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

List of Tables

List of Figures

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.). Western redcedars were culturally important to indigenous peoples who valued the strong wood and utilized it for construction and everyday necessities ("Western red cedar," n.d.). 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.

Remote sensing is a process for identifying the physical aspects of a large area of land by collecting measured light reflection using airplanes or satellites. This method provides access to lot of information about the region of interest that would otherwise be limited from a human's ground-level persepective. Remote sensing is used to detect characteristics on land as well as the sea. Some remote sensing land applications include tracking forest fires, volcanic eruptions, city growth, and also forest changes.

Satellites orbit the earth, carrying sensors that record levels of electromagnetic radiation detected by the reflection of the sun on the earth. The data collected is the measured radiation from different waves of the spectrum (visible light, infrared, ultraviolet, etc.). Depending on the surface, sunlight get absorbed or reflected back at the orbiting satellite. The remotely sensed data is available at certain resolutions depending on the instruments used. Spatial resolution is the size of a pixel from a

satellite image, and a detailed image has smaller spatial resolution so smaller pixels. Spectral resolution is a measure of the size of the wavelengths, where smaller bands and smaller wavelengths improve the spectral resolution of an image and hench improve the level of detail. Satellite images need to be corrected to combine the bands into an image depending on the analysis.

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]

Chapter 1

Data

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`).

1.0.1 Imaging

Satellite images come from Planet.com's PS2.SD instrument found on Dove-R satellites that were launched in 2017. The PS2 telescope is equipped with a 2D frame detector with 6600 pixels across by 4400 pixels down and a spacing of 1100 pixels. For this instrument, one pixel = 5.5um. To separate the light into red, blue, green, and near-infrared (NIR) channels, the telescope has a high-performance butcher-block filter made of 4 individual pass-band filters. Planet.com equates this pass-band filter with that of Sentinel-2.

As it orbits the Earth, the satellite captures continuous strips of single frame images that are split into a RGB frame and NIR frame. The butcher-block pass separately photographs the red, blue, green, and NIR bands and then combines all four to form an image. Images are are downloaded already fully processed and ready to be analyzed. The process includes corrections for radiometric calibration, terrain distortions corrections, elevation corrections, and atmospheric corrections.

The images used in this research were taken on September 2nd, 2020 at an altitude of 475 km. Images are taken from the summer months to reduce the chance of rain clouds and capture peak greenness of the trees. The compiled multi-spectral image is composed of 4 bands: blue, green, red, and NIR, 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.

Planet.com provides a way to download free images that are already processed and corrected for analysis. Some drawbacks to using this data are the limited number of downloads available per free account, the low spatial resolution, which decreases the performance of statistical models trying to predict tree species location, and the low

spectral resolution, so the satellite sensor cannot detect wavelengths in detail.

1.0.2 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. Measurements in feet were made using diameter tape. 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. Location of tree is recorded on a tablet with Collector for ArcGIS and recorders manually select the tree from a satellite image on the screen. The tree's location is mapped over the satellite images in QGIS for analysis.

- Species: Tree genus, species, and common name. The tree species is used to train the model and predict the tree species of pixels in the test data.

- Crown size: Crown width from north to south, crown width from east to west, and the base height of the crown in feet. The crown of the tree is the tree's above ground leaves, so the crown width is the longest horizontal distance that can be measured between the leaves of the tree. A measuring wheel is used to measure this distance.

- Tree Condition: Categorical variable with four categories (from best to worst), good, fair, poor, and dead. A tree was considered good if the tree is strong and has no apparent issues, fair if the tree is average condition with possibly a few dead branches, poor if the tree has major wounds and dead major canopy loss, and dead if the tree has no live leaves.

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 includes a variable for identifying trees including Western Redcedars as well as variables for filtering to include larger trees canopies (at least 6m in diameter) in the dataset, which helps construct a training set of tree polygons over a satellite image. 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.

1.0.3 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 1.1 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.

