Analyzing the Correlations among Tree Characteristics and their Surroundings

Authors

- **Hadil Helaly**
  · ⬡ [hadilhelaly](#)
  Department of Civil and Environmental Engineering

- **Emma Golub**
  · ⬡ [emmaagolub](#)
  Department of Civil and Environmental Engineering

- **Riley Blasiak**
  · ⬡ [blasiak2](#)
  Department of Civil and Environmental Engineering

- **Rupesh Rokade**
  · ⬡ [RupeshRokade16](#)
  Department of Civil and Environmental Engineering

Introduction

The Urban tree database, which was collected by the US Forest Service Research Archive of the US Department of Agriculture, includes data about tree growth in urban areas across 17 cities and 13 states over the span of 14-years (from 1998-2012). The states included in the study are: Arizona, California, Colorado, Florida, Hawaii, Idaho, Indiana, Minnesota, New Mexico, New York, North Carolina, Oregon, and South Carolina. The data come from measurements taken to over 14,000 street and urban park trees, and the data can be obtained by downloading the 1.08 MB compressed "data publication" file from [Link](#). Some measurements of interest include tree age, location, height, crown diameter, leaf area, foliar biomass, and utility line interference. Tree age, for example, was determined from interviews with residents, street construction dates, aerial and historical photos, the city's urban forester, and laboratory cores developed by the Lamont-Doherty Earth Observatory's Tree Ring Laboratory.

The downloaded folder includes 9 data sheets in CSV format. The most interesting data files are 1. TS1_Regional_information.csv, 2. TS2_Regional_species_and_counts.csv, and 3. TS3_Raw_tree_data.csv. First, the "TS1_Regional_information.csv" file contains information about region code, city, state, airport codes, and collection year. Second, the "TS2_Regional_species_and_counts.csv" file contains information (columns) regarding region, scientific and common names of trees, tree type, and 9 columns of dbh_class, which represent a species diameter at breast height and are used to predict tree height, crown diameter, crown height, and leaf area. The file contains a total of 347 rows. Finally, the "TS3_Raw_tree_data.csv" file includes 14487 observations (rows) of raw tree data. For each observation, 41 different variables were collected (columns). A detailed description of each of these 41 variables is as followed:

- DbaseID = Unique id number for each tree.
- Region = 16 U.S. climate regions, abbreviations are used.
- City = City/state names where data collected.
- Source = Original *.xls filename (not available in this data publication).
- TreeID = Number assigned to each tree in inventory by city.
- Zone = Number/ID/name of the management area or zone that the tree is located in within a city; or nursery if young tree data collected there.
- Park/Street = Data listed as Park, Street, Regional Big Tree, or Nursery (for young tree measurements).
- SpCode = 4 to 6 letter code consisting of the first two letters of the genus name and the first two letters of the species name followed by two optional letters to distinguish two species with the same four-letter code (See _Regional_species_and_counts.csv for a list of the SpCodes and corresponding scientific names.)
- ScientificName = Botanical name of species.
- CommonName = Common name of species.
- Tree Type = 3 letter code where first two letters refer to life form (BD=broadleaf deciduous, BE=broadleaf evergreen, CE=coniferous evergreen, PE=palm evergreen) and the third letter is mature height (S=small which is < 8 meters, M=medium which is 8-15 meters, and L=large which is > 15 meters).

- Address = From inventory, street number of building where tree is located.
- Street = From inventory, the name of the street the tree is located on. (NOTE: zero values denote data were not recorded in that city. These values were left unchanged because they originated from city inventories.)
- Side = From inventory, side of building or lot tree is located on (F=front, M=median, S=side, P=park). (NOTE: zero values denote data were not recorded in that city. These values were left unchanged because they originated from city inventories.)
- Cell = From inventory, the cell number (i.e., 1, 2, 3, …), where protocol determines the order trees at same address are numbered (e.g., driving direction or as street number increases).
- OnStreet = From inventory (omitted if not a field in city's inventory), for trees at corner addresses when tree is on cross street rather than addressed street. FromStreet = From inventory, the name of the first cross street that forms a boundary for trees lining un-addressed boulevards. Trees are typically numbered in order (1, 2, 3 …) on boulevards that have no development adjacent to them, no obvious parcel addresses.
- ToStreet = From inventory, the name of the last cross street that forms a boundary for trees lining un-addressed boulevards.
- Age = Number of years since planted. (NOTE: zero values represent newly planted trees, < 1 year old.)
- DBH (cm) = Diameter at breast height (1.37 meters [m]) measured to nearest 0.1 centimeters (tape). For multi-stemmed trees forking below 1.37 m measured above the butt flare and below the point where the stem begins forking, as per protocol.
- TreeHt (m) = From ground level to tree top to nearest 0.5 m (omitting erratic leader).
- CrnBase (m) = Average distance between ground and lowest foliage layer to nearest 0.5 m (omitting erratic branch).
- CrnHt (m) = Calculated as TreeHT minus Crnbase to nearest 0.5 m. (NOTE: zero values indicate no live crown was present, hence no other tree dimension data were available.)
- CdiaPar (m) = Crown diameter measurement taken to the nearest 0.5 m parallel to the street (omitting erratic branch).
- CDiaPerp (m) = Crown diameter measurement taken to the nearest 0.5 m perpendicular to the street (omitting erratic branch).
- AvgCdia (m) = The average of crown diameter measured parallel and perpendicular to the street.
- Leaf (m2) = Estimated using digital imaging method to nearest 0.1 squared meter (m2).
- Setback = Distance from tree to nearest air-conditioned/heated space (may not be same address as tree location): 1=0-8 m, 2=8.1-12 m, 3=12.1-18 m, 4=> 18 m.
- TreeOr = Taken with compass, the coordinate of tree taken from imaginary lines extending from walls of the nearest conditioned space (may not be same address as tree location).
- CarShade = Number of parked automotive vehicles with some part under the tree's drip line. Car must be present (0=no autos, 1=1 auto, etc.).
- LandUse = Predominant land use type where tree is growing (1=single family residential, 2=multi-family residential [duplex, apartments, condos], 3=industrial/institutional/large commercial [schools, gov't, hospitals], 4=park/vacant/other [agric., unmanaged riparian areas of greenbelts], 5=small commercial [minimart, retail boutiques, etc.], 6=transportation corridor).
- Shape = Visual estimate of crown shape verified from each side with actual measured dimensions of crown height and average crown diameter (1=cylinder [maintains same crown diameter in top and bottom thirds of tree], 2=ellipsoid, the tree's center [whether vertical or horizontal is the widest, includes spherical], 3=paraboloid [widest in bottom third of crown], 4=upside down paraboloid [widest in top third of crown]).
- WireConf = Utility lines that interfere with or appear above tree (0=no lines, 1=present and no potential conflict, 2=present and conflicting, 3=present and potential for conflicting). (NOTE: -1 denotes data were not collected.)
- dbh1 = Dbh (centimeters [cm]) for multi-stemmed trees; for non-multi-stemmed trees, dbh1 is same as Dbh (cm).
- dbh2 = Dbh (cm) for second stem of multi-stemmed trees.
- dbh3 = Dbh (cm) for third stem of multi-stemmed trees.
- dbh4 = Dbh (cm) for fourth stem of multi-stemmed trees.
- dbh5 = Dbh (cm) for fifth stem of multi-stemmed trees.
- dbh6 = Dbh (cm) for sixth stem of multi-stemmed trees.
- dbh7 = Dbh (cm) for seventh stem of multi-stemmed trees.
- dbh8 = Dbh (cm) for eight stem of multi-stemmed trees.

Additionally, a fourth data set may be of later interest for estimating leaf area, species dominance at a spatial scale, and carbon storage estimates. The TS5_Foliar_biomass_leaf_samples.csv contains urban foliar samples data by species for 17 U.S. cities. A total of 261 rows are provided.

The breadth of this dataset allows for a myriad of problems to be explored. The primary data that will be utilized for this project is the "TS3_Raw_tree_data.csv" file, as this contains the most columns which will result in more feasible predictions during the machine learning portion of the project. This data can be used to analyze correlations between tree characteristics and their surroundings. One potential research question using the "TS3_Raw_tree_data.csv" file is: how does utility line interference affect the growth of a certain type of tree in one state versus a different state. the preliminery 14 variables that can be used in the proposed analysis include "Address", "Age", "Shape", "WireConf", "Setback","CarShade", "DBH", "TreeHt", "CrnBas" "CrnHt", "CdiaPar", "CDiaPerp", "AvgCdia", "Leaf".

After tidying the dataset, we can compare the effect of the WireConf, Setback, CarShade on the remaining variables of similar trees. Since we also contain the addresses of the trees, along with visualizing graphs from results of the comparisons, we can create maps to understand the variance of these effects across different cities. Further, a machine learning model can be created to possibly target and predict the above results for a city that is not mentioned in the dataset and predict the missing values in the dataset.

Exploratory Data Analysis

Tree growth depends on many factors, some of which are included in this dataset, while others are outside the scope of this work. Resulting from the following data analysis, different regions were found to have trees of the same species and ages, but variable heights. This can be attributed to climate specific variables such as yearly temperature, precipitation, flooding, and even wind speeds [1]. Variables within the scope of this project that could have an impact on tree growth include: power line interference, setback from conditioned spaces, and land use. Additionally, variables describing tree growth such as diameter at breast height, leaf volume, tree height, and tree age also depend on one another [2,3].

Throughout this exploratory analysis, four main questions were developed to guide data exploration, which involved data wrangling to produce visualizations of potential correlations among selected variables of interest. These questions are as follows:

1. How does setback (tree distance from heated/air-conditioned spaces) show in different cities and/or regions? (i.e., correlation with tree height and location)

2. How does growth rate (i.e., height per age of tree) differ for each region, land use, city, etc.?

3. How do power lines impact the growth of trees? (i.e., number of trees, leaf area, tree height, power lines)

4. What are the correlations between tree type, land use, height, leaf area, car shade, DBH, CdiaPar, and CDiaPerp for urban tree planning by region and/or city?

