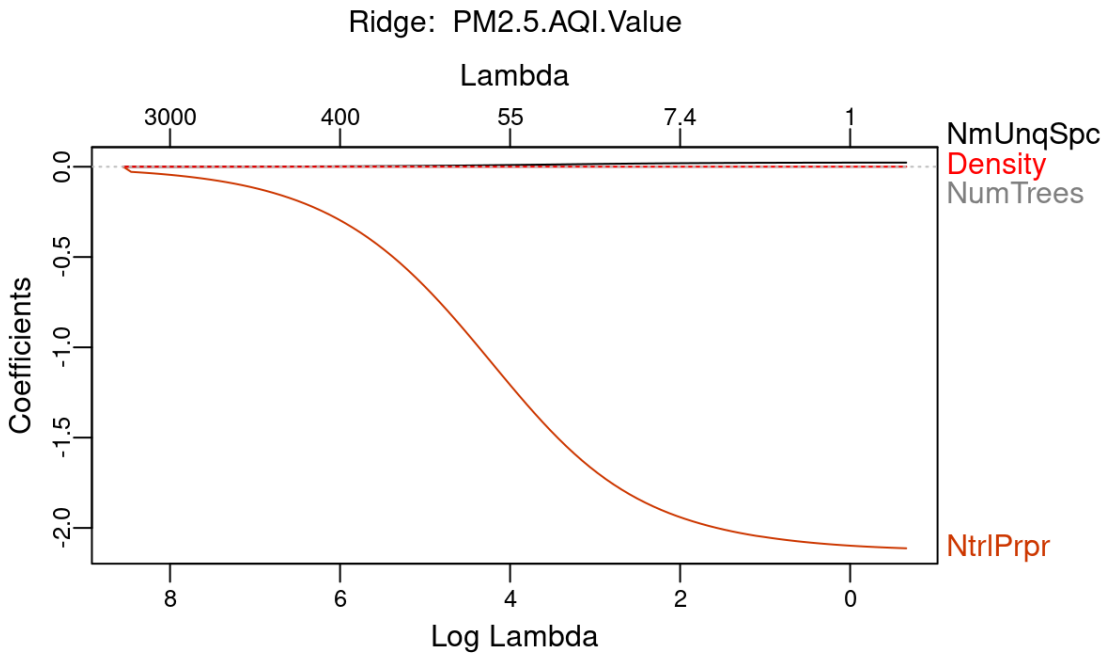

Figure 13: Coefficient values for various λ values in Ridge regression

