Thesis

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Sarah Maebius

December 6th, 2020

Approved for the Division
(Mathematics)

_____

Tom Allen

Acknowledgements

Deeply grateful to...

Table of Contents

Introduction **1**
0.1 Research Problem and Background 2
0.2 Overview 3
0.2.1 A Statistical Learning Approach to Identifying Location of Western Redcedars 3
0.3 Sources 3
0.4 Data Chapter 4
0.4.1 Imaging 4
0.4.2 Ground Data 4
0.4.3 Training Data 5

Chapter 1: Mathematics and Science **9**
1.1 Math 9
1.2 Chemistry 101: Symbols 9
1.2.1 Typesetting reactions 10
1.2.2 Other examples of reactions 10
1.3 Physics 10
1.4 Biology 10

Conclusion **11**

Appendix A: The First Appendix **13**

Appendix B: The Second Appendix, for Fun **17**

References **19**

List of Tables

1 Common tree names included in the tidy data and their total counts. 6
2 Variables and pixel values included in the pixels dataset 7
3 Number of pixels per tree type 7

List of Figures

1 Plot comparing North to South Crown Width (ft) to West to East Crown Width (ft). A tree with a higher north-south crown width has a higher west-east crown width, so without loss of generality the north-south crown width variable is used to filter the dataset for larger trees only. 6
2 8

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Introduction

0.1 Research Problem and Background

Western redcedar trees are evergreen trees that typically grow up to 75 feet tall and are located over the Pacific Northwest, making it an organism with a tolerance for shaded regions with moist environments. These trees are native to the land and have served many purposes to people and animals living in the vicinity of the trees, including medicinal, building, and habitat functions (Peterson, n.d.). However, over the past decade, reports of dead western redcedars have been increasing, suggesting that something other than natural causes is killing off this species. In general, this species experiences several hardships in surviving in the Pacific Northwest, with common causes of death such as forest fires, clearcutting, small animals eating the saplings, and harsh weather including strong winds that easily uproot the trees (Peterson, n.d.). Dying redcedar trees can be identified by their branches which turn brownish yellow or fall off completely. Another sign is that the top of a dying redcedar will turn brown and lose leaves ("Western redcedar Dieback," n.d.). Losing this native tree would have a detrimental effect on animals in the area who rely heavily on the trees for their lifestyle. Scientists have speculated that western redcedar decline might be caused by recent dry summers, the spread of tree disease, insects, or other weather related events ("Western redcedar Dieback," n.d.). Since this is a recent issue, there is not a lot of resources explaining the decline. This research aims to provide more insight into the cause of the western redcedar decline by first predicting the location of the western redcedar trees in Portland and then predicting their condition in terms of health. Having more insight helps to prevent further tree deaths and save the western redcedar species, which also extends to similar tree species and provides more knowledge about environmental changes in the Pacific Northwest region.

Modelling in this research will be conducted by combining information gathered from remote sensing images with ground level information. There are several sources publishing research done using satellite imagery for land classification and for predicting tree species, which will be the groundwork for this project's application of remote sensing models to the specific topic of western redcedar mortality.

Plenty of literature has been published in utilizing satellite imagery for predictive models, however, many issues arise in classifying land type through remote sensing mainly due to image quality. A single image can be composed of image strips taken over the course of multiple flight paths. Consequently, these images are taken at different times of day, which compiles into a single image with a lot of variation in values due to changes in the weather as well as the different angles of the sun's position (Castelluccio, Poggi, Sansone, & Verdoliva, 2015). Ideally, a solid classifying model surpasses any error introduced by imperfection in the satellite images. Common approaches to classification on satellite images include support vector machines, random forests, and decision trees. One study on land type classification explores the performance of convolutional neural networks (CNN) as classification models (Castelluccio et al., 2015).