Initially, setback was investigated to understand the effects it has on the height of trees. Setback is defined as the distance from the tree to the nearest air-conditioned or heated space (which may not be the same address as the tree location), with values of 1,2,3,4, which are defined as 0m to 8m, 8.1m to 12m, 12.1m to 18m, and > 18m, respectively. . After filtering out all the missing values from the dataset, a bar graph was plotted for the mean setback across various locations. (Figure 1)
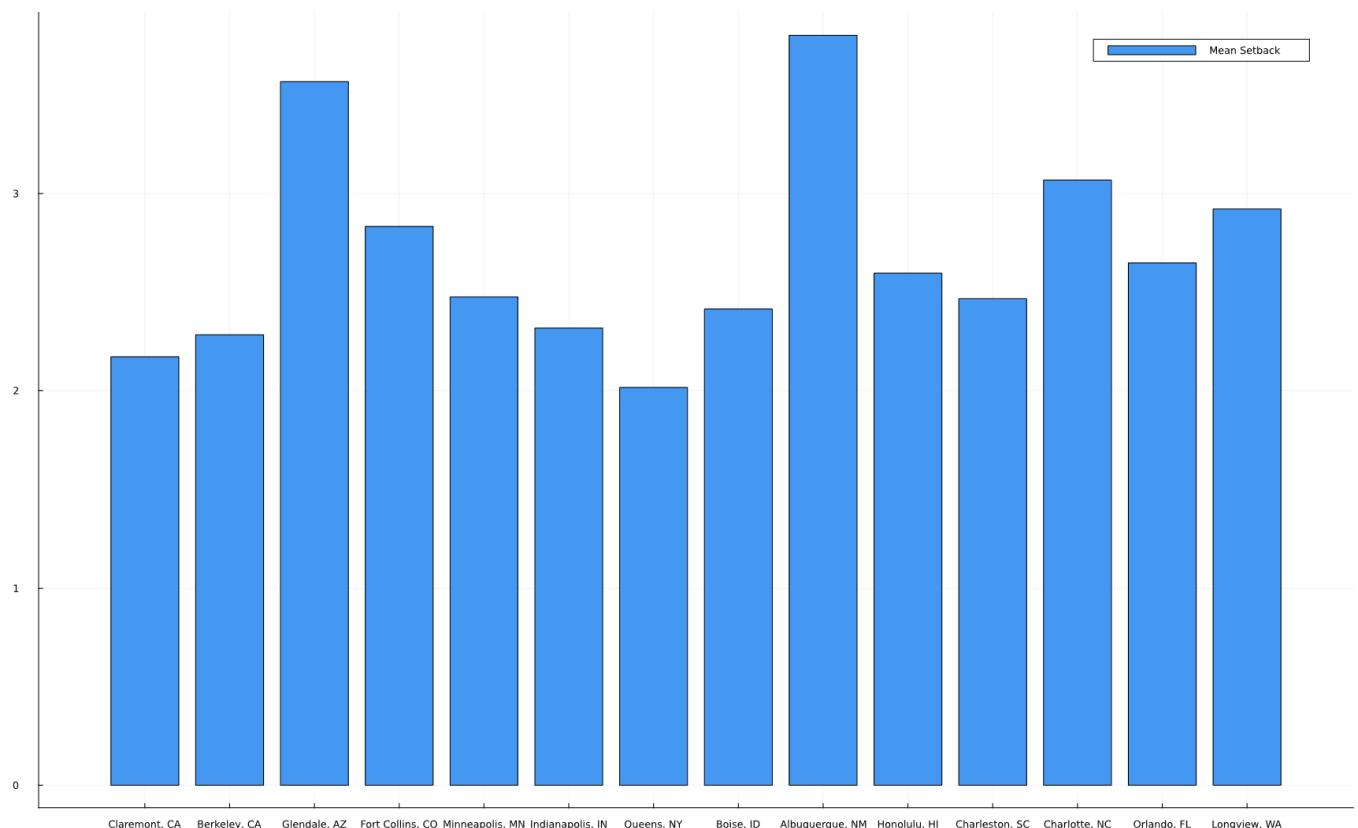

Figure 1: The Mean Setback across different Cities

It was identified that the cities with the highest mean setback in descending order are 1) Albuquerque (3.80), 2) Glendale (3.56), 3) Charlotte (3.06), and 4) Longview (2.92). Similarly, the cities with least mean setback in ascending order are 1) Queens, (2.01), 2) Claremont (2.17), 3) Berkeley (2.28), and 4) Indianapolis (2.31). Next, similar tree species from the top four mean setback values and bottom four mean setback values were identified. This helped to establish a similar medium for tree height comparison. However, it was found that no common species were present between the two groups. Therefore, a random city (Charlotte) was analyzed, where similar species having the same age were grouped together.
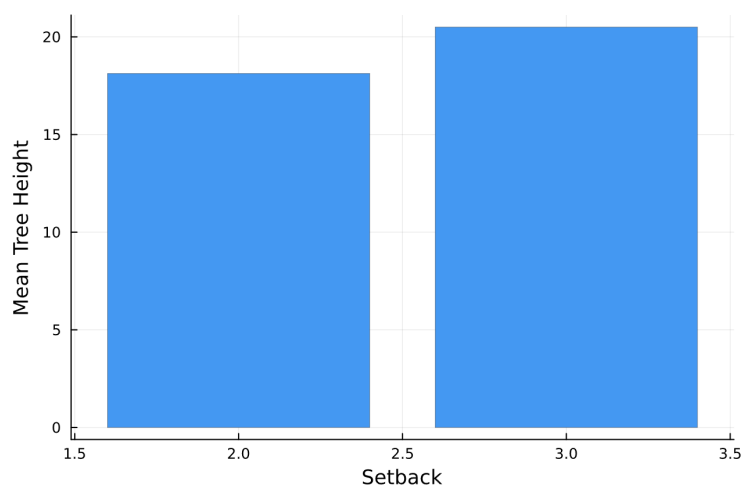

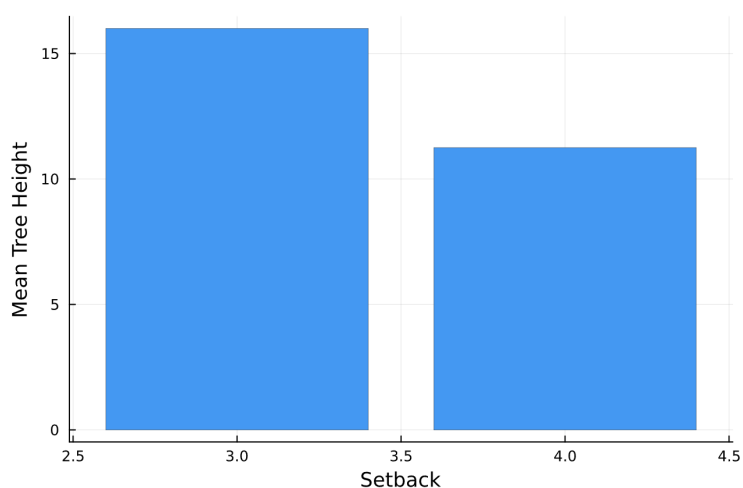**Figure 2:** Mean Tree Height vs Setback for Silver Maple trees

Figure 3: Mean Tree Height vs Setback for River Birch trees

Using their mean heights, it was observed that setback and tree height presented no correlation as seen in the following 2 cases:

1. Silver Maple trees of age 35 with Setback of 2 and 3. (Figure 2)

2. River Birch trees of age 15 with Setback of 3 and 4. (Figure 3)

Next, the tree height parameter was explored by selecting two random cities - Longview, WA and Modesto, CA, and the species that were selected due to their existence in both locales were - Sweetgum, Cherry Plum, and European White Birch.
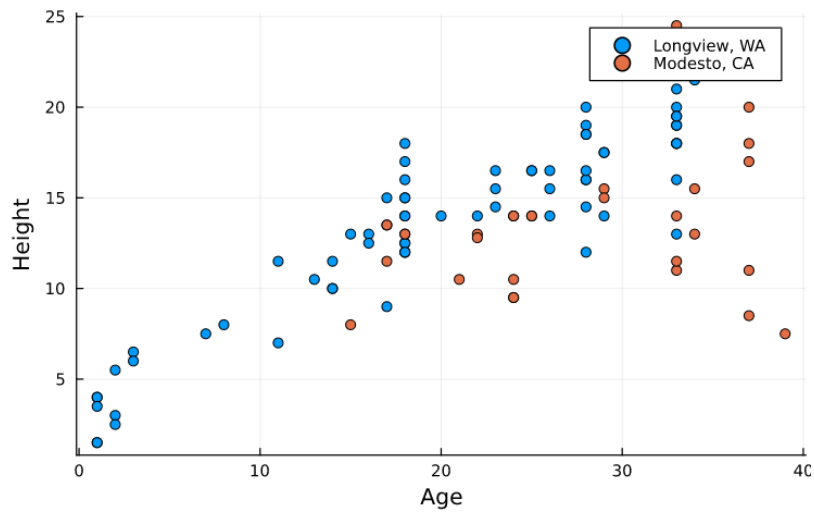
Figure 4: Age versus Height of Sweetgum trees in Longview, WA and Modesto, CA.

Figure 4 shows that in Longview, WA, Sweetgum trees are taller than Modesto, CA trees at any age.
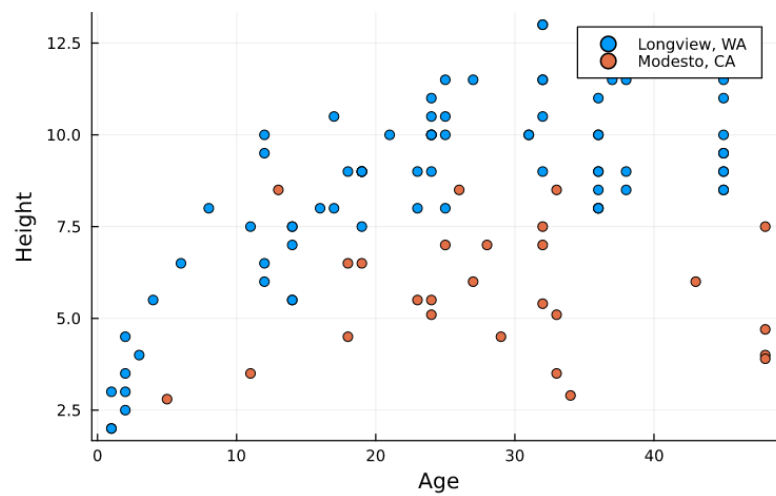

Figure 5: Age versus Height of Cherry Plum trees in Longview, WA and Modesto, CA.

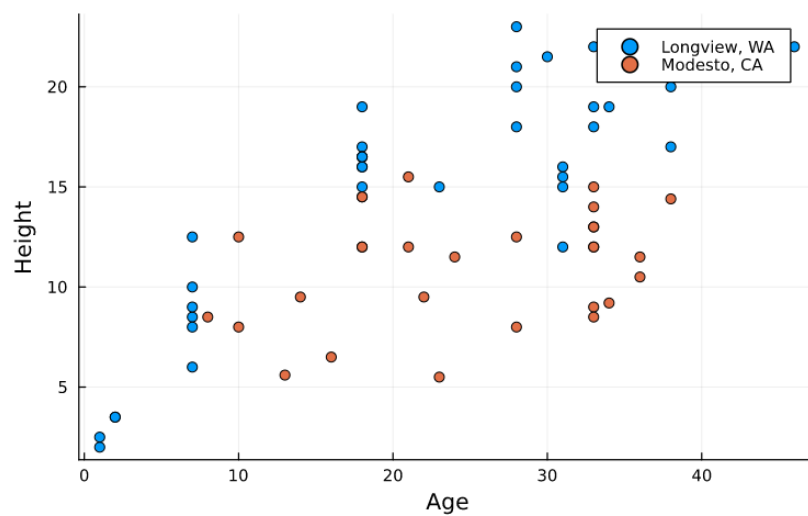Figure 5 shows that in Longview, WA, Cherry Plum trees are taller than Modesto, CA trees at any age.

