Data Chapter

There are two sources of data in this project. Data at a pixel level come from satellite images downloaded from Planet.com (https://www.planet.com), and data at the ground level come from a library called "pdxTrees" in RStudio (https://github.com/mcconvil/pdxTrees).

Imaging

The images from Planet.com were taken by satellites on September 2nd, 2020 at an altitude of 475km. The compiled multi-spectral image is composed of 4 bands: blue, green, red, and near-infrared, with center wavelengths 490nm, 565nm, 665nm, and 865nm respectively, with an average bandwidth of 41nm. The average spatial extent is around 25km by 8km. The images cover the entire Portland area, with specific measurements of 26km from west to east and 30km from north to south. The spatial resolution of the pixels is approximately 3 meters. Downloaded images are then opened in QGIS where further analysis on the images can be made.

Using Planet.com for gathering satellite images has the added benefit of providing a limited number of downloadable images by making a free account. However, one drawback to using this data is that it has a poor resolution, which will decrease the performance of statistical models trying to predict tree species location.

https://www.planet.com/products/satellite-imagery/files/1610.06_Spec%20Sheet_Combined_Imagery_Product_Letter_ENGv1.pdf https://developers.planet.com/docs/data/sensors/

Ground Data

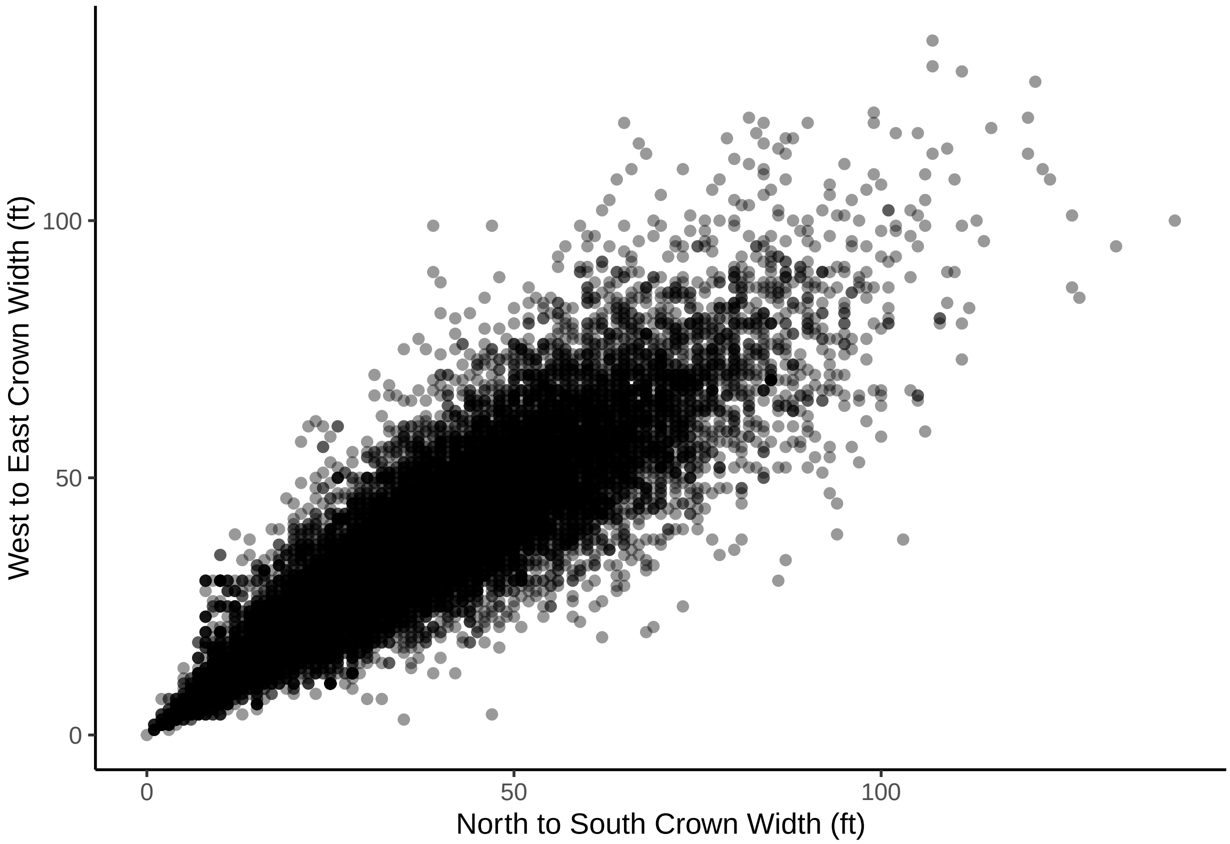
Ground level data is available from the 'pdxTrees' package in RStudio. This data was collected as part of the Portland's Parks and Recreation's Urban Forestry Tree Inventory Project, which collected park tree information from 2017 to 2019. Originally, the inventory project started with the goal of improving the city's tree management plans. Under the guidance of US Forestry staff, volunteers around the city's neighborhoods are trained in identifying tree species and provided with tools necessary to record tree measurements. The Portland park trees inventory consists of over 25,000 trees, each with information about tree location, size, species, and health. For the purposes of this project, the following variables were selected: Spatial information: Longitude and latitude of tree Species: Tree genus, species, and common name Crown size: Crown width from north to south, crown width from east to west, and the base height of the crown in feet Tree Condition: Categorical variable with four categories (from best to worst), good, fair, poor, and dead To prepare the ground level data for training the model, the data was filtered by species and crown size. The following species were identified and included: maples, oak, douglas fir, western red cedar, oregon bigleaf, sequoia. Since the data is used to identify trees in the satellite images, the ground data was also filtered to include only the trees with a crown width from north to south of 20 feet or more, this ensures that there will be around 9 to 20 pixels for each tree.