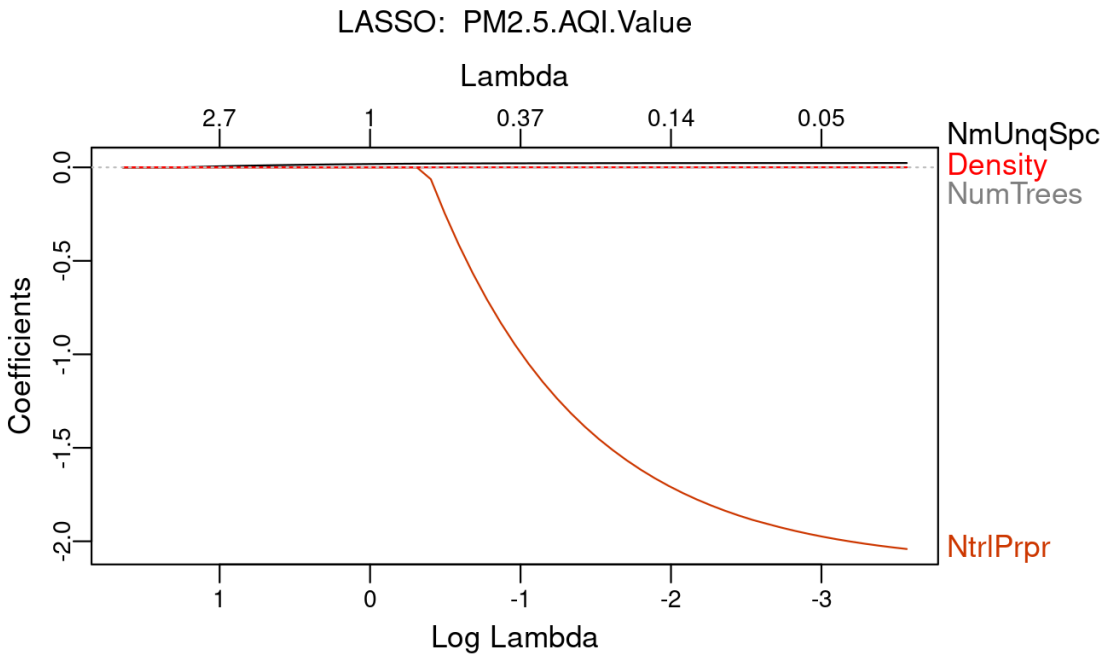

Figure 14: Coefficient values for various λ values in LASSO regression

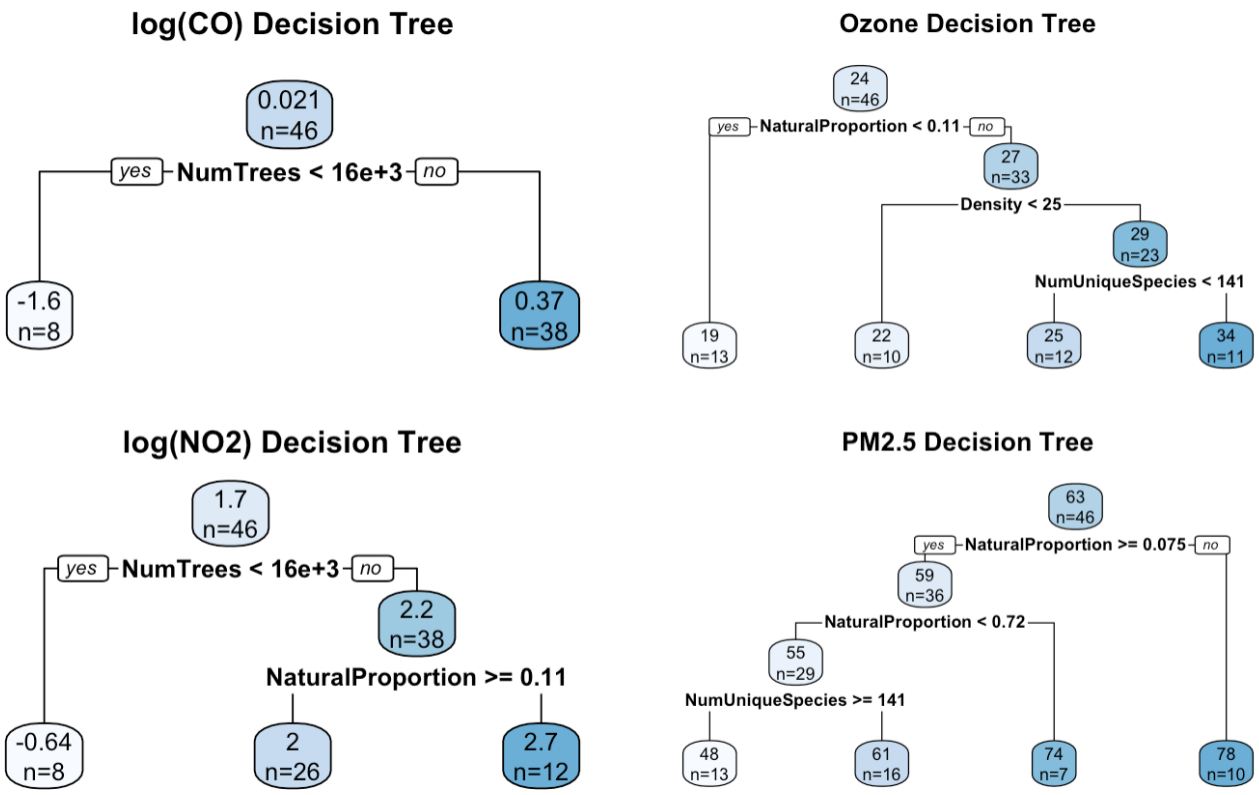