Another related study identifies 7 tree species by applying a hyperspectral CNN model. This study was completed using field data with measured information about tree count, tree species, and mortality status and hyperspectral imagery data. The

hyperspectral data is converted into the form of a tree canopy height model with circular polygons centered at individual tree canopies. The study analyzes the performance of CNNs for both hyperspectral imagery data and a Red-Green-Blue (RGB) subset of the hyperspectral imagery data. The results of the experiment conclude that training a CNN on the hyperspectral data outperforms the CNN trained on RGB data in classifying land type on the UC Merced Land Use dataset. The RGB model does not perform as well as the hyperspectral model in terms of distinguishing tree classes, but does perform around the same level of accuracy in terms of genus classification. For this project, the data comes from Planet.com and only has RGB data available for the Portland region.

This thesis follows along with the methods in (Fricker et al., 2019) for combining the satellite imagery data with ground data by creating spatial polygons and extracting pixel-level information to train classification models. The work in this thesis differs from the methods in the cited literature by using random forests and support vector machines as classification models instead of CNNs and uses the RGB model instead of hyperspectral model. Finally, this thesis applies established classification methods for satellite imagery data to answer the specific question about the cause of death of western redcedars in the Pacific Northwest. First models predicted the locations for each tree species, then the health conditions of the western redcedars are modelled to identify any patterns in the species mortality over the past decade.

0.2 Overview

0.2.1 A Statistical Learning Approach to Identifying Location of Western Redcedars

This work combines RGB imagery data from Planet.com with ground level tree data from the RStudio `pdxTrees` library and applies random forest and support vector machine classification methods to predict the location of tree species (specifically western redcedars) in Portland, Oregon and model the condition of western redcedars to ultimately understand the cause of death in this species. [insert key findings here]

0.3 Sources

[1] `https://plants.usda.gov/plantguide/pdf/cs_thpl.pdf`

[2] `https://ppo.puyallup.wsu.edu/plant-health-concerns/redcedar/`

[3] `https://www.treespnw.com/resources/2018/11/7/are-the-western-redcedars-dying`

[4] `https://arxiv.org/pdf/1508.00092.pdf` (Land Use Classification in Remote Sensing Images by Convolutional Neural Networks)

[5] `https://www.fs.fed.us/psw/publications/north/psw_2019_north009_fricker.pdf` (A Convolutional Neural Network Classifier Identifies Tree Species in Mixed-Confier Forest from Hyperspectral Imagery)

0.4 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 library `pdxTrees` in RStudio (`https://github.com/mcconvil/pdxTrees`).

0.4.1 Imaging

The images from Planet.com were taken by satellites on September 2nd, 2020 at an altitude of 475 km. The compiled multi-spectral image is composed of 4 bands: blue, green, red, and near-infrared, with center wavelengths 490 nm, 565 nm, 665 nm, and 865 nm respectively, with an average bandwidth of 41 nm. The average spatial extent is around 25 km by 8 km. The images cover the entire Portland area, with specific measurements of 26 km from west to east and 30 km 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 its low 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/`

0.4.2 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 tree species were identified and included:

maples, oak, Douglas-fir, Western Redcedar, and sequoia. Since the data is used to identify trees in the satellite images, the ground data was also filtered to only include 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 per tree.

Portland's Tree Inventory provides valuable data for this project because it includes a variable for identifying Western Redcedars as well as variables for identifying larger trees in the dataset, which helps construct 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.

0.4.3 Training Data

Characteristics of Training Data

The training data at the ground-level was filtered by tree size and tree type. 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 1 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.

Table 1: Common tree names included in the tidy data and their total counts.

Tree Name	Tree Count
Douglas-Fir	6485
Norway Maple	1406
Western Redcedar	652
Bigleaf Maple	464
Giant Sequoia	315
English Oak	135