Figure 6: Age versus Height of European White Birch trees in Longview, WA and Modesto, CA.

Figure 6 shows that in Longview, WA, European Birch trees are taller than Modesto, CA trees at any age.

Next, the growth rate (Tree Height / Age) vs Tree ID yielded the following graph:
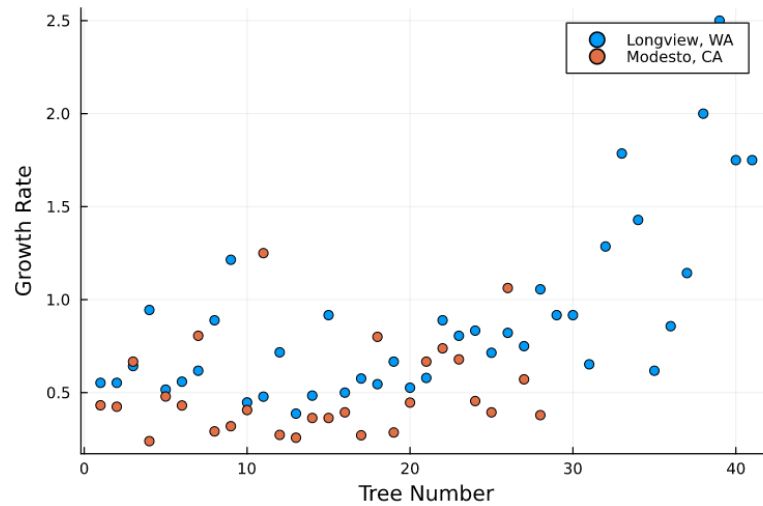
Figure 7: Growth rate of European White Birch in Longview, WA and Modesto, CA.

This figure shows that the growth rate is not constant for one tree type, and may vary throughout the tree's lifetime. Here it can be seen that generally, the growth rate in Longview, WA is greater than that of Modesto, CA.

These figures illustrate a relationship between location and height of trees. This relationship may be attributed to different temperature, precipitation, or other factors outside of the dataset.

To further understand the differences for tree height across all the locations in the dataset, a barplot of tree heights grouped by species was produced over all locations to study typical heights associated with each tree type. From Fig 8, one can observe how some trees (i.e., blue gum eucalyptus, valley oak) present the highest tree heights compared to others, such as the common crapemyrtle or the pinyon pine, which present much smaller heights. City planners might use this information to decide which trees to include in their city landscape plans.
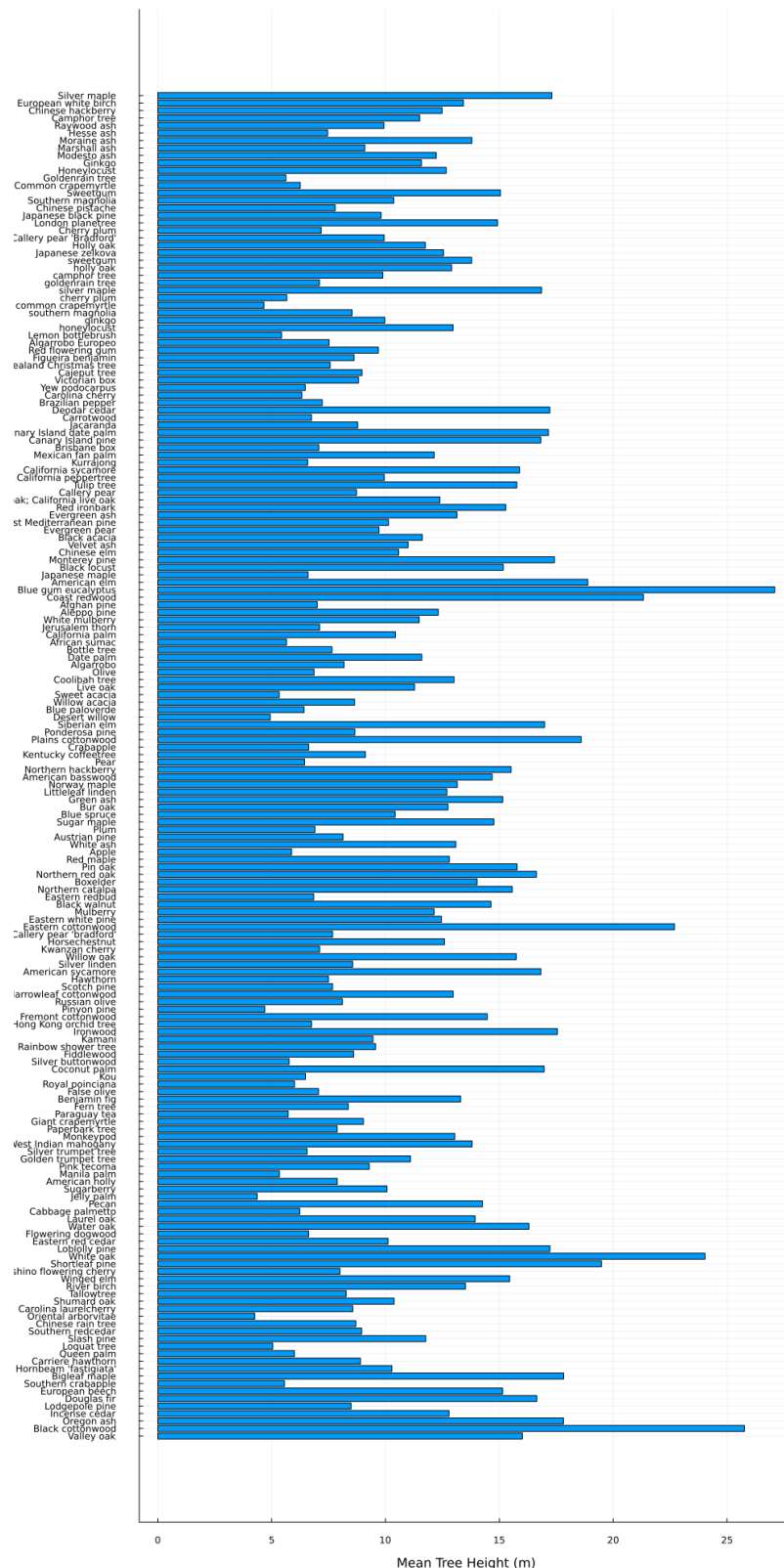
Figure 8: Tree Height by Species.

Then, barplots of tree height by city and region were investigated to develop a deeper understanding of spatial tree height distributions. The following figures present how the average tree height varies by city and region.
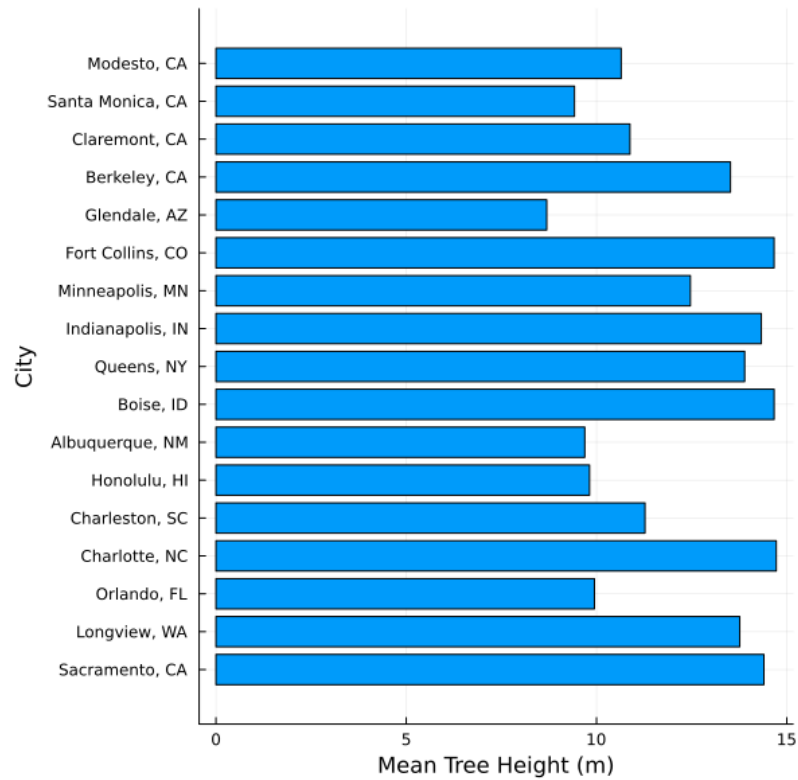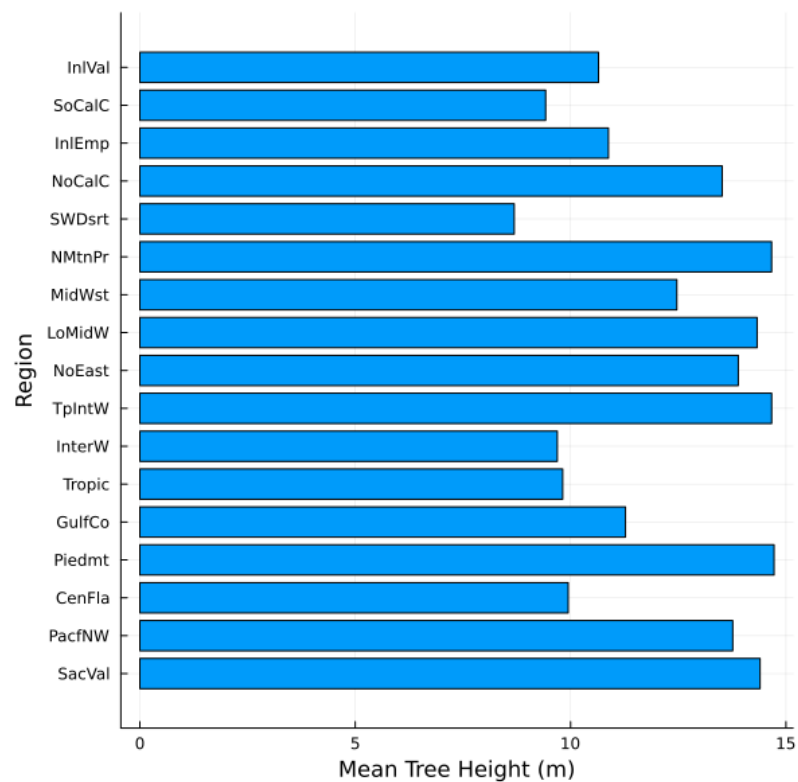
Figure 9: Tree Height by City.