Figure 15: Decision trees for CO, Ozone, NO2, and PM2.5

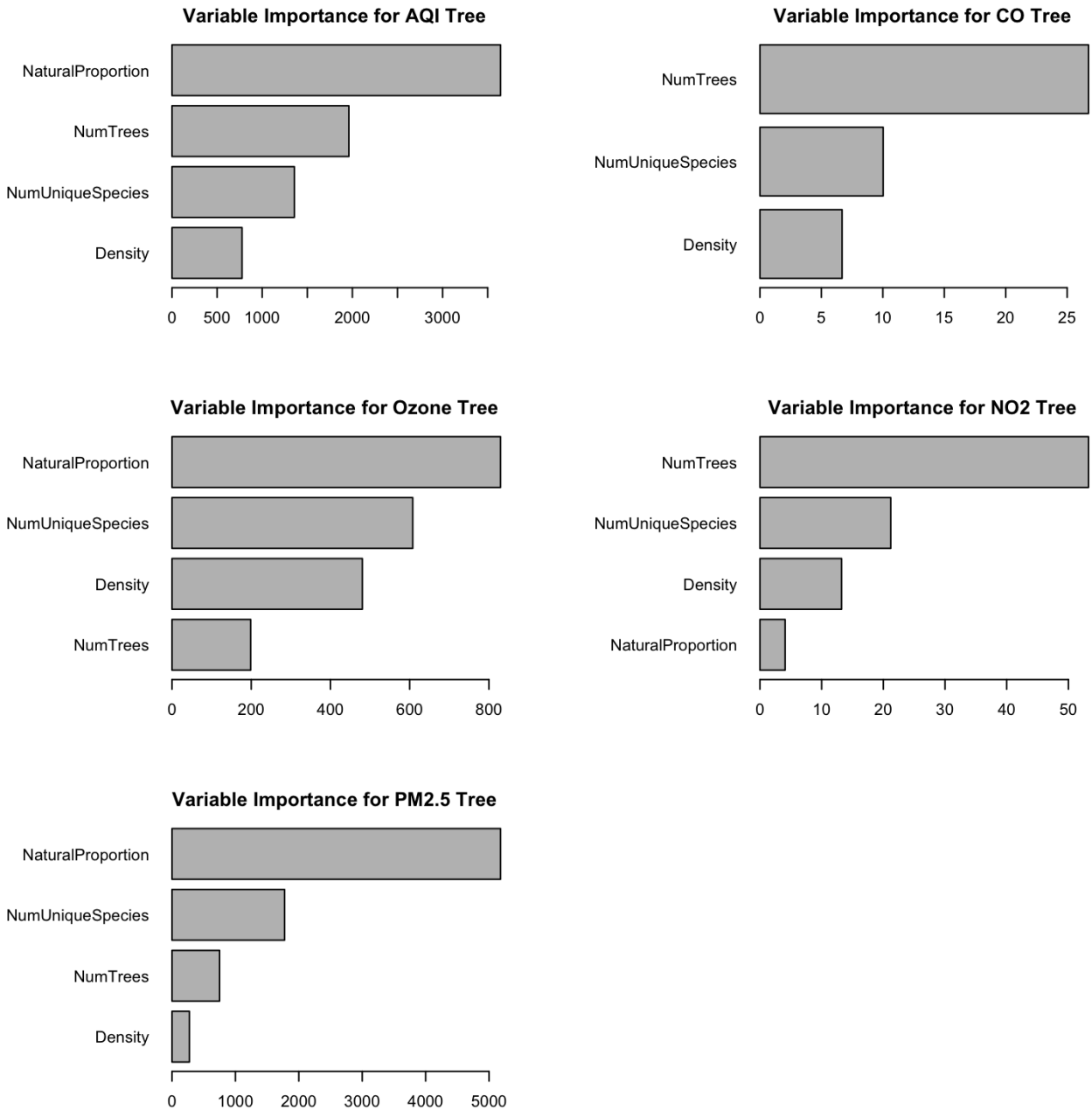