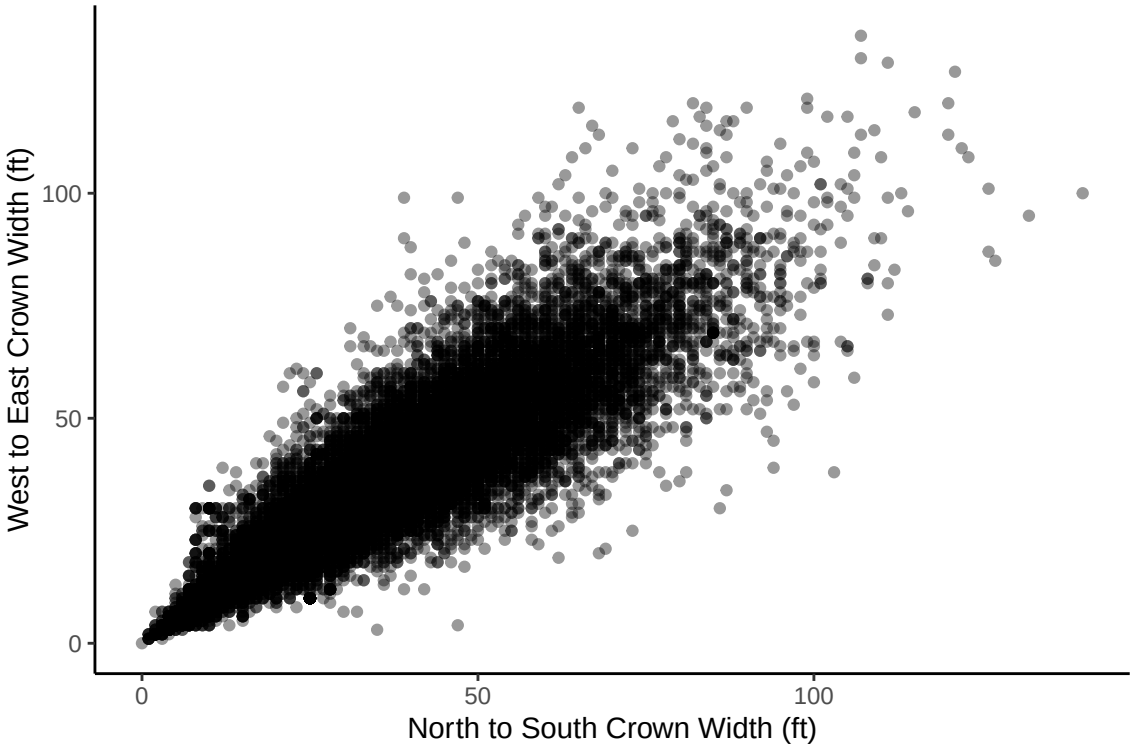

Figure 1: Plot comparing North to South Crown Width (ft) to West to East Crown Width (ft). A tree with a higher north-south crown width has a higher west-east crown width, so without loss of generality the north-south crown width variable is used to filter the dataset for larger trees only.

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. Table 1 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.

Table 2: Variables and pixel values included in the pixels dataset

ID	red	green	blue	ir	id_type	Genus	Species	Cmmn_Nm	Cr_W_NS	Cr_W_EW	Crw_B_H	Conditn	St_Wdth	ndvi
43	4707	4125	3187	6627	acpl	Acer	ACPL	Norway Maple	28	28	8	Fair	NA	0.1694018
43	4384	3860	2864	7111	acpl	Acer	ACPL	Norway Maple	28	28	8	Fair	NA	0.2372336
43	4446	3803	2807	7443	acpl	Acer	ACPL	Norway Maple	28	28	8	Fair	NA	0.2520818
43	4619	3911	2859	7801	acpl	Acer	ACPL	Norway Maple	28	28	8	Fair	NA	0.2561997
43	4695	4060	3201	5996	acpl	Acer	ACPL	Norway Maple	28	28	8	Fair	NA	0.1216911
43	4344	3719	2728	6747	acpl	Acer	ACPL	Norway Maple	28	28	8	Fair	NA	0.2166622

Table 3: Number of pixels per tree type

Tree Name	Pixels Count
Bigleaf Maple	885
Douglas-Fir	1380
English Oak	1025
Giant Sequoia	1436
grass	2923
Norway Maple	2340
Western Redcedar	1364