Figure 10: Tree Height by Region.

From these figures, it can be observed that there are distinct spatial differences among tree height distributions. For example, trees tend to be taller in more mountainous regions and shorter in desert regions, and this aligns well with the results observed in Fort Collins, CO and Albuquerque, NM, respectively.

Moreover, an exploration of the presence of utility lines and their impact on the growth of trees was conducted. For this analysis, four variables were selected and filtered to find the correlation between the presence of utility lines and the growth of trees. These variables include "WireConf," "Age," "TreeHt," and "DBH." The "WireConf" variable is a categorical variable that presents whether the utility lines interfere with or appear above a tree. This variable might include one of

five values: 0= no lines, 1= lines present and with no potential conflict, 2= lines present and conflicting, and 3= lines present with potential for conflicting, while any values with "-1" denote data that was not collected. The "Age" variable is a numerical variable that presents the number of years since the tree was planted. The "TreeHt (m)" variable is a numerical variable that presents tree height from ground to the treetop to the nearest 0.5 m. The "DBH" variable is a numerical variable that presents the diameter of the tree at breast height (1.37 meters [m]) measured to the nearest 0.1 centimeters.

The first step in analyzing the effect of wire conflict on the dataset was to group the data by "WireConf" to discover how many trees in the database were affected. Figure 11 shows the percentage of trees in the database in each category after excluding all trees that do not have data. Figure 11 shows that 71% of trees in the database are not in areas that have utility lines conflicting with them, which will help to compare tree growth when no utility lines are present vs when utility lines are present.
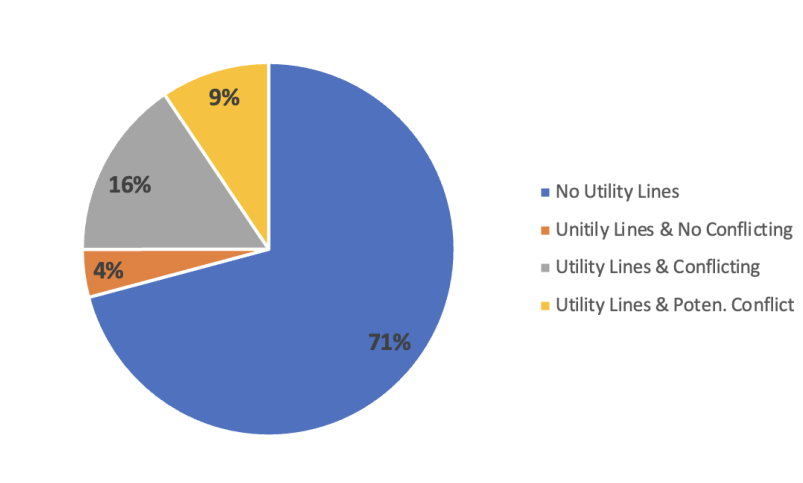

Figure 11: Number of Trees in Each Category in The Database.

The second step was to calculate the average height of trees for each of the aforementioned categories as shown in Figure 12. The average tree height in all categories varies from 10 to 13 meters, which does not clarify the impact of the growth of trees with the presence of the utility lines. Therefore, further investigation is needed.
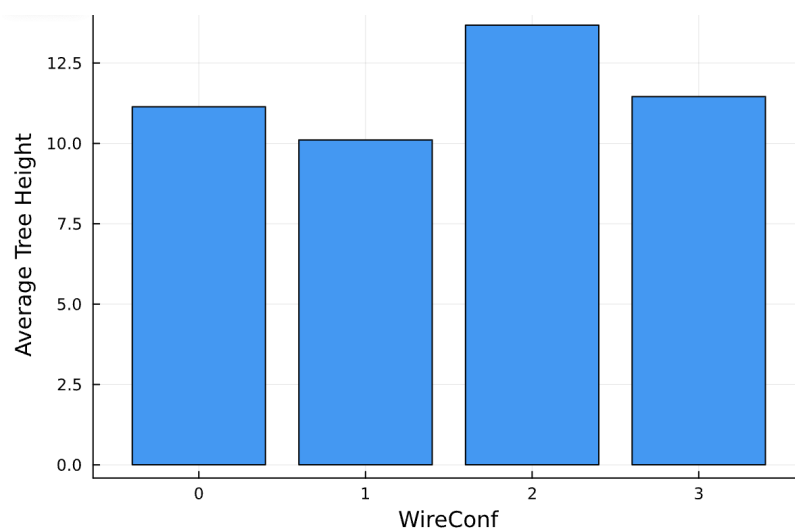
Figure 12: Average Tree Height Based on Wire Conflict.

The third step was to find the correlation between the age and the height of trees for each of the aforementioned categories. Figure 13 shows that there is a strong correlation between tree age and average tree height in all categories. The calculated correlation in all categories is higher than 0.7. Additionally, in all categories, the correlation is almost the same under the age of 50 years then, it started to be slightly different in each category as shown in Figure 14.

Figure 13: The Correlation between Tree Age and Average Tree Height Based on Wire Conflict

Figure 14: The Correlation between Tree Age and Average Tree Height Based on Wire Conflict

