Analyzing the Correlations among Tree Characteristics and their Surroundings

This manuscript (permalink) was automatically generated from uiceds/cee-492-term-project-fall-2022-her@c352d59 on October 29, 2022.

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

The following will provide a narrative description and characterization of the tree dataset, interspersed with summary statistics and plots. Throughout this exploratory analysis, four main questions were investigated to guide data exploration:

1. How do power lines impact the growth of trees? (i.e., number of trees, leaf area, tree height, power lines)
2. How does setback (tree distance from heated/airconditioned spaces) show in different cities and/or regions? (i.e., correlation with tree height, leaf size, location)
3. What are the correlations between tree type, land use, height, leaf area, carshade, DBH, CdiaPar, and CDiaPerp for urban tree planning by region and/or city?
4. How does growth rate (i.e., height per age of tree) differ for each region, land use, city, etc.?

For each of these questions, the data was wrangled and filtered to generate visualizations of potential correlations among selected variables of interest.

Question 1

In this part, the research team were exploring if the presence of utility lines has an impact on the growth of trees. To answer this question, four variables were selected to be analyzed and filtered to find the correlation between the presence of utility lines and the growth of tree that include "WireConf" "Age", "TreeHt", and "DBH". The "WireConf" variable is a categorical variable that presents if the utility lines interfere with or appear above a tree. This variable might include one of five values, 0=no lines, 1=present and no potential conflict, 2=present and conflicting, 3=present and potential for conflicting, and -1 denotes data were not collected. The "Age" variable is a numerical variables that presents number of years since 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 diameter of tree at breast height (1.37 meters [m]) measured to nearest 0.1 centimeters.

The first step in our analysis is to group data by "WireConf" to discover how many trees in our database were affected. Figure 1 shows the percentage of trees in the database in each category after excluding all trees that do not have data, where 1= no lines, 2 = present and no potential conflict, 3 = present and conflicting, and 4 = present and potential for conflicting. It is clear that the majority of the trees are not in areas that have utility lines conflicting with trees which will help the research team to examine the growth of trees when there are no utility lines and compare it with the growth of trees when utility lines are present.

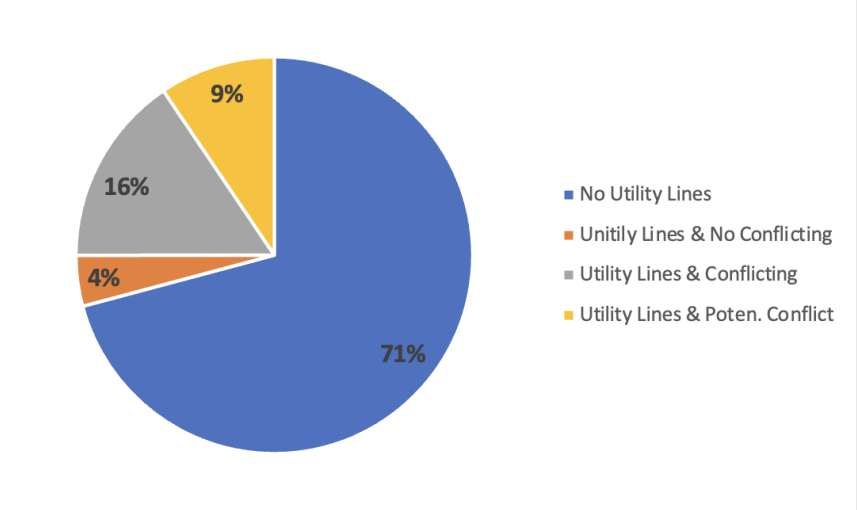


Figure 1: Number of Trees in Each Category in The Database.

The second step is to calculate the average height of trees for each of the aforementioned categories as shown in Figure 2 . The average tree height in all categories is varies from 10 to 13 meters which does not clarify the impact of the growth of tree with the present of the utility line. Therefore, further investigation is needed.

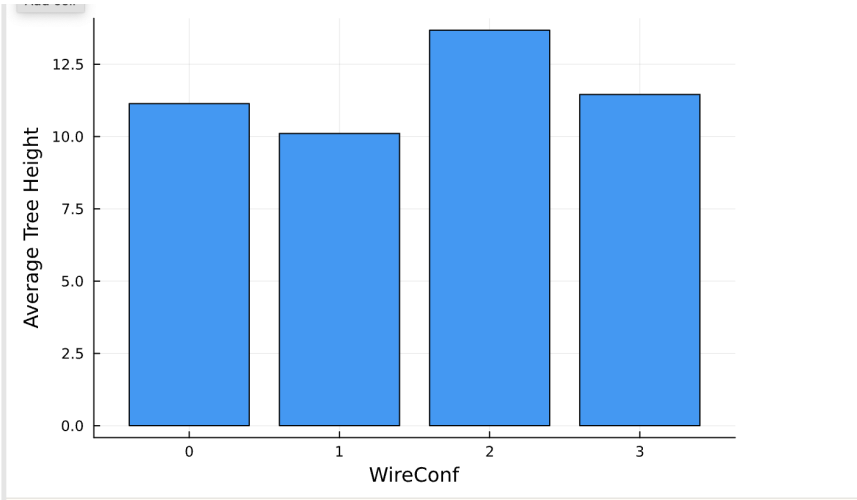


Figure 2: Average Tree Height Based on Wire Conflict.

The third step is to find the correlation between the age of trees and the height for each of the aforementioned categories, as shown in Figure 3. It is clear 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 as shown in Figure 4.