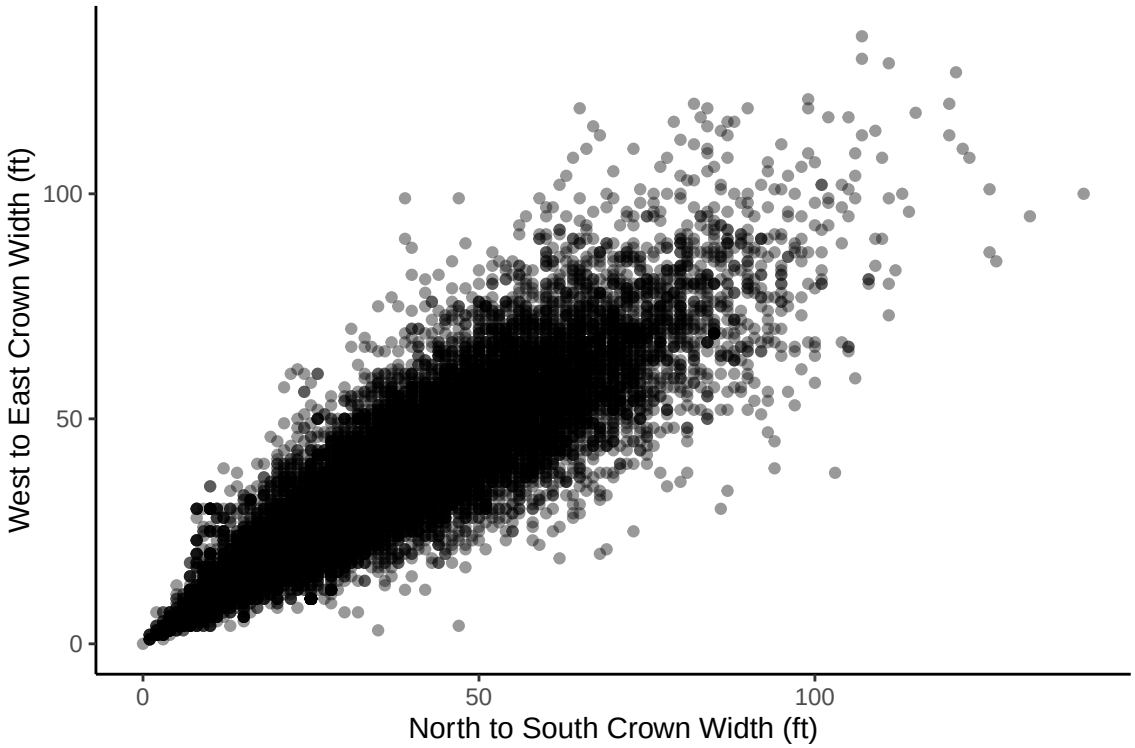

Figure 1.1: Plot comparing North to South Crown Width (ft) to West to East Crown Width (ft). A tree with a higher north-south crown width has a higher west-east crown width, so without loss of generality the north-south crown width variable is used to filter the dataset for larger trees only.

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 1.1 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 1.1: 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

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. Ideally, the raster strips contains training trees of each species, but due to limitations of our ground-level data this is not the case. 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.

Polygon Limitations

The polygons are drawn with the intention of outlining pixels for a known tree species so that the extracted pixels truly represent that species and so that there is a sufficient amount of pixels per species, however, limitations of the satellite images' resolution make some error inevitable. Many parks trees have canopies that overlap, which cause some polygons to contain pixels from trees of different species. Other pixel errors might come from including pixels that are cast in shadow by the polygon's training tree or by surrounding infrastructure. To avoid these errors, polygons with uncertain canopies are cross-checked with the satellite filter on Google Maps. For example, a point in QGIS that is indicated as a Douglas-fir tree might appear as one tree on the low-resolution image from Planet.com, but a closer look at Google Maps shows that it is actually two trees in close proximity. Having low spatial resolution decreases the number of pixels available within a polygon, but that has to be balanced out with the need for a large number of pixels. With regards to the number of pixels

Table 1.2: 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 1.3: 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

per tree species, having around the same number of tree polygons per species results in differing amounts of pixels per species due to the different average sizes of tree canopies for different species. After drawing the images, the pixels totals per tree species is computed to ensure that the training pixels data contains at least 800 pixels per tree species.

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 1.2 displays the first few entries of the pixels dataset. Table 1.3 contains the summary of the total number of pixels for each tree type. Figure 1.2 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.

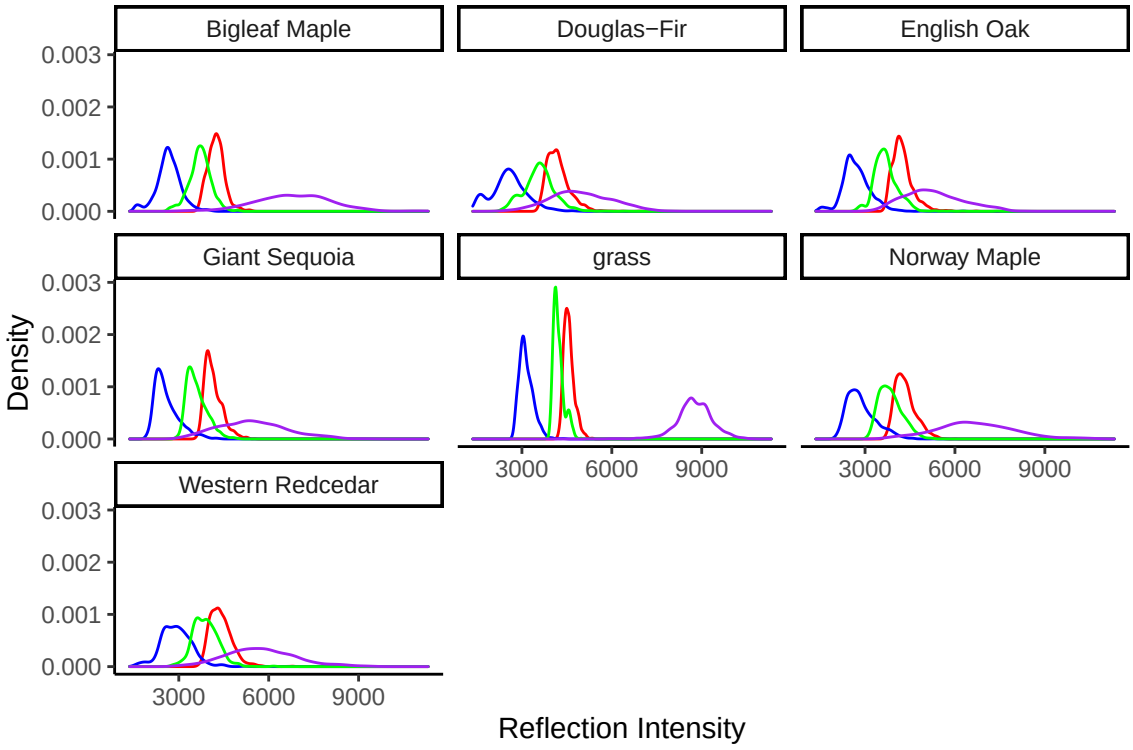
Bigleaf Maple
Douglas−Fir
English Oak
Giant Sequoia
grass
Norway Maple
Western Redcedar
Density
0.003
0.002
0.001
0.000
3000
6000
9000
Reflection Intensity