The fourth step was to analyze the correlation between the average diameter of a tree and its age in each category. Figure 15 shows that there is a moderate-to-strong correlation between the average DBH and tree age in all categories. The calculated correlation in all categories is higher than 0.8, see Figure 16.
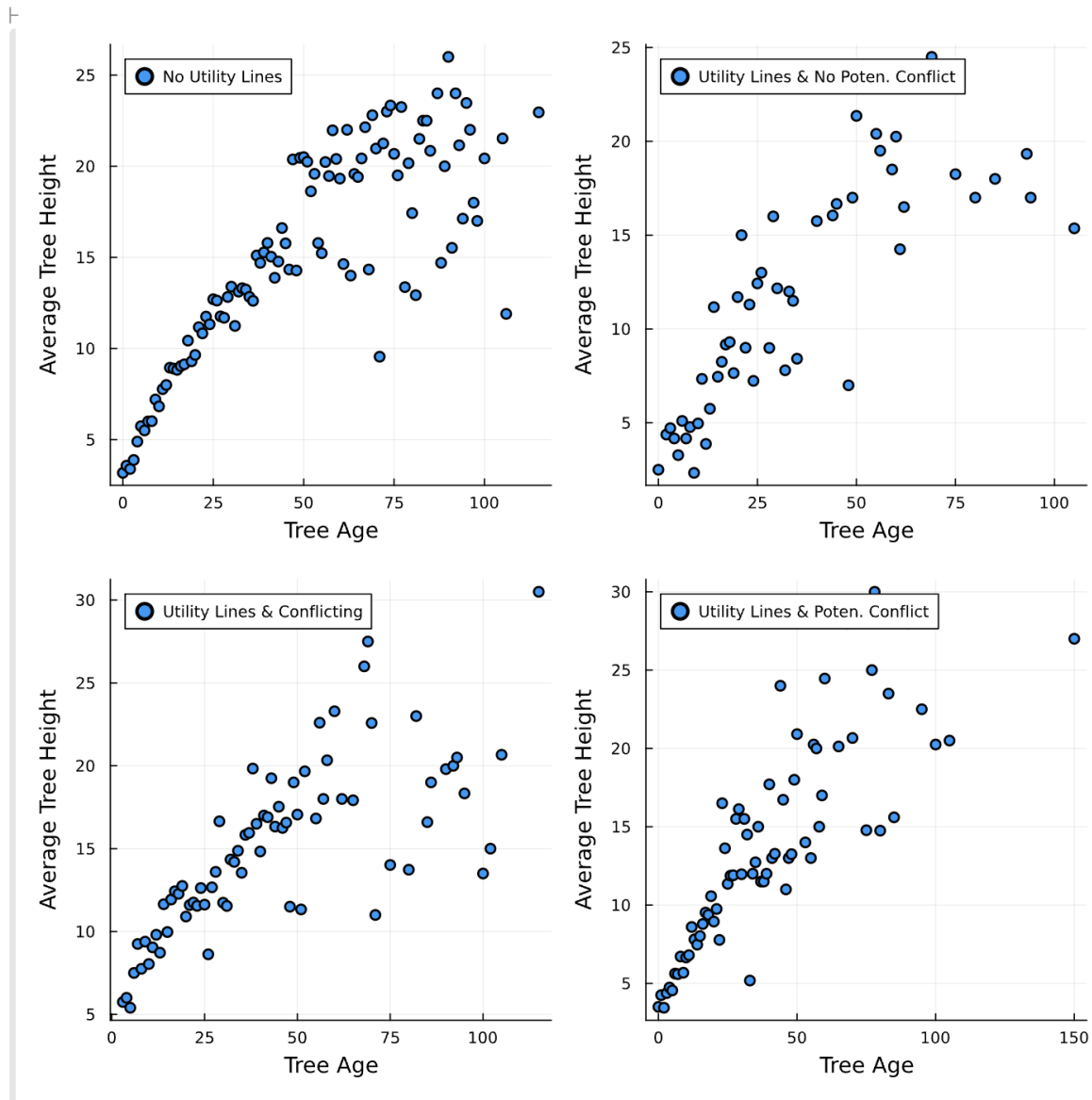
Figure 15: The Correlation between Tree Age and Average Diameter of Trees based on Wire Conflict
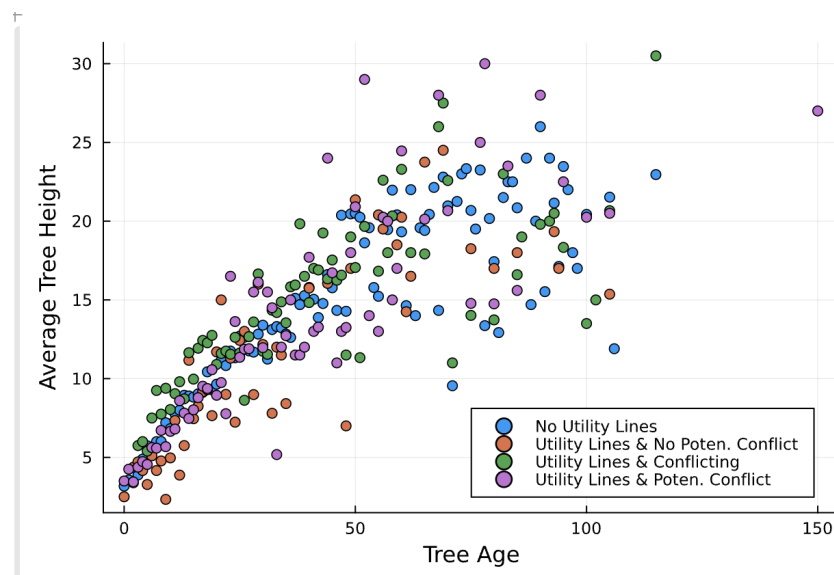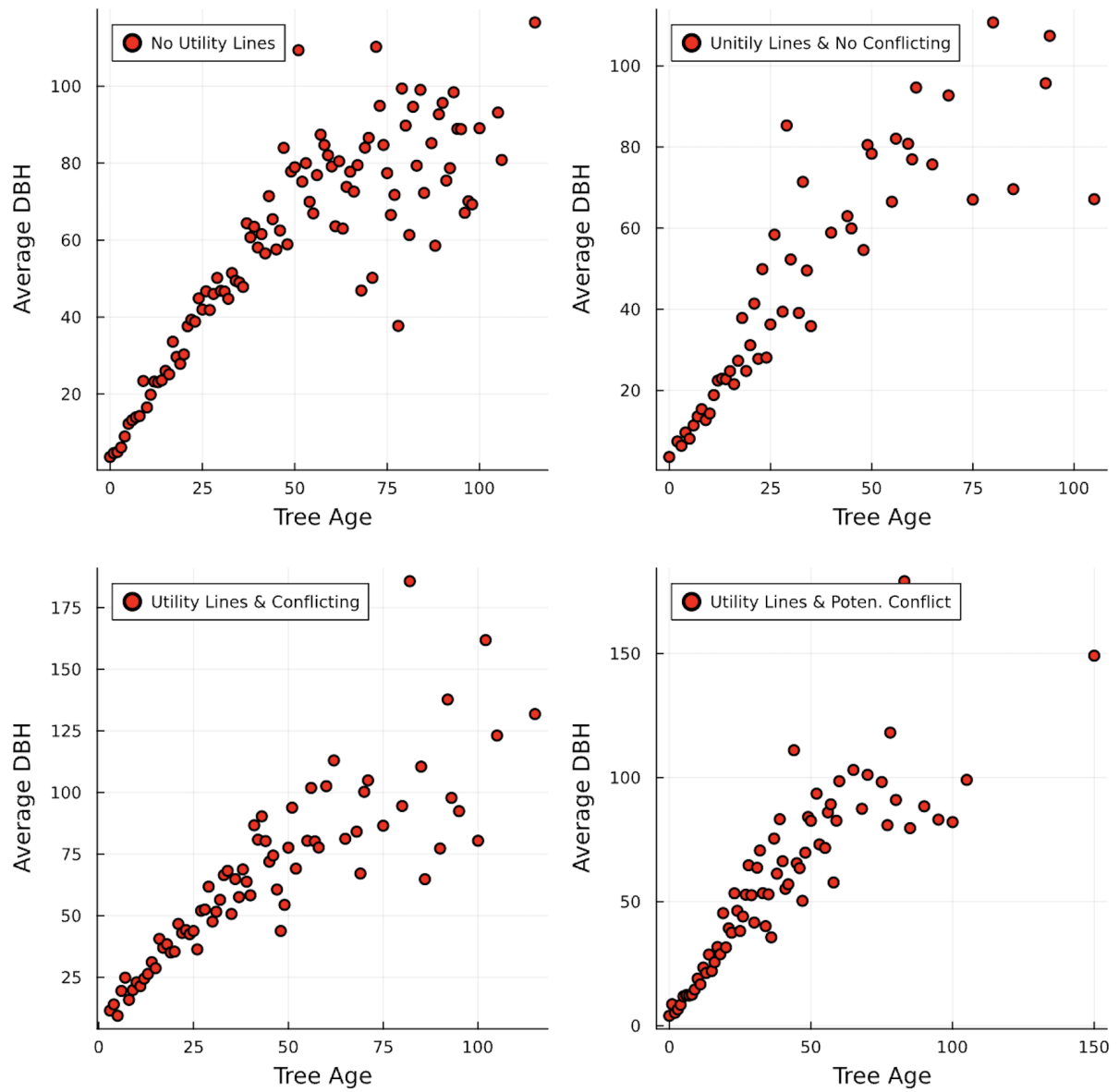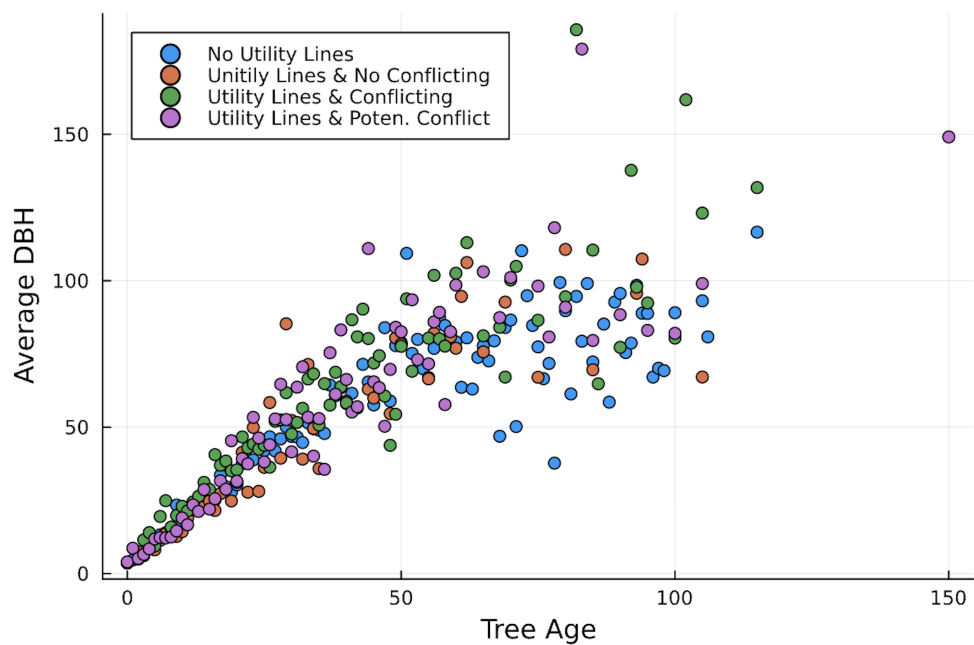

Figure 16: The Correlation between Tree Age and Average Diameter of Trees Based on Wire Conflict

The last step was to find the correlation between the height and diameter of trees to validate its use in estimating the tree height based on its diameter. Figure 17 presents the correlation between the two aforementioned variables. The figure shows a moderate correlation between tree height and its diameter, and the calculated correlation is 0.78.
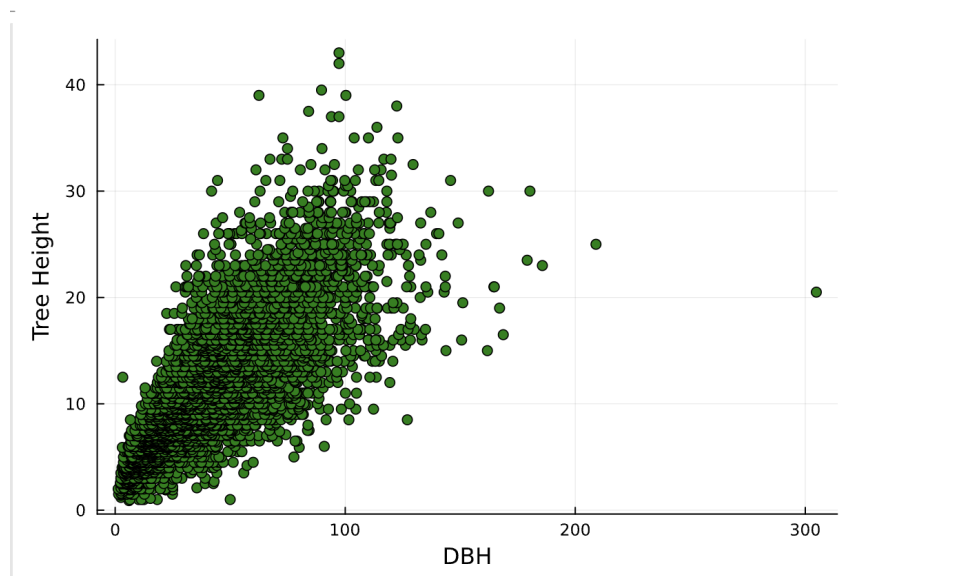

Figure 17: The Correlation between Height and Average Diameter of Trees

Therefore, the presence of utility lines does not have a great impact on the growth of trees.

Next, the relationships among tree species, tree height, land use, and location were explored to identify any plausible correlations for the purpose of urban tree planning. One may consider how urban city planners select particular species of tree to plant within specific land use types. For example, perhaps an urban planner might select a particular tree species based on average height or canopy size (leaf area) in order to provide suitable landscaping along a street and provide sufficient shade to city goers without intercepting overhead telephone lines or buildings. Furthermore, these data were grouped by city and region to investigate spatial differences among the variables.

Next, a bar plot depicting the average land use (which was calculated by rounding the mean land use type across species, where land use contains the following categories: 1=single family residential, 2=multi-family residential, 3=industrial/institutional/large commercial, 4=park/vacant/other, 5=small commercial, 6=transportation corridor) was created to visualize which species might be more commonly associated with a land use type. Based on the results in Fig. 18, it appears that some tree species are more frequently linked to specific land use types (i.e., evergreen ash trees to small/commercial land uses or both willow acacia and japanese maple to single family residential land uses).
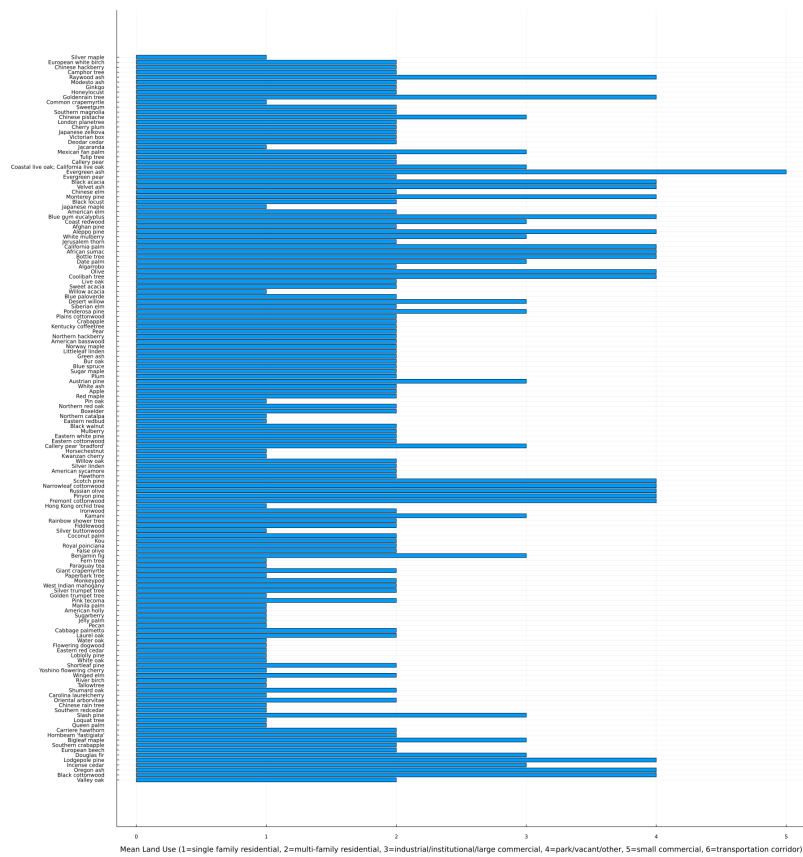
Figure 18: Tree Species by Average Land Use.