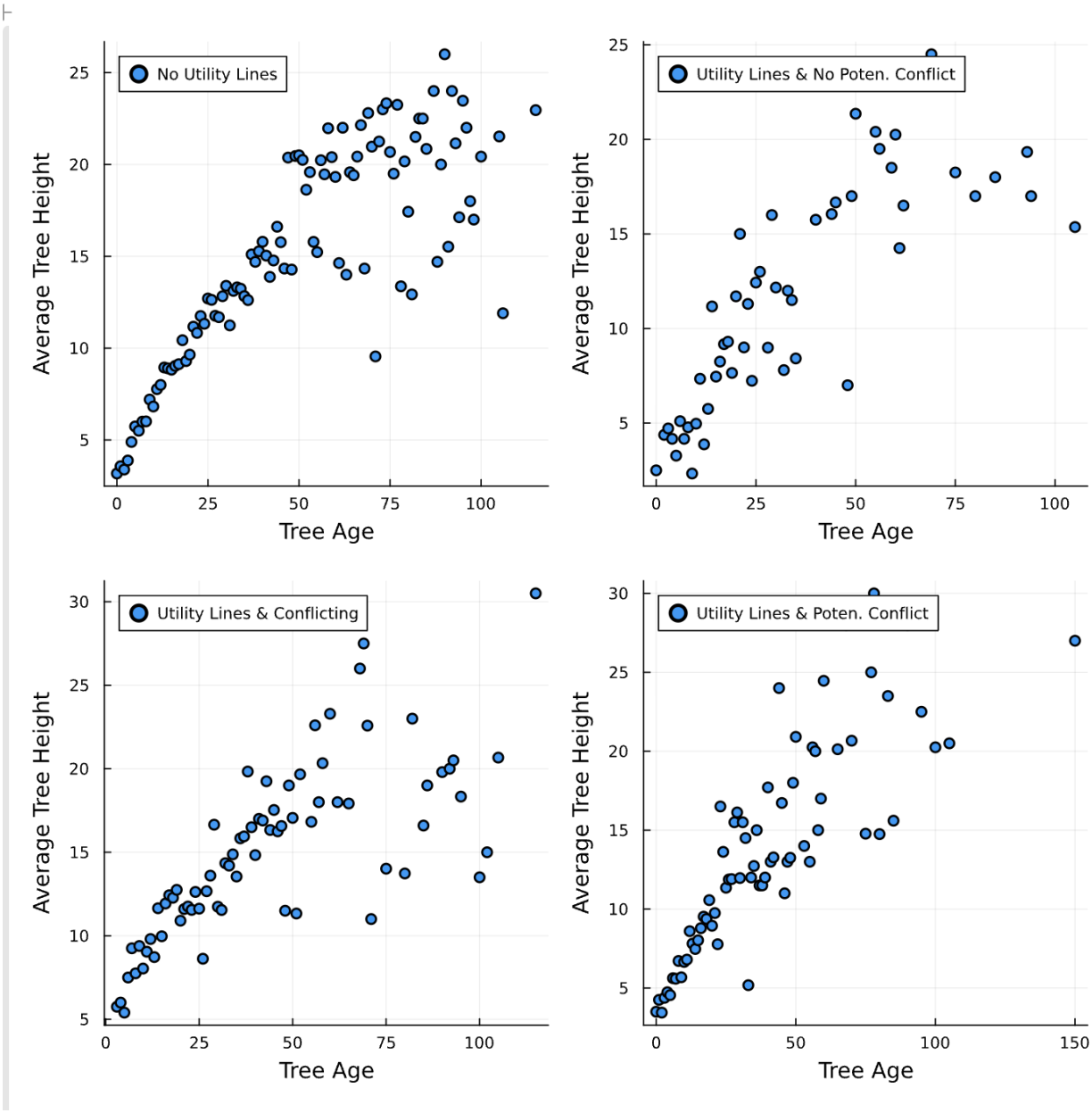


Figure 3: The Correlation between Tree Age and Average Tree Height Based on Wire Conflict

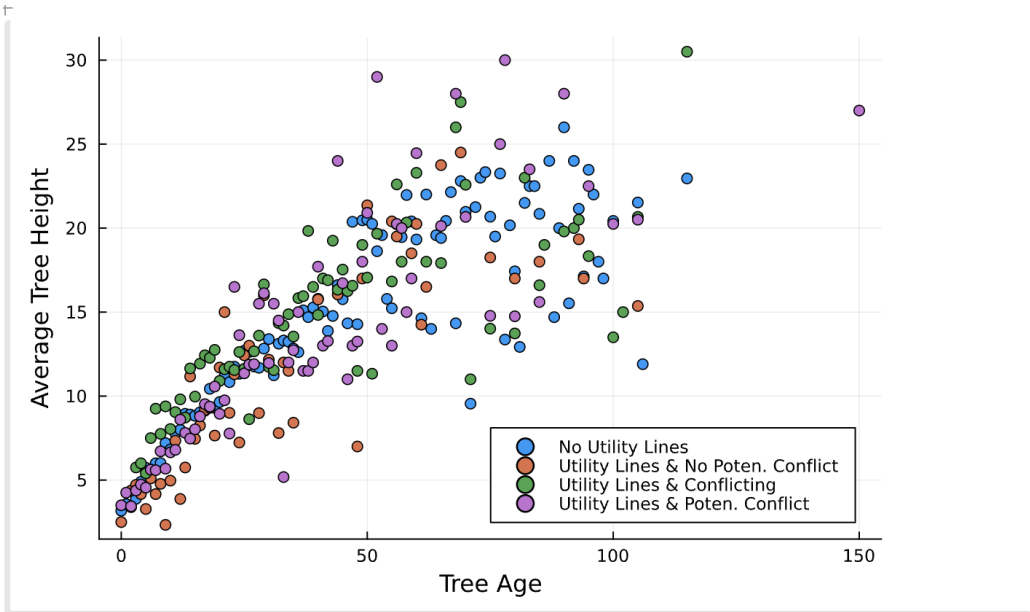


Figure 4: The Correlation between Tree Age and Average Tree Height Based on Wire Conflict

The fourth step is to analyze the correlation between the average diameter of tree and its age in each category. Figure 5 shows that there is a 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 6.

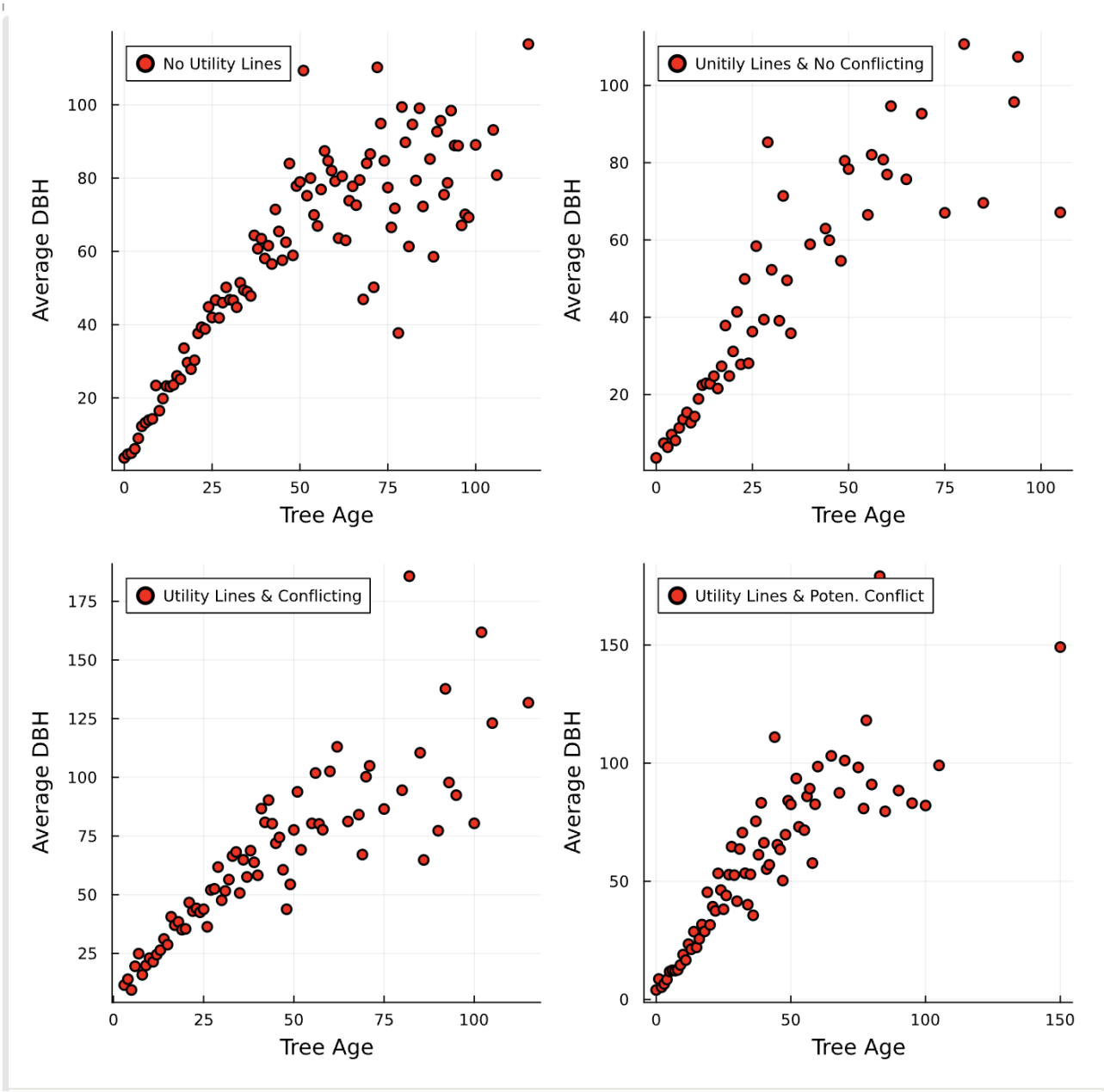


Figure 5: The Correlation between Tree Age and Average Diameter of Trees based on Wire Conflict