Figure 16: Variable importance for decision trees

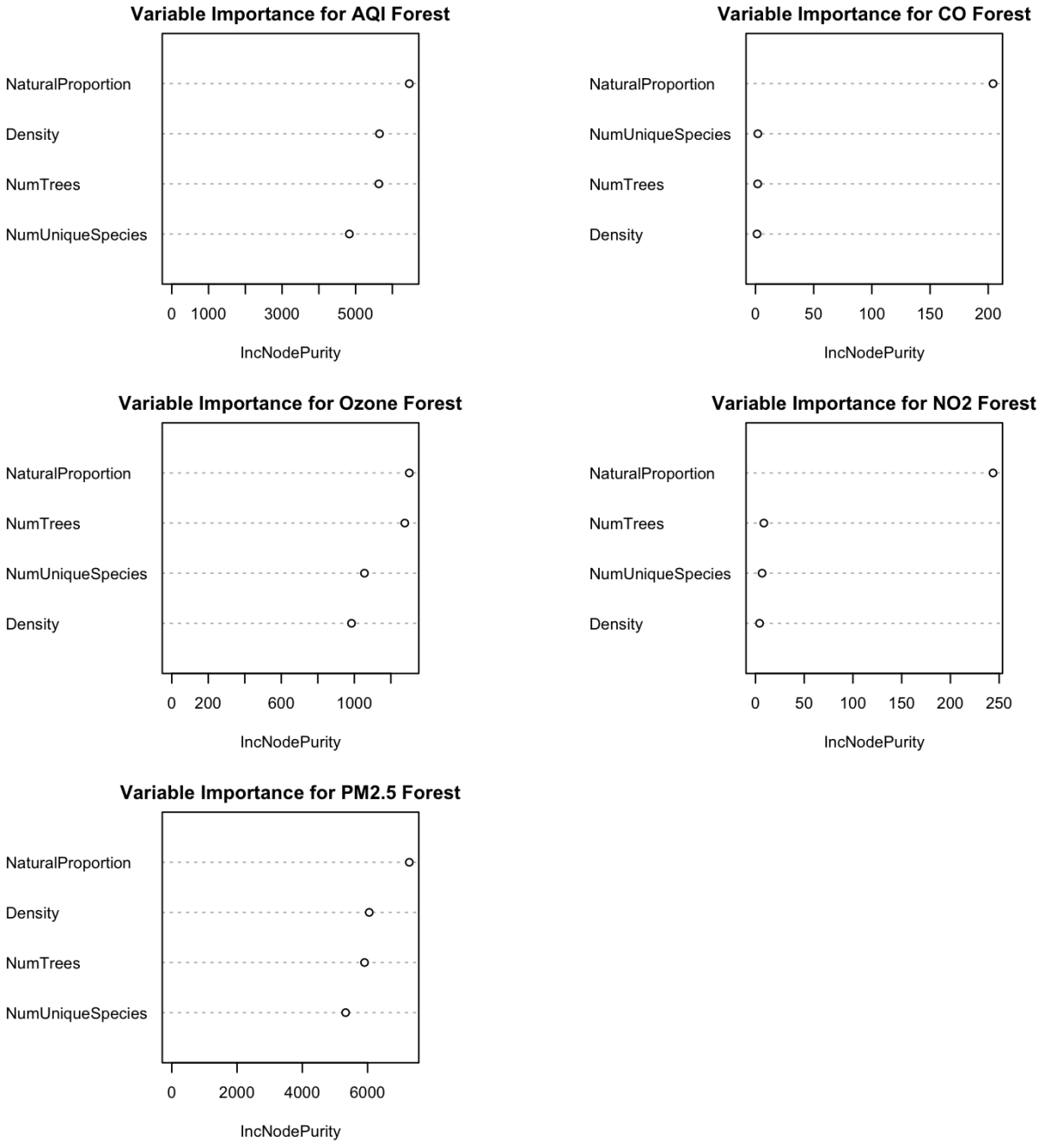

Figure 17: Variable importance for random forest