Portland's Tree Inventory is appropriate for this project, because it provides the variables necessary for identifying Western Redcedars as well as the larger trees in the dataset, which is ideal for constructing a training set of tree polygons in a satellite image. One way this data would be even better suited for the purposes of this project would be if the street trees subdata also contained variables for tree canopy size.

Training Data

Characteristics of Training Data

The training data at the ground-level was filtered by tree size and tree type. Only trees with at least 20 feet in north to south crown width were included in the training dataset, to ensure that the polygons around each tree contain around 9 to 20 pixels. Figure @ref(fig:crownWidth) plots the north to south crown width values against the west to east crown width values to present the ranges of crown widths in the entire pdxTrees parks dataset and to justify only using the north to south crown width for filtering the dataset since the two variables have a strong positive correlation.



Six tree types were determined to have large enough canopies on average and appear frequently enough to construct a training dataset with plenty of observations for each species of tree: Douglas-Fir, English Oak, Giant Sequoia, Maple, Western Redcedar. Figure @ref(fig:pointsTable) provides the number of trees under each common name category. The largest group contains 6485 observations of Douglas-Fir trees, and the smallest group contains 135 observations of English Oak trees.

```
## # A tibble: 6 x 2
##   Common_Name      counts
##   <chr>             <int>
## 1 Douglas-Fir        6485
## 2 Norway Maple       1406
## 3 Western Redcedar    652
## 4 Bigleaf Maple       464
## 5 Giant Sequoia       315
## 6 English Oak         135
```