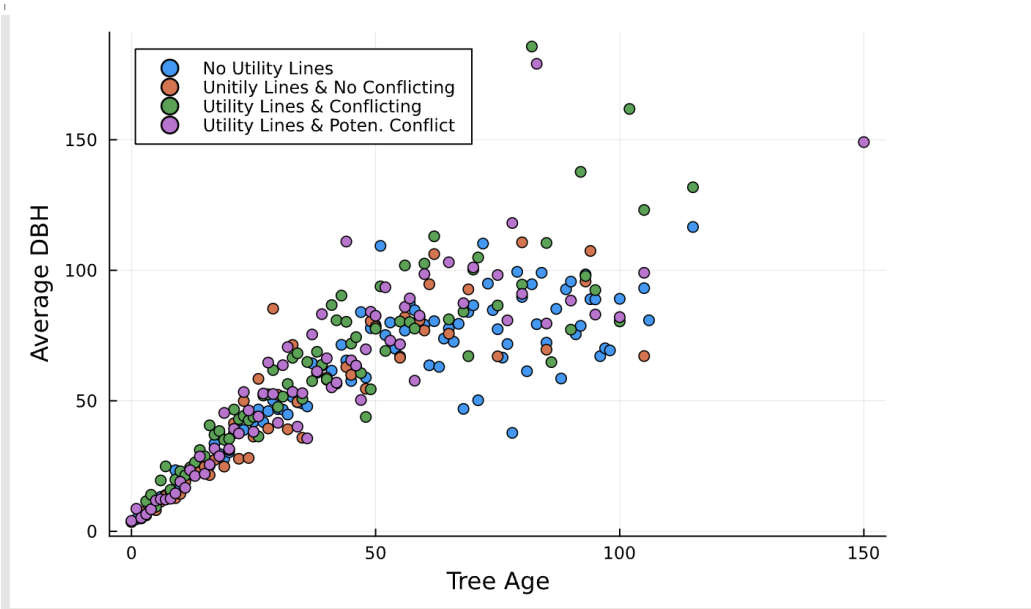


Figure 6: The Correlation between Tree Age and Average Diameter of Trees Based on Wire Conflict

The last step is to find the correlation between the height and diameter of trees to see if the research team can use that in estimating the height of trees based of its diameter. Figure 7 present the correlation between the two aforementioned variables. It is clear that there is a strong correlation between tree height and its diameter. The calculated correlation is 0.78.

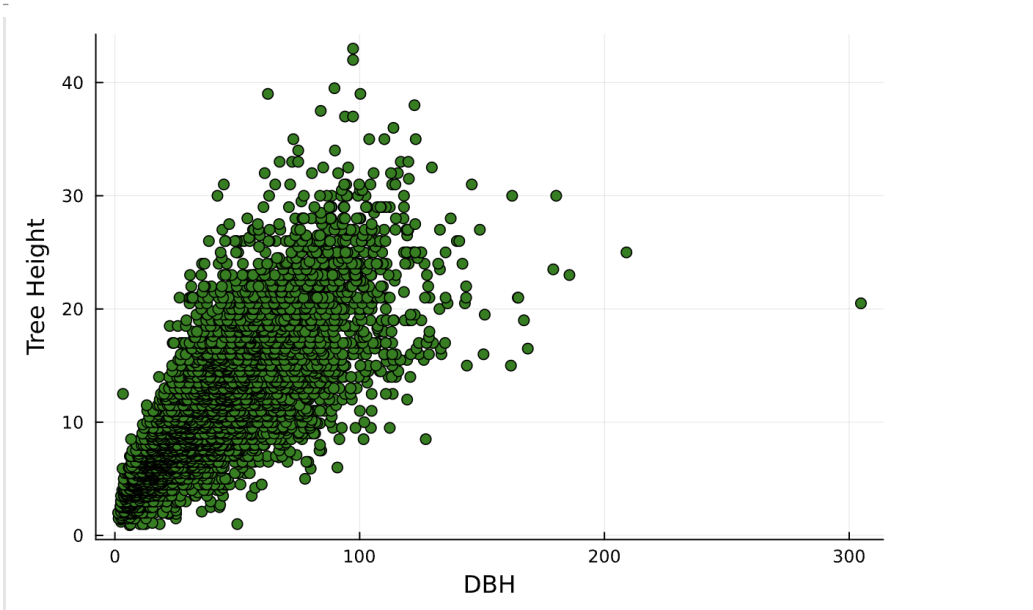
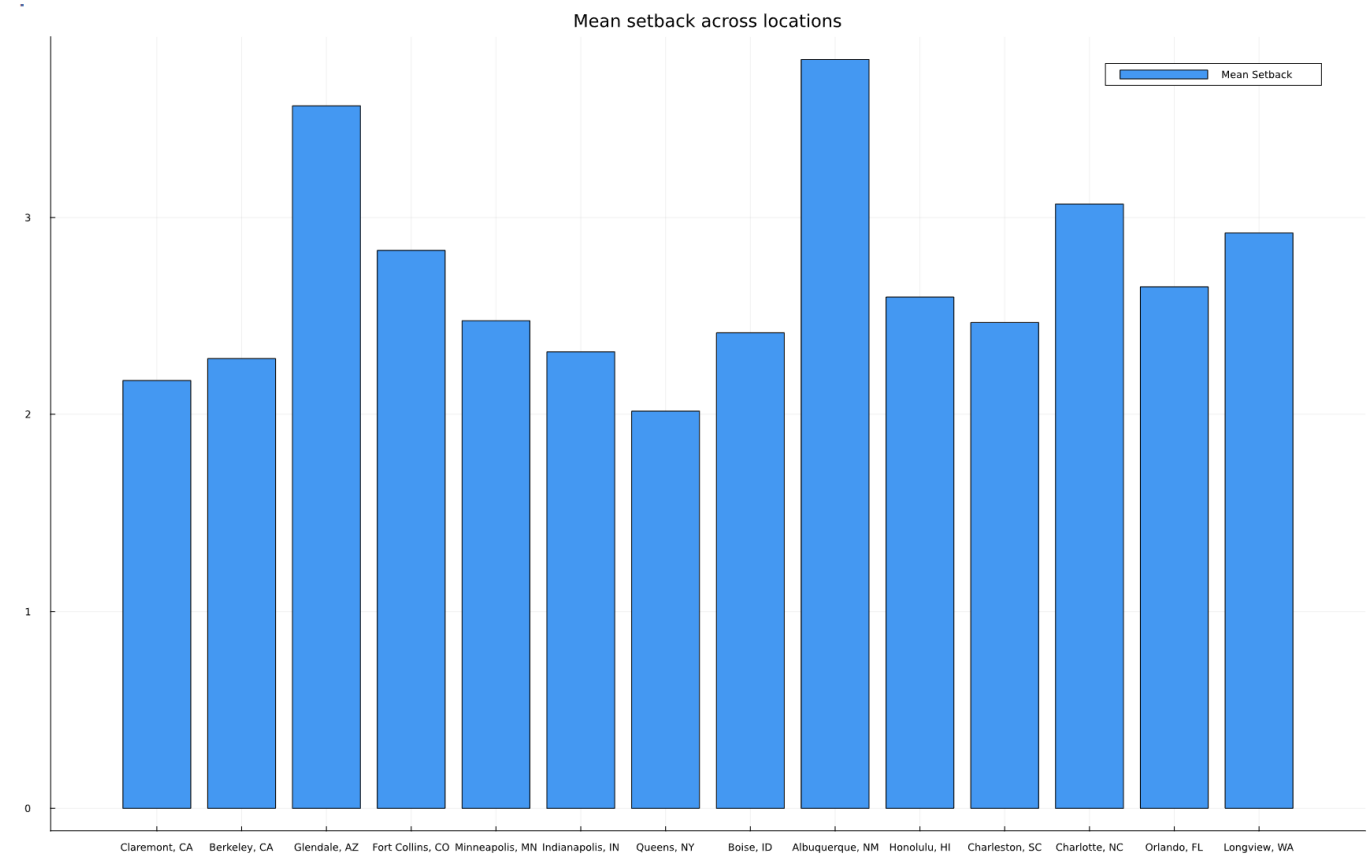


Figure 7: The Correlation between Height and Average Diameter of Trees

Therefore, it is clear that the present of utility line does not have a great impact on the growth of trees. However, the research team has found that there is a strong correlation between tree age and both height and diameter of tree that can be used in developing a regression model that can predict the age of trees based on their height and diameter.