Figure 1.2

–>

Chapter 2

Methods

2.1 Current Methods

Since the Western redcedar decline is recent, there are currently no methods for identifying tree species like Western redcedars for the Pacific Northwest using satellite imaging. However, a similar study identifies tree species for a strip of land in California by combining ground-level data with satellite images (Fricker et al., 2019). Their methods involved drawing polygons around rastered images that are layered with ground data as explained in Chapter 1, but the pixels from the polygons are used to train a convolutional neural network (CNN) model instead of the random forests and support vector machines in this study. The CNN model was then evaluated with k-fold cross validation. CNN models are appropriate in a classification setting and when working with satellite images because they account for spatial relations, which is likely to appear in classifying tree species. This study follows the methods of preparing the data for modelling as well as the cross validation method to evaluate the performance of the random forests and support vector machines models. Random forests use random feature selections to create decision trees and increase the number of correctly classified observations. Random forests have the advantage of being simplier to train and still performs at a similar level as CNN models.

2.2 Training Models

For both the random forest model and the support vector machines model, the tidy pixels data is arbitrarily divided in half into a training dataset and a testing dataset and the selected variables for training the models are the four bands (red, green, blue, infrared) and an NDVI variable. The `randomForest` function from the `randomForest` package (Liaw & Wiener, 2002) is used to train a random forest model on the training dataset to predict the tree species. The function takes parameters for data, the predictive variable, and the subset data that should be used to train the model. By default, the function grows 500 trees and samples cases with replacement.

The `svm` function from the `e1071` package (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2019) is used to train a support vector machine model on the training

dataset to classify pixels into types of trees. By default, data are scaled to zero mean and unit variance and the kernel is radial.

2.3 Preparing Raster Images

The training model has to be applied to the entire raster image to predict the location of Western redcedars, however, extracting all the pixels from raster images is a slow process. Filtering the rasters to keep vegetation and to recognize the difference between grass pixels and tree pixels alleviates the computational intensity of extracting all the pixels.

2.3.1 Masking Vegetation

The satellite images in RStudio need to be masked to reduce the number of pixels that the model has to classify and prevent the chance of a non-vegetation surface being predicted as a tree. A common mask applied to raster images is a Normalized Difference Vegetation Index (NDVI) mask. This index is a measure of the greenness of a pixel, with higher values indicating vegetation and lower values indicating infertile areas such as a rock. The formula for NDVI is $NDVI = \frac{NIR-Red}{NIR+Red}$. Figure [insert figure title] displays the density of NDVI values over the raster images. To remove the pixels that are not vegetation, an NDVI threshold is determined to be [insert threshold] to create a mask dividing the pixels into vegetation (1) and non-vegetation (0). When the trained model is applied to the raster image, it is applied to the masked raster image that only keeps pixels with mask values of 1.

2.3.2 Grass Polygons

Further preparation for applying the model to the entire mask involved creating grass polygons in QGIS to add a grass attribute in the model. Five fields of grass were outlined as polygons in QGIS. In RStudio, the pixels were extracted from the grass polygons for a total of 2,923 grass pixels. Training the model to also predict grass ensures that a field of grass will not be classified as a tree when applied to the entire raster image.

2.4 Modelling Tree Species in Portland

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

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). *E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien*. Retrieved from https://CRAN.R-project.org/package=e1071

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

Western red cedar. (n.d.). Retrieved from https://www.oregonencyclopedia.org/articles/western_red_cedar/#.YB7fUy1h1N0

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