Additionally, the correlations among tree height, DBH, crown height, and leaf area were explored to illustrate quantitative factors that urban planners might consider when redesigning a site. Moreover, the US Forest Service Research Archives, from which the raw tree data was obtained, describes how variables such as tree age can be used to predict a species diameter at breast height (dbh), which can in turn predict tree height, crown diameter, crown height, leaf area, and tree age [2]. Extending the investigation to include these considerations, tree height, DBH, crown height, and leaf area variables were selected and their correlations were calculated. Figures 19, 20,21 and 22 depict marginal histograms, which are useful in explaining the distributions of each variable as well as how they are correlated.
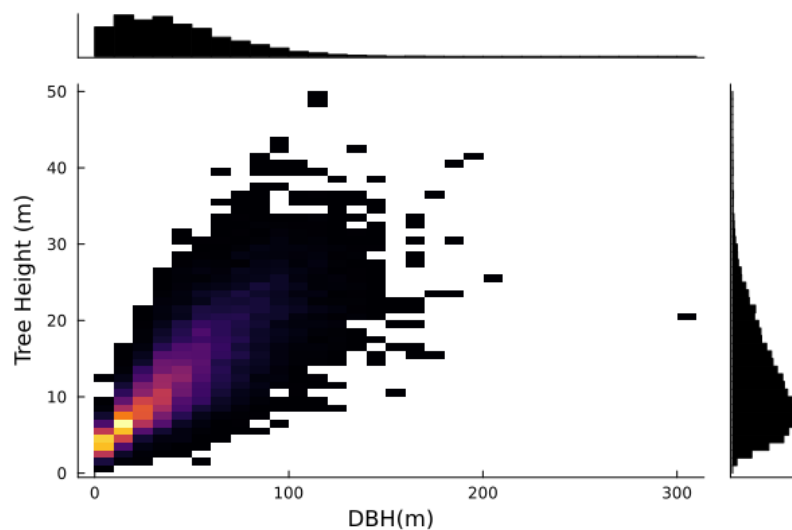
Figure 19: Marginal Histogram of DBH and Tree Height.
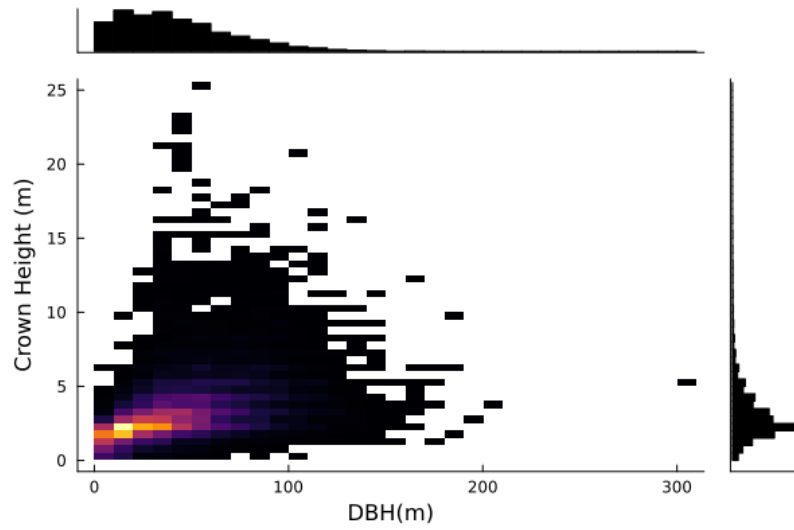
Figure 20: Marginal Histogram of DBH and Crown Height.
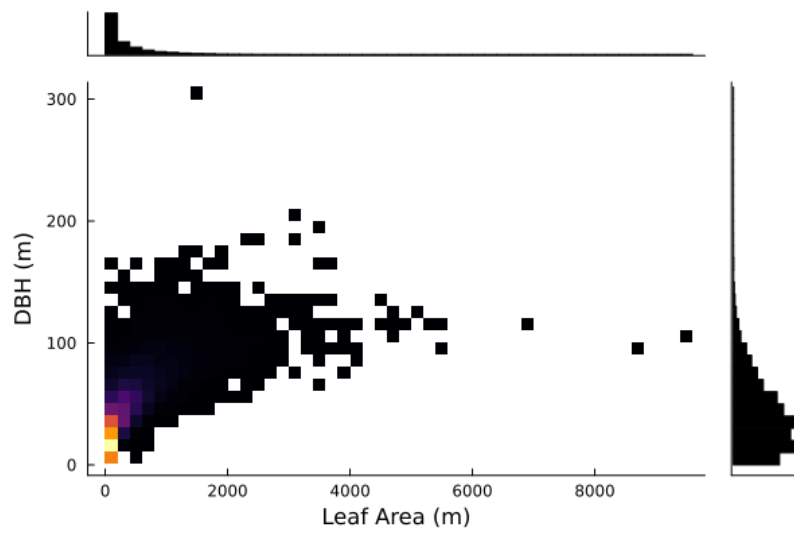
Figure 21: Marginal Histogram of Leaf Area and DBH.
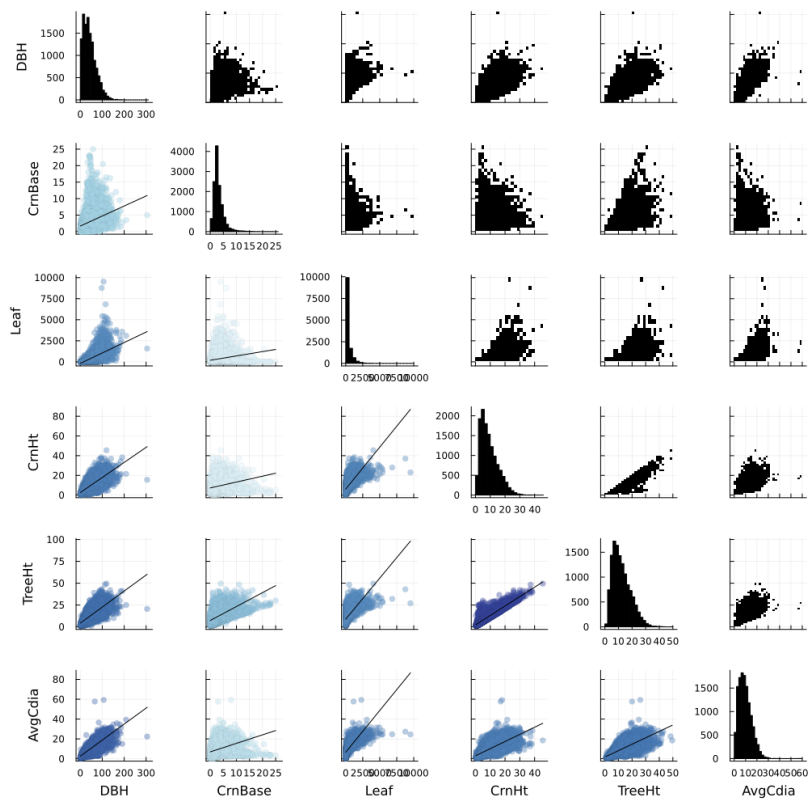
Figure 22: Correlation Plot of DBH, Tree Height, Crown Height, Leaf Area, and others.

To investigate these above relationships further, average DBH by tree heights was grouped by cities to illustrate how the two variables are related in different cities. The following figures visualize these relationships and show a moderate-to-strong positive correlation between average DBH and tree height across different cities. Several cities were randomly chosen out of all available cities. The correlations between average DBH and tree height are also listed below.
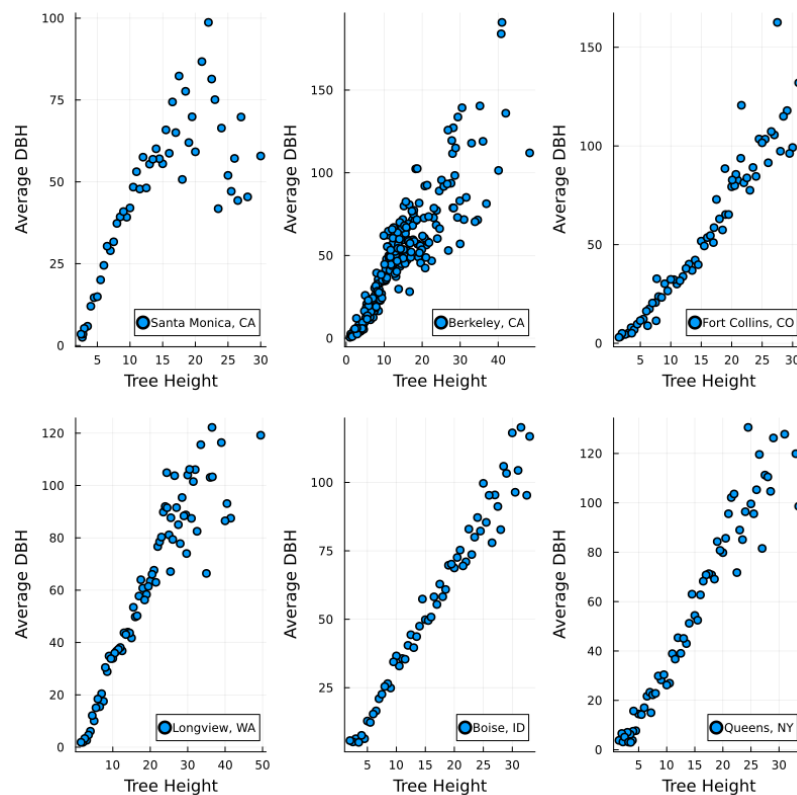
Figure 23: Average DBH vs Tree Height by City.
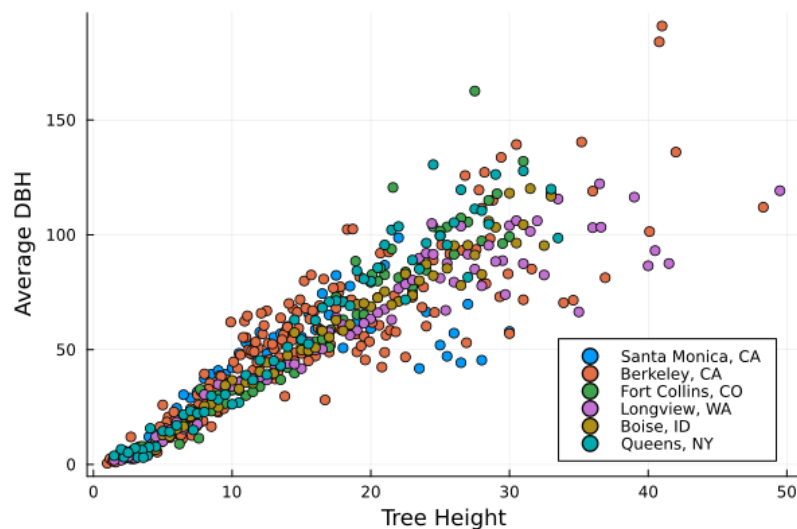
Figure 24: Average DBH vs Tree Height by City - Combined.