Question 2

One of the promising variables in the dataset was identified to be the 'setback.' Setback is defined as the distance from tree to nearest air-conditioned/heated space (may not be same address as tree location) with its units explained as follows: 1=0m to 8m, 2= 8.1m to 12m, 3= 12.1m to 18m, 4= > 18m. The analysis on the data explored the effect of setback on the height of trees. After filtering out all the missing values from the dataset, a bar graph was plotted for the mean setback across various locations.



The Mean Setback across different Cities

It was identified that the cities with the highest mean setback (in descending order) were: 1) Albuquerque (3.80385), 2) Glendale (3.56843), 3) Charlotte (3.06892), 4) Longview (2.92153) Similarly, the cities with least mean setback were (in ascending order) : 1) Queens (2.01564), 2) Claremont (2.17143), 3) Berkeley (2.28313), 4) Indianapolis (2.31699)

Next step was to identify similar species between the trees from top 4 mean setback and bottom 4 mean setback. This would help establish similar grounds for tree height comparison. However it was found out that there were no common species among the two groups. Hence, a random city (Charlotte) was taken into consideration, where, similar species having the same age were grouped together.

The Mean Height for different Setback

The Mean Height for different Setback

Using their mean heights, it was observed that setback and tree height did not show any correlation.

Question 3

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 to provide sufficient shade to city goers without interscepting overhead telephone lines or buildings. [Site from evidence]. Furthermore, these data were grouped by city and region to investigate spatial differences among the variables. Perhaps southern California cities like Santa Monica plant different trees compared to those in Boise, Idaho for different purposes. The following visualizations were produced to study these qualitative and quantitative relationships.

First, 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 on which trees to include in their city landscape plans depending on whether short or tall trees would best suit their site.

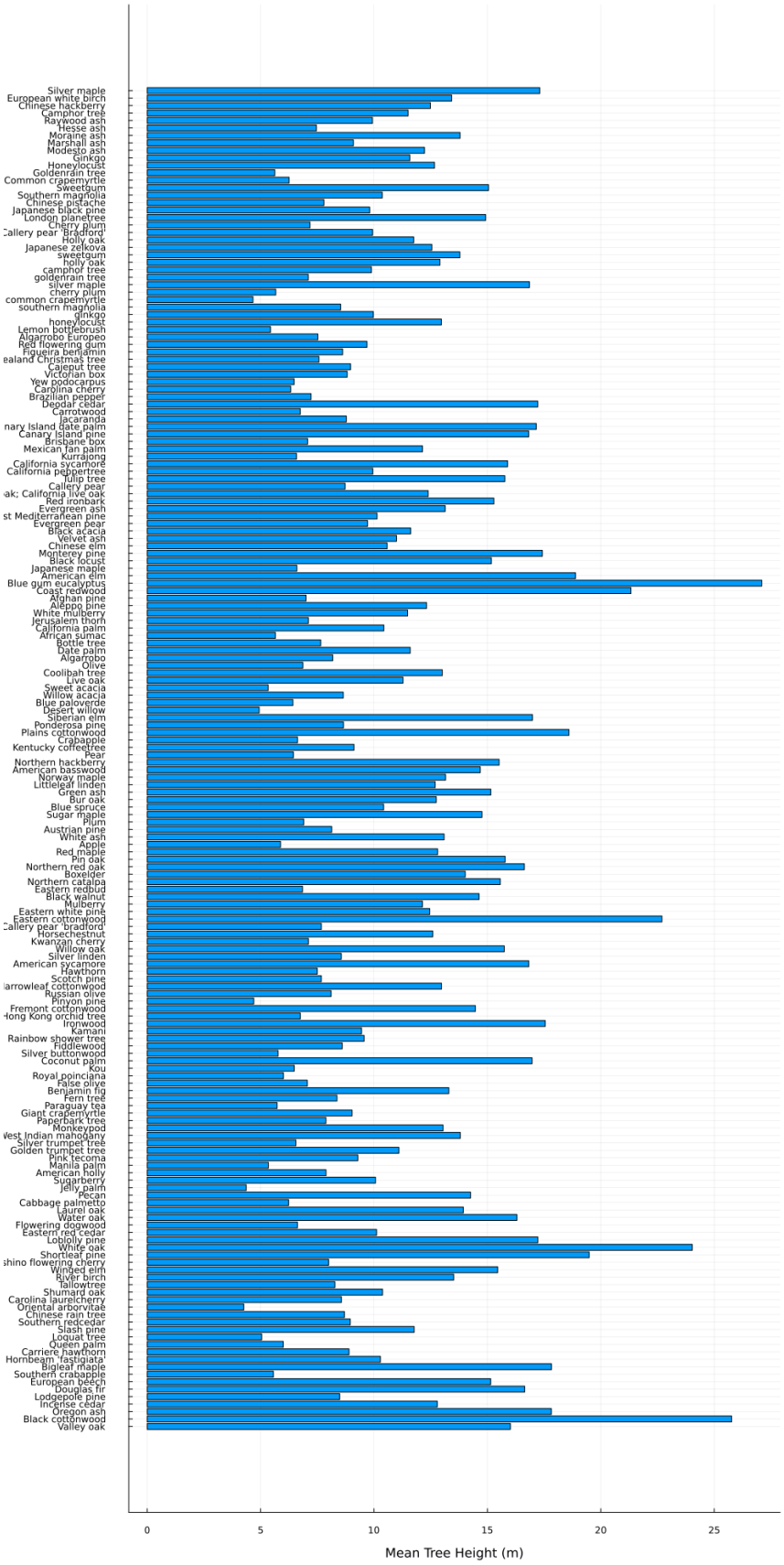


Figure 8: Tree Height by Species.

Secondly, 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.

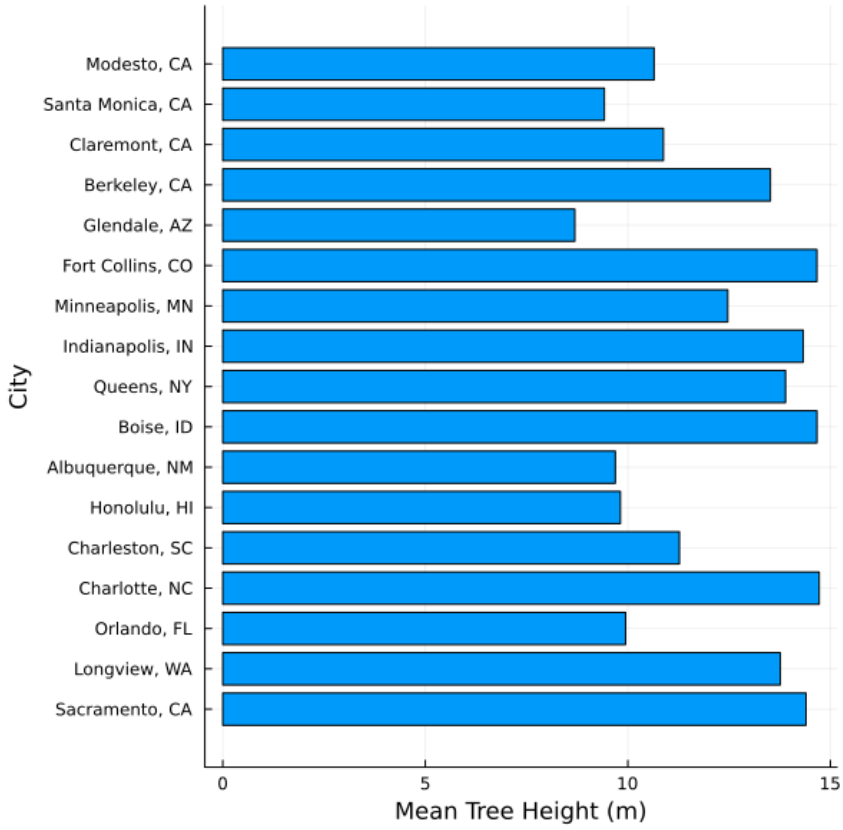


Figure 9: Tree Height by City.