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. The point shapefile layers were displayed over the raster images downloaded from Planet.com in QGIS, which provided a guide point for locating individual trees. Then using QGIS's drawing tool, a polygon is carefully drawn around the tree canopy. Most polygons turned out to be four to six-sided polygons to retain the general shape of the tree canopy. Trees polygons were only created if the outline of the tree canopy was clearly visible or surrounded by other trees of the same species in order to avoid including pixels from the wrong species in that polygon. Also, the shadows of the trees are visible in the raster images, so the polygons were drawn with the intention of not including the tree shadow. 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 (red, green, blue, infrared). Ultimately this turned into 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 2 displays the first few entries of the pixels dataset. Table 3 contains the summary of the total number of pixels for each tree type. Figure 2 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.

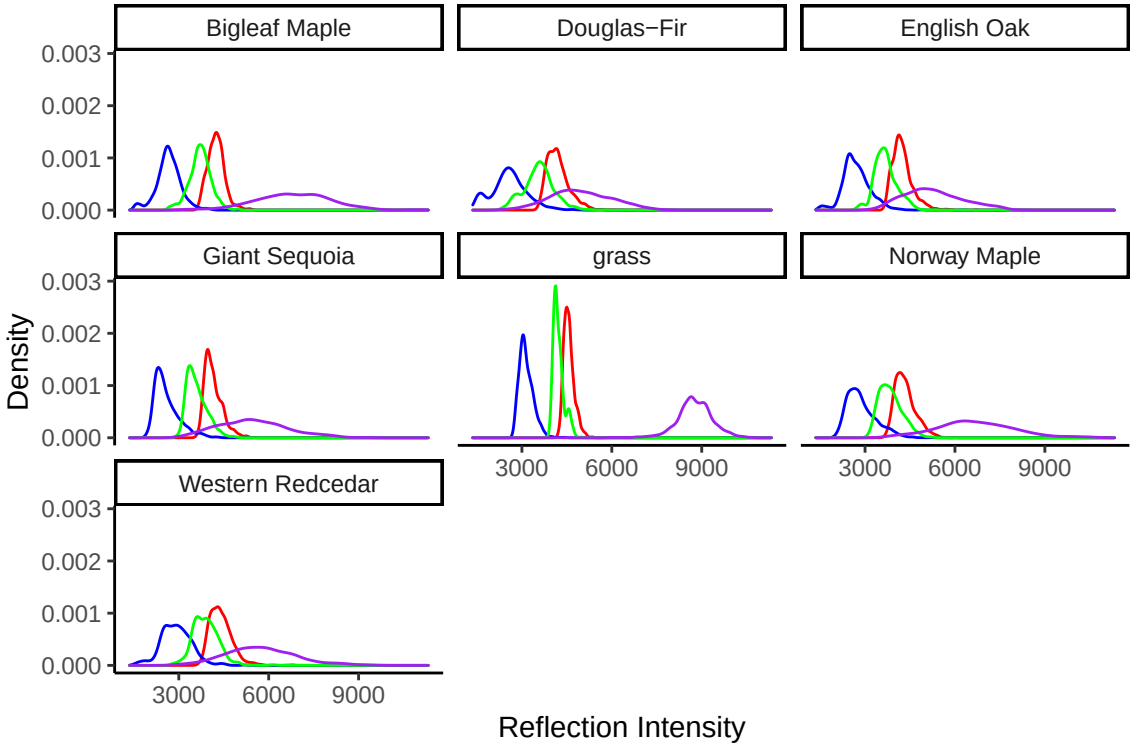

Figure 2

Chapter 1

Mathematics and Science

1.1 Math

TEX is the best way to typeset mathematics. Donald Knuth designed TEX when he got frustrated at how long it was taking the typesetters to finish his book, which contained a lot of mathematics. One nice feature of *R Markdown* is its ability to read LaTeX code directly.

If you are doing a thesis that will involve lots of math, you will want to read the following section which has been commented out. If you're not going to use math, skip over or delete this next commented section.