Correlation Coefficients:

1. DBH vs Tree Height Overall: 0.802

2. DBH vs Tree Height for randomly chosen cities:

   Santa Monica, CA: 0.713

   Berkeley, CA: 0.889

Fort Collins, CO: 0.959

Longview, WA: 0.933

Boise, ID: 0.985

Queens, NY: 0.970

Leaf Area vs DBH: 0.713

3. DBH vs Crown Base Height: 0.422

In summary, this exploratory analysis has shown both strong and insignificant correlations among raw tree data variables. The relationship between tree height and setback revealed insignificant correlation , while moderate-to-strong correlations between tree age and both height and diameter of tree exist. Additionally, correlation between tree age and its diameter is stronger than that of tree age and height. Overall, relationships among tree size and tree growth are significant because they can be used by urban forest managers, landscape architects, and city planners to select suitable trees given limited growing space or an intended purpose. Predicting the most suitable trees for a site has the potential to reduce costly future conflicts between trees and infrastructure [2].

Predictive Modeling

Based on some of the above correlations and supporting evidence from the US Forest Service Research Archives [2], there are strong correlations among diameter at breast heights (dbh) and tree age, tree height, leaf area, crown height, and average crown diameter. Therefore, several predictive models using 3 different machine learning techniques were explored to select the most suitable model to predict dbh based on the aforementioned variables. Those techniques include (1) decision-tree algorithm, (2) regression, and (3) neural networks. To enable a reliable performance evaluation procedure, the collected data was divided into two separate datasets for each developed predictive model: (1) training dataset that includes 70% of all the available data that will be used in developing the model, and (2) testing dataset that includes 30% of all available data that will be used for evaluating the performance of the developed model. The following three sections provide a detailed description of these aforementioned 3 machine learning techniques.

Decision-Tree Algorithm

A regression decision tree model was run with two, three, four, and five independent variables in order to predict DBH. The output of these iterations are shown in Figure 25 :

Figure 25: Decision Tree using 2,3,4, and 5 independent variables: Tree Height, Age, Leaf Area, Crown Height, and Average Crown Diameter

Figure 25: Decision Tree using 2,3,4, and 5 independent variables: Tree Height, Age, Leaf Area, Crown Height, and Average Crown Diameter

With each addition of independent variables, the mean coefficient of determination increased, and the mean squared error decreased. But overall, the decision tree model for this data is not robust and does not do a great job at fitting the data. The correlation coefficient gets increasingly closer to one with additional variables, meaning the linear relation is better with more variables and DBH than a few variables and DBH. This finding shows that a linear regression model might be a better method for modeling this dataset.

A classification decision tree was also tested to see if the tree type could be predicted using crown height, age, DBH, and tree height. This model was run with three folds similarly to the previous decision tree. The resulting model had very low accuracy of 0.369198, 0.377247, and 0.376442. These values of accuracy suggest that predicting tree type in this manner is not reliable.

Regression

This technique depends on developing multiple linear regression models among a dependent variable and independent variables. Aiming to improve the performance of a predictive model, we constructed a simple regression model that uses one independent variable (tree age) to predict the dependent variable (average dbh). To visualize the model performance, we calculated the coefficient of determination, Root Mean Square Error (RMSE), and model accuracy. As shown in Figure 26, the model achieved an R-squared of 74%, RSEM of 21.12, and a lpw accuracy of 12.4%.

```
model1 =
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}}, GLM.DensePredCho

DBH_mean ~ 1 + Age

Coefficients:
─────────────────────────────────────────────────────────────────────────
                 Coef.   Std. Error      t  Pr(>|t|)  Lower 95%  Upper 95%
─────────────────────────────────────────────────────────────────────────
(Intercept)  32.7814      3.3241       9.86    <1e-14   26.1662    39.3965
Age           0.607107    0.0402932   15.07    <1e-24    0.526921   0.687293
─────────────────────────────────────────────────────────────────────────
```

Figure 26: Predictive Model using one independent variable

We investigated further to find an explanation for the model's poor performance. We plotted the average dbh in the y-axis and age in the x-axis to visualize the training and testing datasets. We noticed that two data points were outliers, as shown in Figure 27.
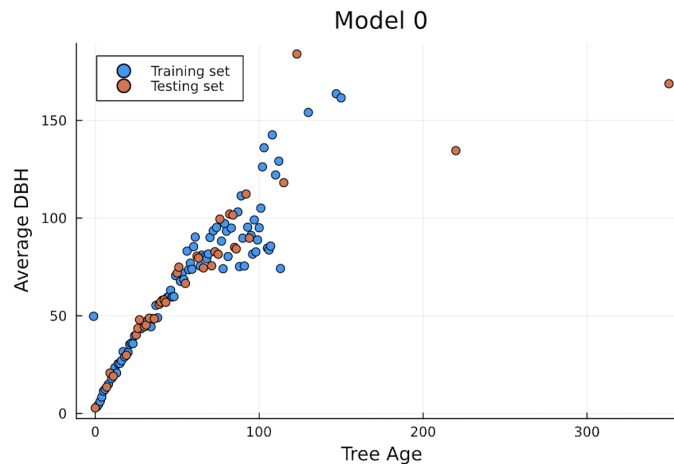
Figure 27: Determing outliers

A second model was performed using the same dependent and independent variables after excluding the outliers. These two outliers were tree ages above 200 years. The model achieved an R-squared of 88%, RMSE of 10.91, and accuracy of 87%, as shown in Figure 28.

```
model2 =
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}}, GLM.DensePredCho

DBH_mean ~ 1 + Age

Coefficients:
─────────────────────────────────────────────────────────────────────────
                 Coef.   Std. Error      t  Pr(>|t|)  Lower 95%  Upper 95%
─────────────────────────────────────────────────────────────────────────
(Intercept)  13.1109      2.76822      4.74    <1e-05    7.60084   18.6209
Age           0.962126    0.0401725  23.95    <1e-37    0.882165   1.04209
─────────────────────────────────────────────────────────────────────────
```

```julia
  model2 = lm(@formula(DBH_mean ~ Age), train2) #first try predict dph
```

```julia
  #predict2= (13.9761 .+ 0.980197 .* test2.Age)[:]
```

```julia
predict2 =
  [16.9594, 20.8079, 21.77, 23.6943, 24.6564, 33.3155, 35.2398, 40.0504, 41.9747, 43.8989, 4
  predict2= (13.1109 .+ 0.962126 .* test2.Age)[:]
```

```julia
10.910070096393538
  rmse(predict2, test2.DBH_mean[:])
```

```julia
R2_2 = 0.88
  R2_2= round.(r2(model2),digits=2)
```

```julia
accuracy_model2 = 0.8736328172544511
  accuracy_model2= acc(predict2, test2.DBH_mean[:])
```

Figure 28: Predictive Model after Deleting Outliers

A third model was performed using two independent variables instead of one: tree age and average tree height. This model achieved an R-squared of 94%, RMSE of 8.08, and accuracy of 92%, as shown in Figure 29.

```
model3 =
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}}, GLM.DensePredCho

DBH_mean ~ 1 + Age + TreeHt_mean

Coefficients:
─────────────────────────────────────────────────────────────────────────
                  Coef.   Std. Error      t  Pr(>|t|)   Lower 95%  Upper 95%
─────────────────────────────────────────────────────────────────────────
(Intercept)   -5.30844     3.15756    -1.68     0.0967   -11.5946   0.977778
Age            0.568004    0.0531009  10.70     <1e-16     0.462289  0.67372
TreeHt_mean    2.4595      0.305188    8.06     <1e-11     1.85192   3.06708
─────────────────────────────────────────────────────────────────────────
```
  · model3= lm(@formula(DBH_mean ~ Age+TreeHt_mean), train3)

```
predict3 =
▶[2.62237, 6.05855, 15.0031, 17.4977, 23.3167, 27.6365, 31.7028, 33.4565, 43.0589, 43.7155,
```
  · predict3 = (-5.30844 .+ 0.568004 .* test3.Age .+ 2.4595.* test3.TreeHt_mean)[:]
                                                                            ⏵ 107 µs

```
8.085818351750811
```
  · rmse(predict3, test3.DBH_mean[:])

```
R2_5 = 0.94
```
  · R2_5= round.(r2(model3),digits=2)
                                                                            ⏵ 12.4 µs

```
accuracy_model3 = 0.9197077479207694
```
  · accuracy_model3= acc(predict3, test3.DBH_mean[:])

Figure 29: Predictive Model Using 2 Independent Variables

A fourth model was performed using three independent variables: tree age, average tree height, and average leaf area. The model achieved an R-squared of 95%, RMSE of 9.01, and accuracy of 92%, as shown in Figure 30.

```
model4 =
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}}, GLM.DensePredCho

DBH_mean ~ 1 + Age + TreeHt_mean + Leaf_mean

Coefficients:
─────────────────────────────────────────────────────────────────────────
                  Coef.   Std. Error      t  Pr(>|t|)   Lower 95%  Upper 95%
─────────────────────────────────────────────────────────────────────────
(Intercept)   -1.3087      3.29214    -0.40     0.6921   -7.86418   5.24678
Age            0.603128    0.0608102   9.92     <1e-14     0.482039  0.724216
TreeHt_mean    1.80091     0.364622    4.94     <1e-05     1.07486   2.52697
Leaf_mean      0.00908568  0.0036555   2.49     0.0151     0.00180666  0.0163647
─────────────────────────────────────────────────────────────────────────
```
  · model4= lm(@formula(DBH_mean ~ Age+TreeHt_mean+Leaf_mean), train4)

  · #predict4 = (-2.59117 .+ 0.509142 .* test4.Age .+ 2.18325.* test4.TreeHt_mean .+
    0.00941475 .* test4.Leaf_mean)[:]

```
predict4 =
▶[9.83018, 12.8424, 17.5271, 21.8187, 25.0372, 26.587, 29.8762, 35.1606, 37.4005, 41.4902,
```
  · predict4 = (-1.3087   .+ 0.603128 .* test4.Age .+ 1.80091.* test4.TreeHt_mean .+
    0.00908568 .* test4.Leaf_mean)[:]
                                                                            ⏵ 117 µs

```
9.01108213518639
```
  · rmse(predict4, test4.DBH_mean[:])

```
R2_6 = 0.95
```
  · R2_6= round.(r2(model4),digits=2)

```
accuracy_model4 = 0.9180848363958132
```
  · accuracy_model4= acc(predict4, test4.DBH_mean[:])