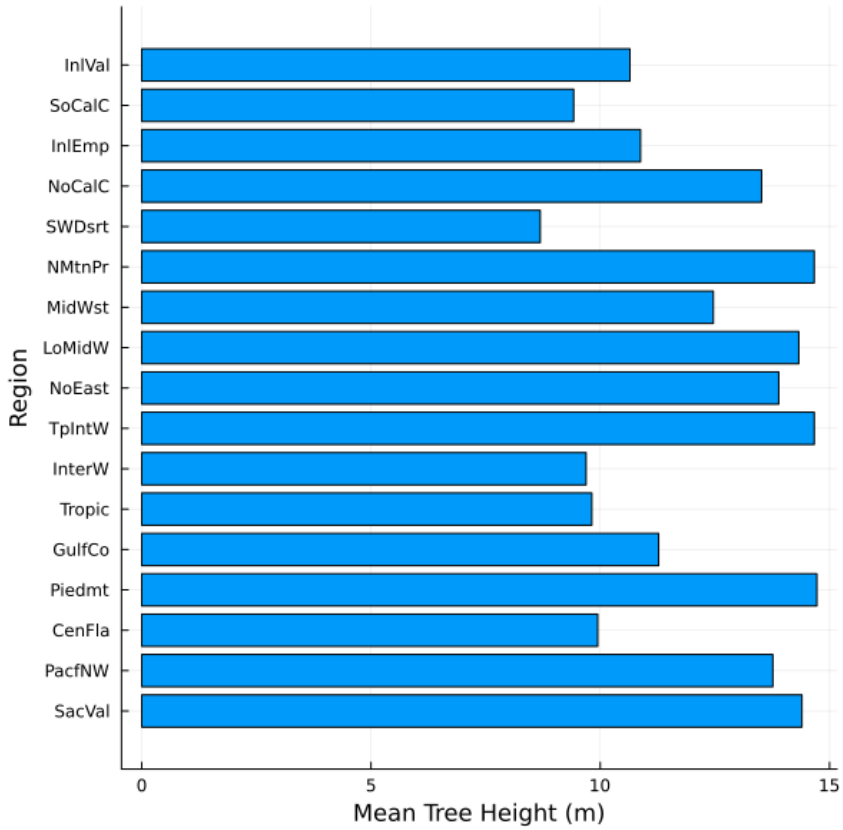


Figure 10: Tree Height by Region.

Thirdly, a barplot 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. 11, 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).

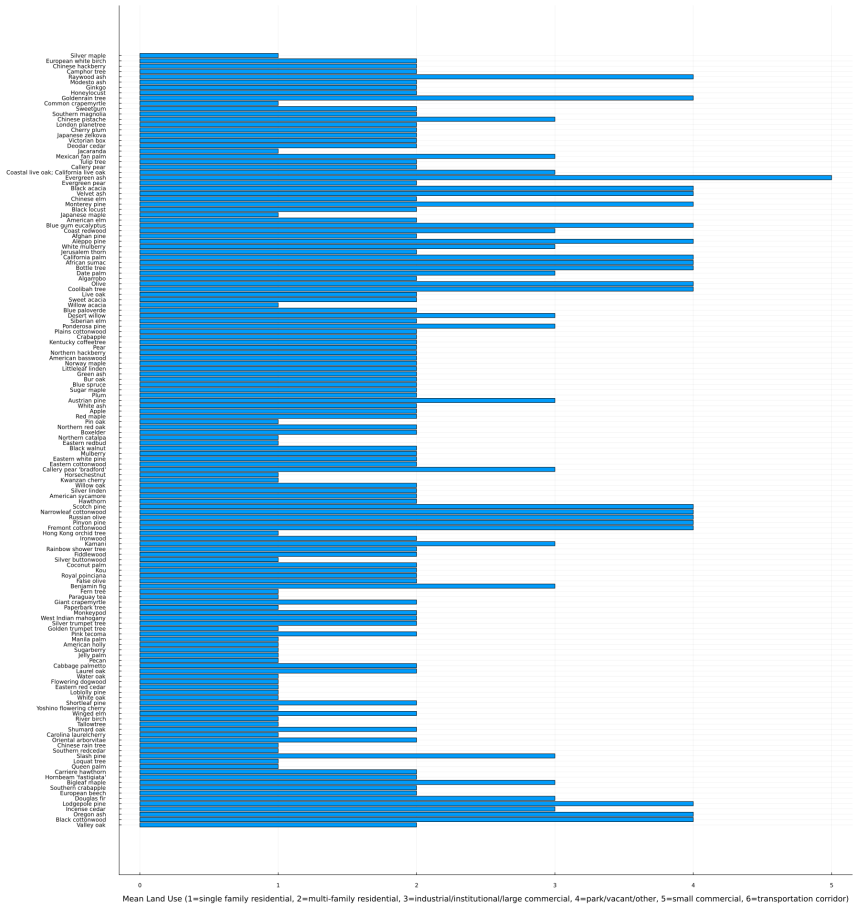


Figure 11: Tree Species by Average Land Use.

Additionally, the correlations among tree height, DBH, crown height, and leaf area were further 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 (https://data.nal.usda.gov/dataset/urban-tree-database) [note: citations will be updated formally!]. Extending the investigation to include these considerations, tree height, DBH, crown height, and leaf area variables were selected and their correlations were calculated. The following graphs depict marginal histograms, which are useful in explaining the distributions of each variable as well as how they are correlated.

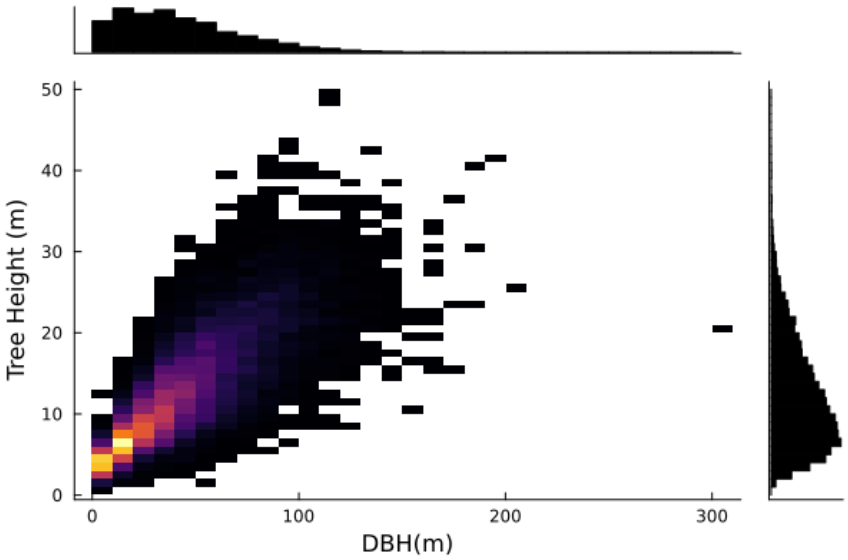


Figure 12: Marginal Histogram of DBH and Tree Height.