Creating Spatial Polygons

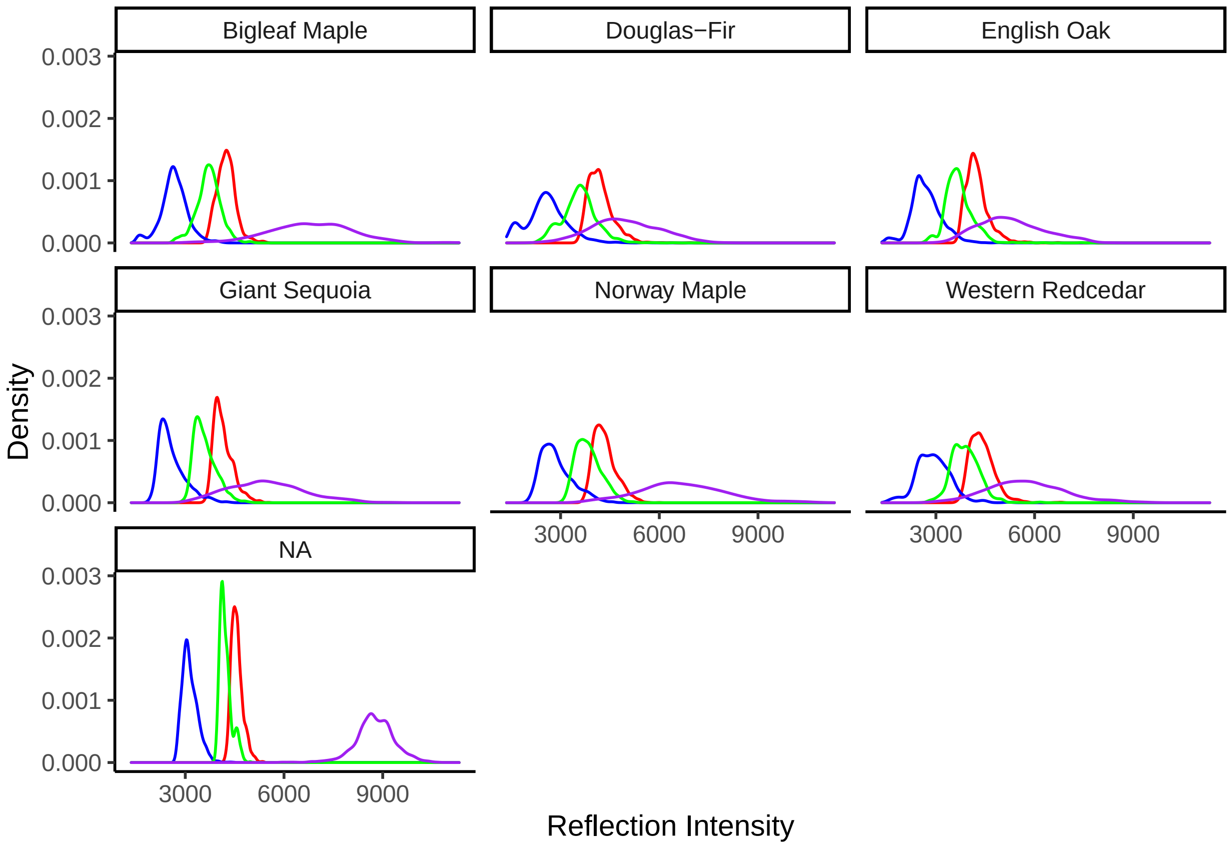
The training dataset was converted into shapefiles for each type of tree and exported into QGIS where polygons were manually drawn around 100 of the trees for each species. Displaying the point shapefiles layer over the raster images downloaded from Planet.com in QGIS provided a guide point for locating individual trees around which a polygon can be created by tracing the tree canopy. Careful attention was paid to avoid drawing polygons around trees with canopies that overlap with other trees of different species along with trees that are cast in shadows or otherwise partially obstructed by surrounded structures. For the five different species of trees, at least 100 polygons were drawn around trees of that type, with each polygon containing at least 6 pixels and at most 20 pixels. The polygon shapefiles were then exported into RStudio where the rest of the analysis was conducted.

Combining Ground-level Data with Pixel-level Data

The first spatial join in RStudio was conducted to match up each Spatial Polygon with a point in the training dataset. Then the raster images were loaded and joined with the polygons to extract the pixel values inside each polygon for all 4 bands. Ultimately, a pixel table with rows representing each pixel with its corresponding polygon, light reflection intensity values for all 4 bands, and the ground information about that tree. Table @ref(fig:pixelCounts) contains the summary of the total number of pixels for each tree type.

```
## # A tibble: 7 x 2
##   Cmmn_Nm          counts
##   <fct>             <int>
## 1 Bigleaf Maple       885
## 2 Douglas-Fir        1380
## 3 English Oak        1025
## 4 Giant Sequoia      1436
## 5 Norway Maple       2340
## 6 Western Redcedar   1364
## 7 <NA>               2923
```

Figure @ref(fig:pixDensity) displays the range of reflection intensity pixel values per tree type. Each tree species has similar ranges of values over the red, green, blue, and infared bands. Of the species included in the training data, the Giant Sequoia trees appear to have higher density counts for the red, green, and blue bands.



Bigleaf Maple
Douglas−Fir
English Oak
Giant Sequoia
Norway Maple
Western Redcedar
NA
0.003
0.002
0.001
0.000
Density
3000
6000
9000
Reflection Intensity