1.2 Chemistry 101: Symbols

Chemical formulas will look best if they are not italicized. Get around math mode's automatic italicizing in LaTeX by using the argument `$\mathrm{formula here}$`, with your formula inside the curly brackets. (Notice the use of the backticks here which enclose text that acts as code.)

So, $Fe_2^{2+}Cr_2O_4$ is written `$\mathrm{Fe_2^{2+}Cr_2O_4}$`.
Exponent or Superscript: O^-
Subscript: CH_4

To stack numbers or letters as in Fe_2^{2+}, the subscript is defined first, and then the superscript is defined.
Bullet: $CuCl \bullet 7H_2O$
Delta: Δ
Reaction Arrows: $\longrightarrow$ or $\xrightarrow{solution}$
Resonance Arrows: $\leftrightarrow$
Reversible Reaction Arrows: $\rightleftharpoons$

1.2.1 Typesetting reactions

You may wish to put your reaction in an equation environment, which means that LaTeX will place the reaction where it fits and will number the equations for you.

$$\mathrm{C_6H_{12}O_6 + 6O_2 \longrightarrow 6CO_2 + 6H_2O} \tag{1.1}$$

We can reference this combustion of glucose reaction via Equation (1.1).

1.2.2 Other examples of reactions

$\mathrm{NH_4Cl_{(s)} \rightleftharpoons NH_{3(g)} + HCl_{(g)}}$
$\mathrm{MeCH_2Br + Mg \xrightarrow[below]{above} MeCH_2 \bullet Mg \bullet Br}$

1.3 Physics

Many of the symbols you will need can be found on the math page `https://web.reed.edu/cis/help/latex/math.html` and the Comprehensive LaTeX Symbol Guide (`https://mirror.utexas.edu/ctan/info/symbols/comprehensive/symbols-letter.pdf`).

1.4 Biology

You will probably find the resources at `https://www.lecb.ncifcrf.gov/~toms/latex.html` helpful, particularly the links to bsts for various journals. You may also be interested in TeXShade for nucleotide typesetting (`https://homepages.uni-tuebingen.de/beitz/txe.html`). Be sure to read the proceeding chapter on graphics and tables.

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(thesisdown))
  devtools::install_github("ismayc/thesisdown")
library(thesisdown)
```

In Chapter ??:

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if (!require(remotes)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("remotes", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste(
        'You need to run install.packages("remotes")',
        "first in the Console."
      )
    )
  }
}
```

```
if (!require(dplyr)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("dplyr", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste(
        'You need to run install.packages("dplyr")',
        "first in the Console."
      )
    )
  }
}
if (!require(ggplot2)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("ggplot2", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste(
        'You need to run install.packages("ggplot2")',
        "first in the Console."
      )
    )
  }
}
if (!require(bookdown)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("bookdown", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste(
        'You need to run install.packages("bookdown")',
        "first in the Console."
      )
    )
  }
}
if (!require(thesisdown)) {
  if (params$`Install needed packages for {thesisdown}`) {
    remotes::install_github("ismayc/thesisdown")
  } else {
    stop(
      paste(
        "You need to run",
        'remotes::install_github("ismayc/thesisdown")',
```

```
        "first in the Console."
      )
    )
  }
}
library(thesisdown)
library(dplyr)
library(ggplot2)
library(knitr)
flights <- read.csv("data/flights.csv", stringsAsFactors = FALSE)
```

Appendix B

The Second Appendix, for Fun

References

Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *arXiv:1508.00092 [Cs]*. Retrieved from http://arxiv.org/abs/1508.00092

Fricker, G. A., Ventura, J. D., Wolf, J. A., North, M. P., Davis, F. W., & Franklin, J. (2019). A Convolutional Neural Network Classifier Identifies Tree Species in Mixed-Conifer Forest from Hyperspectral Imagery. *Remote Sensing*, *11*(19), 2326. http://doi.org/10.3390/rs11192326

Peterson, J. S. (n.d.). WESTERN RED CEDAR, 3.

Western redcedar Dieback. (n.d.). *PPO Home*. Retrieved from https://ppo.puyallup.wsu.edu/plant-health-concerns/redcedar/