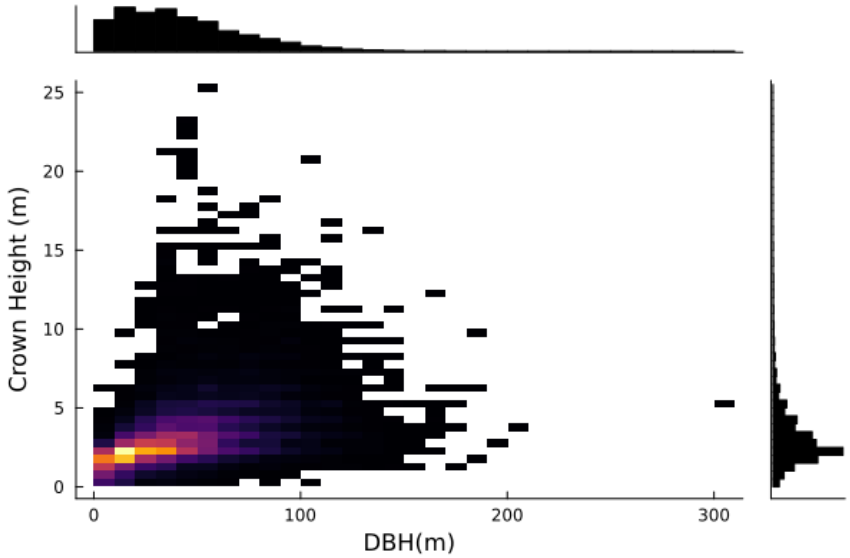


Figure 13: Marginal Histogram of DBH and Crown Height.

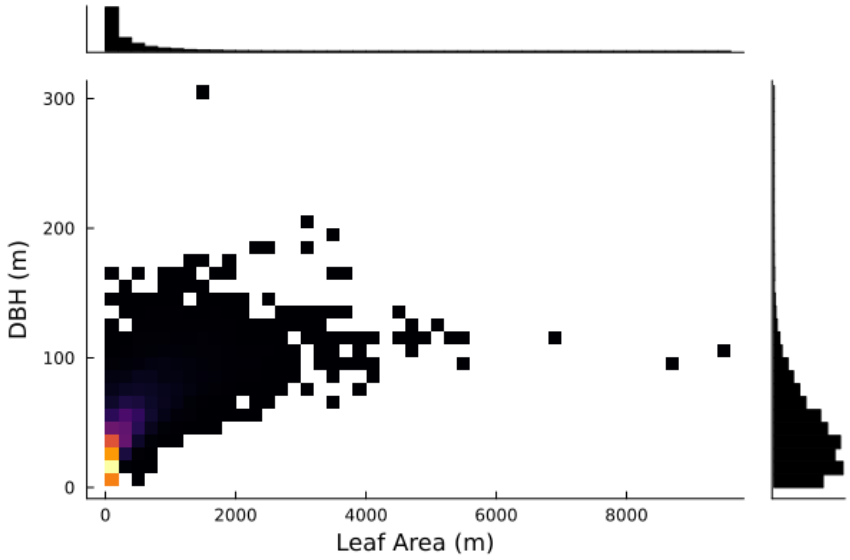


Figure 14: Marginal Histogram of Leaf Area and DBH.

{In depth explanation of above:}

Finally, to investigate these correlations further, average DBH by tree heights were 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.

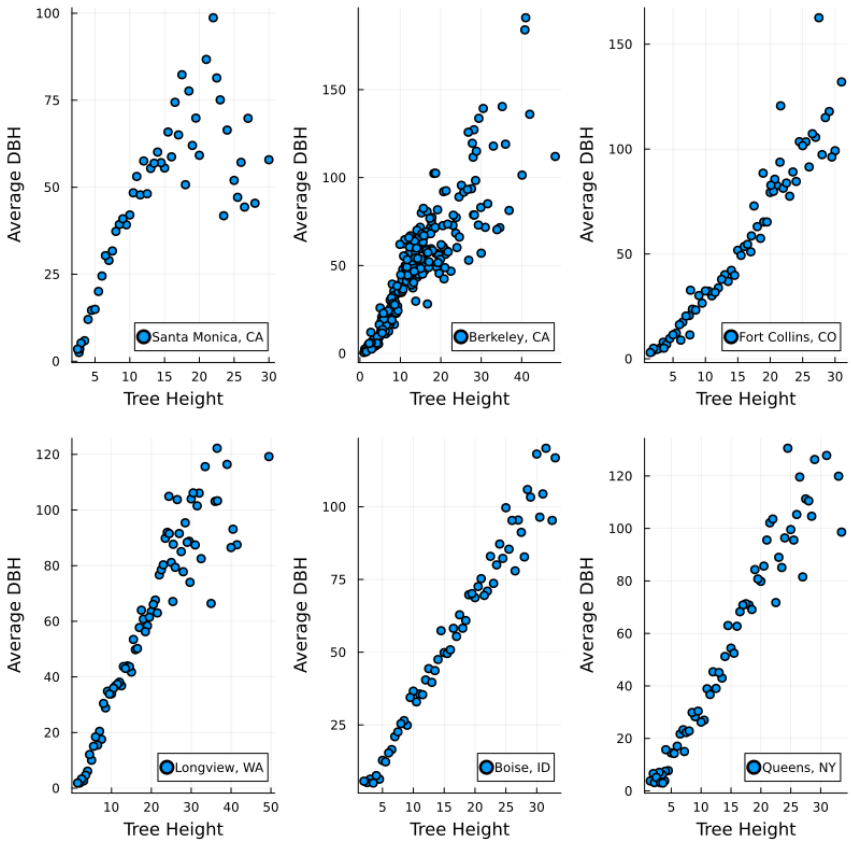


Figure 15: Average DBH vs Tree Height by City.

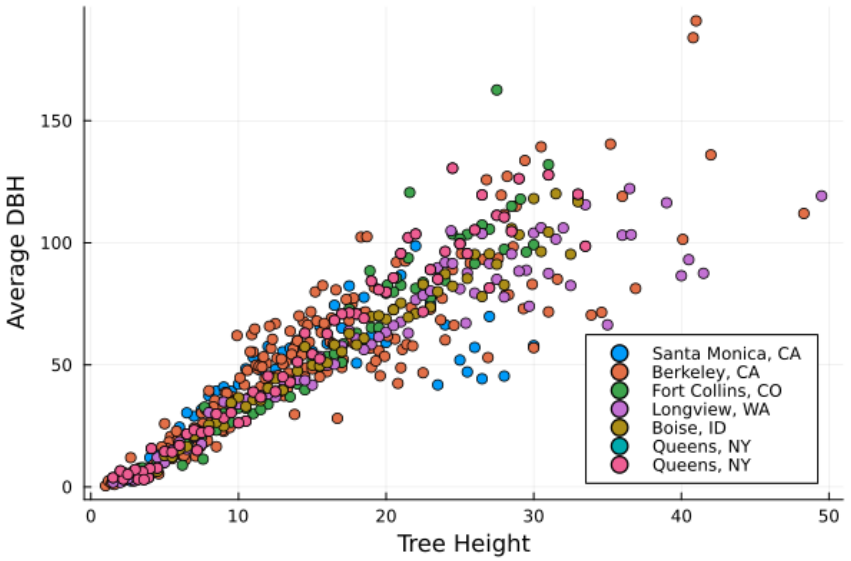


Figure 16: Average DBH vs Tree Height by City - Combined.