Figure 30: Predictive Model Using 3 Independent Variables

The fifth model was performed using four independent variables: tree age, average tree height, average leaf area, and average crown diameter. The model achieved an R-squared of 93%, RMSE of 7, and accuracy of 94.5%, as shown in Figure 31.

```
model5 =
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}, GLM.DensePredCho

DBH_mean ~ 1 + Age + TreeHt_mean + Leaf_mean + AvgCdia_mean

Coefficients:
─────────────────────────────────────────────────────────────────────────
                  Coef.   Std. Error      t   Pr(>|t|)   Lower 95%   Upper 95%
─────────────────────────────────────────────────────────────────────────
(Intercept)    -6.70194      3.68984   -1.82     0.0733   -14.0509     0.647015
Age             0.457378   0.0669582    6.83    <1e-08     0.32402     0.590737
TreeHt_mean     2.34818       1.13877    2.06     0.0426    0.080124     4.61623
Leaf_mean     -0.0461712      1.09158   -0.04     0.9664    -2.22024      2.1279
AvgCdia_mean    0.767212     0.590222    1.30     0.1976   -0.408317     1.94274
─────────────────────────────────────────────────────────────────────────
```

- `model5= lm(@formula(DBH_mean ~ Age+TreeHt_mean+Leaf_mean+AvgCdia_mean), train5)`

```
cor_age_avgdbh = 0.906
```

- `cor_age_avgdbh= round.(cor(train5.AvgCdia_mean, train5.DBH_mean), digits=3)`

  ⊙ 16.5 µs

```
predict5 =
▸[3.73975, 19.1645, 20.3533, 28.1565, 32.9219, 39.3908, 42.6755, 43.4538, 45.0554, 43.1593,
```

- `predict5 = (-6.70194 .+ 0.457378 .* test5.Age .+ 2.34818.* test5.TreeHt_mean .+ -0.0461712 .* test5.Leaf_mean .+ 0.767212 .* test5.AvgCdia_mean)[:]`

```
7.001448879471186
```

- `rmse(predict5, test5.DBH_mean[:])`

```
R2_5 = 0.93
```

- `R2_5= round.(r2(model5),digits=2)`

  ⊙ 47.0 µs

```
accuracy_model5 = 0.9451709898442938
```

- `accuracy_model5= acc(predict5, test5.DBH_mean[:])`

  ⊙ 10.3 µs

Figure 31: Predictive Model Using 4 Independent Variables

Based on the above analysis, the best model that achieved the lowest RMSE and highest accuracy was Model 5.

Neural Network

This final technique involves two main approaches. The first builds a simple linear regression neural network of one tree characteristic input to one tree characteristic output. First, the tree data was filtered into "DBH", "TreeHt", "Age", and "CrnBase." Then, "DBH" was chosen as the input variable and "TreeHt" was selected as the output or dependent variable for prediction. The observations in the dataframe were then shuffled to prepare for splitting the dataset into 50% training and 50% testing data. Following the splitting and reshaping of the data, a linear regression neural network was constructed using the Julia Flux package, one dense layer with one input and one output channel, gradient descent as the optimization approach, and a mean square error (MSE) loss function. After running through 12 epochs or iterations, the neural net predicted 88% of tree heights from given DBH values. The RSME associated with this model was roughly 13.8. This model performance was relatively decent compared to other models, but it was still insufficient to generate valid predictions.

Moving on to a more complex neural network, the second technique tackled more input or independent variables to predict tree species. After isolating relevant tree characteristic input features such as "TreeType", "Age", "DBH", "TreeHt", "CrnBase", "CrnHt", "CdiaPar", "CDiaPerp", "AvgCdia", and "Leaf" and filtering out unwanted missing data, the neural net was structured to predict tree species, or "CommonName." 157 unique tree species names were identified, and these were manipulated to create unique integer indeces mapped to each unique "CommonName." Following appropriate reshaping and data re-structuring to meet the required input format in the Flux Package, a convolutional neural network (CNN) with 5 convolutional layers of varying filter size and 2 dense layers was built and run over 125 epochs with ADAM as the optimization algorithm and cross entropy as the loss function. ADAM, as opposed to stochastic gradient descent, is able to incorporate concepts of momentum rather than randomness to push the gradient descent algorithm out of local minima and isolate the global minima. Unfortunately, the result of this CNN produced a very low accuracy of 4%.

Due to this low accuracy, a more simplified NN on tree species was performed with only 2 dense layers and no convolutional filters. With the same optimization and loss functions that were used in the previous CNN, the result of this neural net also produced a low prediction accuracy of 4.5%. Though this increased slightly, the poor accuracy presents a larger concern regarding the strategy of data prediction. Because only 9 input features were used to predict 157 unique tree species, it is more likely that the resulting model performance was not due to the model itself but rather due to an extremely high ratio of tree species to tree characteristics such as height or DBH. Excess variety in the number of output variables made it difficult to accurately predict tree species with so few input characteristics.

Because of this, a NN was constructed to predict tree type (which has 11 unique tree types) instead of tree species to simplify the number of predicted outputs. Recall how tree types are 3 letter codes, where the first two letters refer to life form (BD=broadleaf deciduous, BE=broadleaf evergreen, CE=coniferous evergreen, PE=palm evergreen) and the third letter refers to the tree's mature height (S=small, which is < 8 meters, M=medium, which is 8-15 meters, and L=large, which is > 15 meters). Starting with only 2 dense layers, this neural network yielded a better but still poor model accuracy of 41%.

To improve this, a CNN was performed on tree type with 5 convolutional layers of increasing and decreasing filter sizes and 2 dense layers. Although CNNs are typically used for problems involving spatial patterns, we tried building one anyway to see if prediction accuracy could be improved. A slightly better accuracy of 47% was in fact achieved, which could suggest how more complex convolutional layers might yield more effective model performances.

The following image shows the accuracy of the CNN using 5 convolutional layers and 2 dense layers to predict tree type over many iterations.
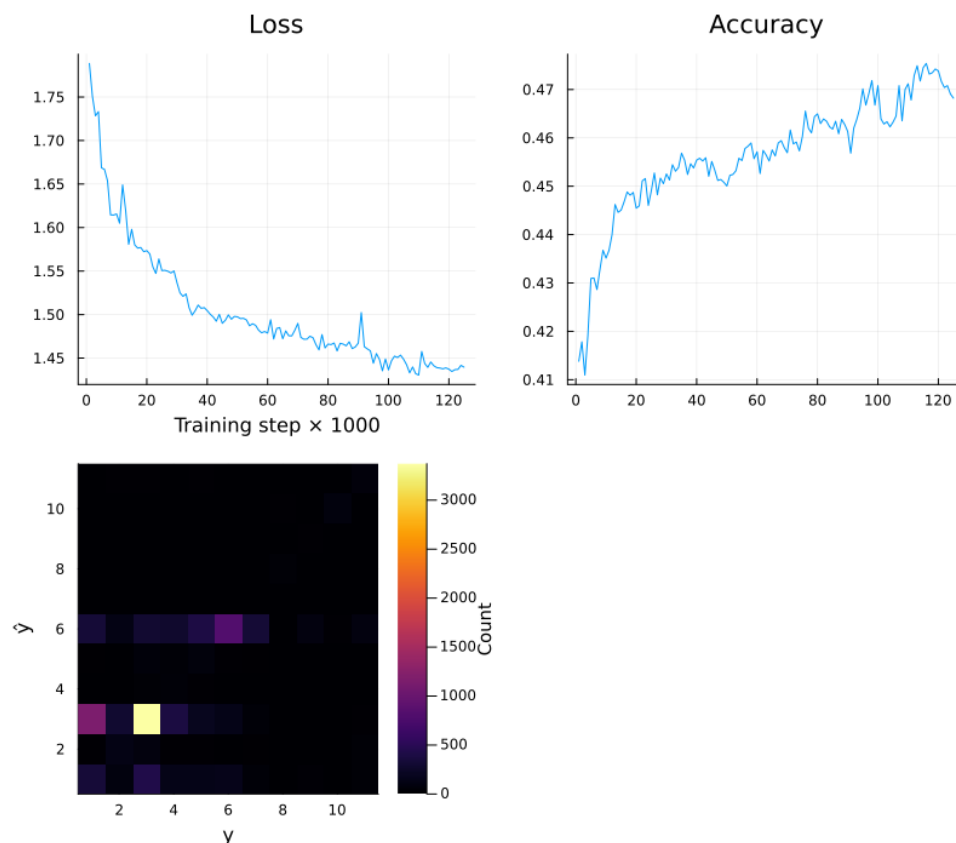

Figure 32: Convolutional Neural Network Predicting TreeType

Summary of Model Comparison

To summarize all models that were performed, the table below shows the inputs, outputs, and associated accuracies for each model in terms of R-squared and RSME.

Discussion and Conclusion

In summary, the variety of characteristics present in the raw tree data set posed quite a challenge when we approached the construction of a well-performing predictive model. Unfortunately, none of the three predictive models (decision tree, linear regression, and neural net) illustrated sufficient performance, though some were better than others, such as the linear regression model using four independent variables (tree age, average tree height, average leaf area, and average crown diameter) and the simple neural net using one input variable (DBH) to predict one output variable (TreeHt). [.... TALK ABOUT CORRELATION MATRIX TO SELECT VARIABLES OF HIGHER CORRELATION HERE? ....]

Interestingly, a general pattern emerged. By increasing the number of independent variables used in any of the models, predictive performance improved and the models' error decreased.

[Add discussion about why raw tree data is hard to accurately predict? Talk about what we hoped to use these models for, and perhaps what might be done in the future to study tree data in a better way? Think bigger picture meaning,

outside of the analyses we did.]

References

[1] Oregon State University. 2022. Environmental factors affecting plant growth. OSU Extension Service. https://extension.oregonstate.edu/gardening/techniques/environmental-factors-affecting-plant-growth#:~:text=Environmental%20factors%20that%20affect%20plant,affect%20plant%20growth%20and%20development

[2] McPherson, E. Gregory; van Doorn, Natalie S.; Peper, Paula J. 2016. Urban tree database. Fort Collins, CO: Forest Service Research Data Archive. Updated 21 January 2020. https://doi.org/10.2737/RDS-2016-0005

[3] Kuuluvainen, T., Mäki, J., Karjalainen, L., & Lehtonen, H. (2002). Tree age distributions in old-growth forest sites in Vienansalo Wilderness, eastern fennoscandia. Silva Fennica, 36(1). https://doi.org/10.14214/sf.556