Correlation Coefficients - DBH vs Tree Height Overall: 0.8023385455282306 - Santa Monica, CA: 0.7132638836192362 - Berkeley, CA: 0.8886494827638055 - Fort Collins, CO: 0.959599952400562 - Longview, WA: 0.9334828561163339 - Boise, ID: 0.9853443320647175 - Queens, NY: 0.9702645394292799 - Leaf Area vs DBH: 0.7132638836192362 - DBH vs Crown Base Height: 0.42209722299954183

{Wrap up}

Question 4

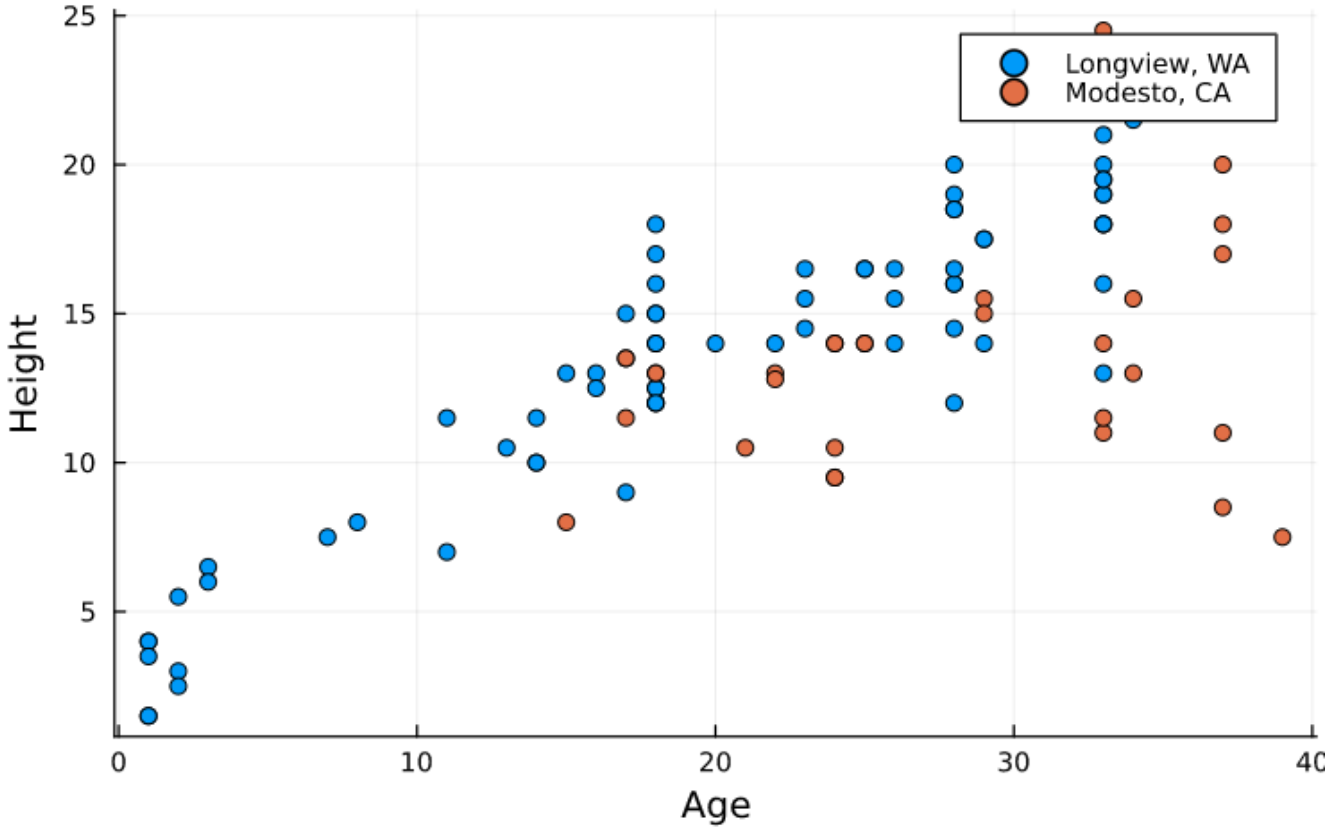
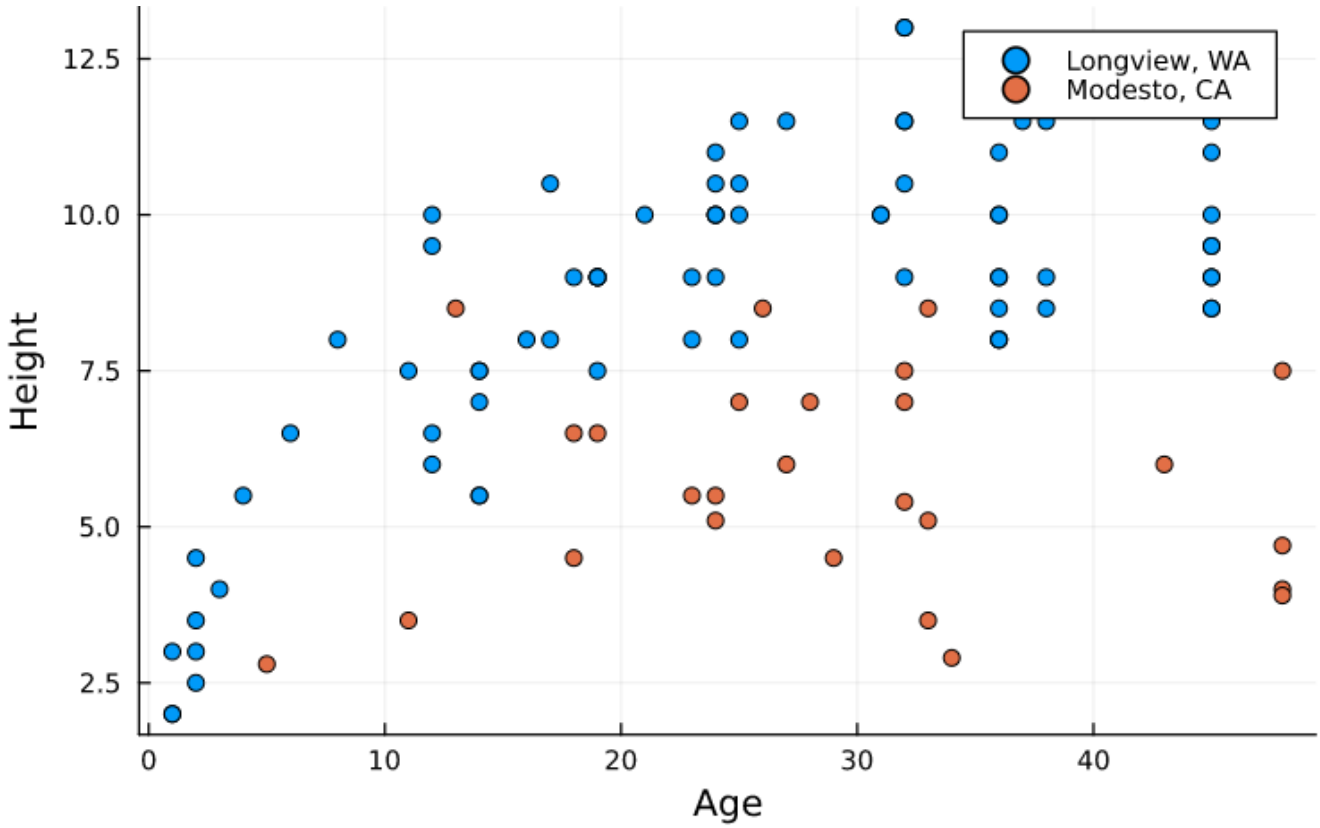


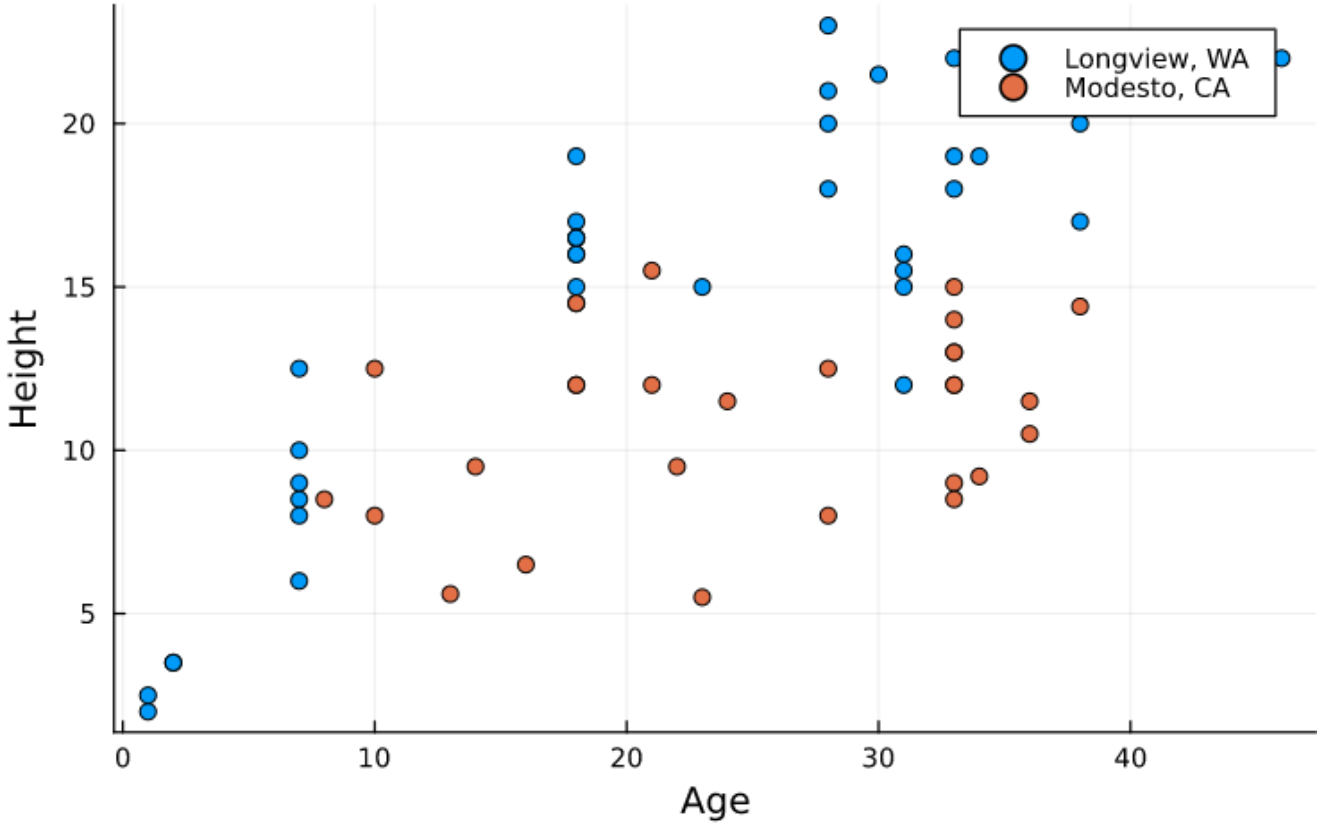
Figure 17: Age versus Height of Sweetgum trees in Longview, WA and Modesto, CA.

This figure shows that Longview, WA trees are taller than Modesto, CA trees at any age.



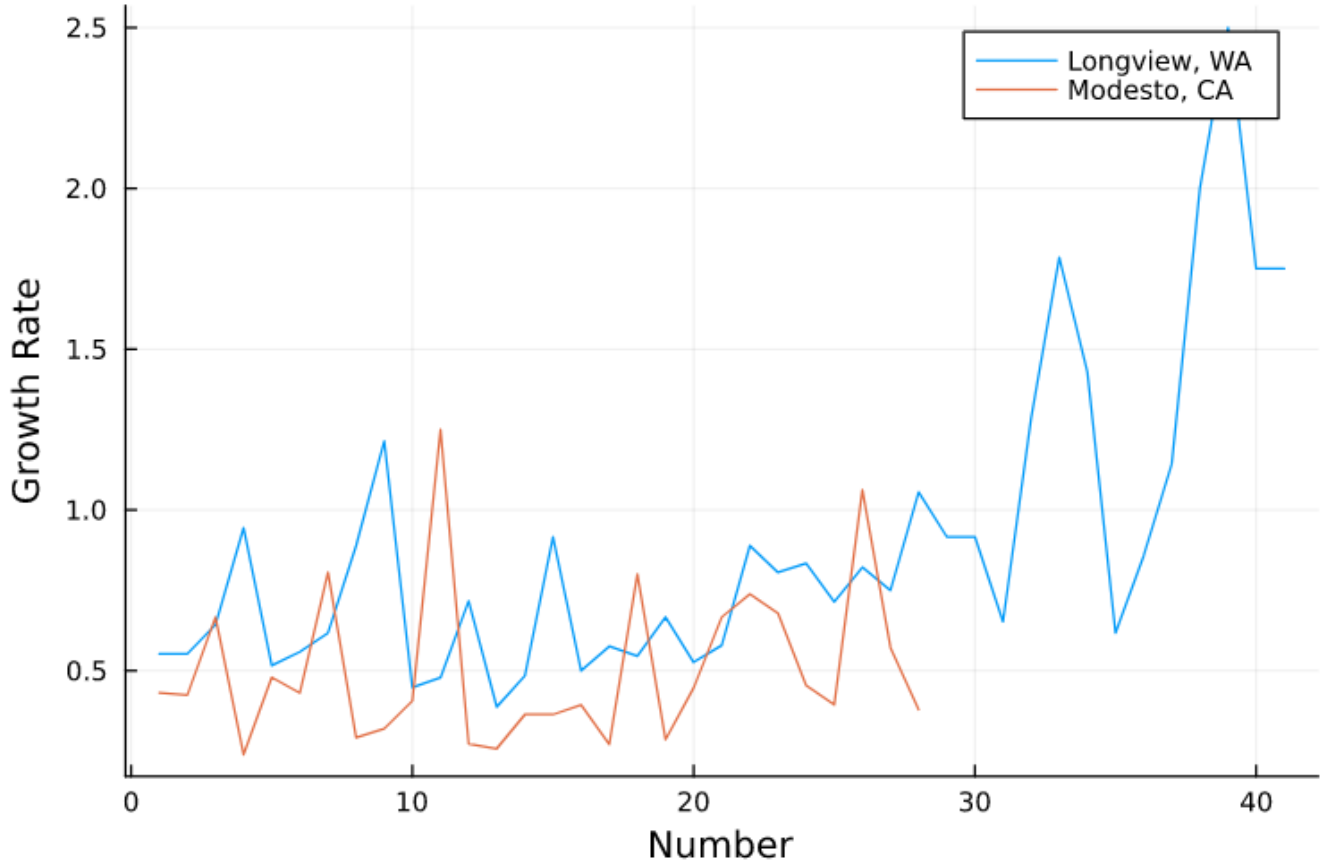
{#fig:Ri_Cherry_hiVSage}

This figure shows that Longview, WA trees are taller than Modesto, CA trees at any age.



{#fig:Ri_Euro_hiVSage}

This figure shows that Longview, WA trees are taller than Modesto, CA trees at any age.



{#fig:Ri_Euro_GrowthRate}

This figure shows that the growth rate is not constant for one tree type, and may vary for the age of the tree or the time at which the tree was planted. Here it can be seen that generally the growth rate in Longview, WA is greater than that of Modesto, CA

These figures show that there is a relationship between location and height of trees. This relationship may be because of temperature, precipitation, or other factors outside of the dataset. Some variables within this dataset that may affect tree height are explored in this section, and include: setback of trees from conditioned spaces, wire interference, and...

References

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