Demand Prediction Modeling for Utility Vegetation Management

by

Wade Allen McElroy

B.S. Mechanical Engineering, University of Nevada Las Vegas, 2009
M.S. Aerospace Engineering, University of Texas at Austin, 2010

Submitted to the MIT Sloan School of Management and the Department of Mechanical Engineering in Partial Fulfillment of the Requirements for the Degrees

of

Master of Business Administration
and
Master of Science in Mechanical Engineering

In conjunction with the Leaders for Global Operations Program at the Massachusetts Institute of Technology

June 2018

Signature of Author____ Signature redacted

MIT Sloan School of Management
Department of Mechanical Engineering
May 11, 2018

Certified by____ Signature redacted

Georgia Perakis, Thesis Supervisor
William F. Pounds Professor of Management
MIT Sloan School of Management

Certified by__ Signature redacted

Konstantin Turitsyn, Thesis Supervisor
Professor, MIT School of Engineering

Accepted by_____ Signature redacted

Maura Herson
Director, MIT Sloan MBA Program
MIT Sloan School of Management

Accepted by_____ Signature redacted

Rohan Abeyaratne
Professor and Graduate Officer
MIT Mechanical Engineering

This page left intentionally blank

Demand Prediction Modeling for Utility Vegetation Management

By

Wade Allen McElroy

Submitted to the MIT Sloan School of Management and the MIT Department of Mechanical Engineering on May 11, 2017 in partial fulfillment of the requirements for the degrees of Master of Business Administration and Master of Science in Mechanical Engineering

Abstract

This thesis proposes a demand prediction model for utility vegetation management (VM) organizations. The primary uses of the model is to aid in the technology adoption process of Light Detection and Ranging (LiDAR) inspections, and overall system planning efforts.

Utility asset management ensures vegetation clearance of electrical overhead powerlines to meet state and federal regulations, all in an effort to create the safest and most reliable electrical system for their customers. To meet compliance, the utility inspects and then prunes and/or removes trees within their entire service area on an annual basis. In recent years LiDAR technology has become more widely implemented in utilities to quickly and accurately inspect their service territory.

VM programs encounter the dilemma of wanting to pursue LiDAR as a technology to improve their operations, but find it prudent, especially in the high risk and critical regulatory environment, to test the technology. The biggest problem during, and after, the testing is having a baseline of the expected number of tree units worked each year due to the intrinsic variability of tree growth. As such, double inspection and/or long pilot projects are conducted before there is full adoption of the technology.

This thesis will address the prediction of circuit-level tree work forecasting through the development a model using statistical methods. The outcome of this model will be a reduced timeframe for complete adoption of LiDAR technology for utility vegetation programs. Additionally, the modeling effort provides the utility with insight into annual planning improvements. Lastly for later usage, the model will be a baseline for future individual tree growth models that include and leverage LiDAR data to provide a superior level of safety and reliability for utility customers.

Thesis Supervisor: Georgia Perakis
Title: William F. Pounds Professor of Management Science, MIT Sloan School of Management

Thesis Supervisor: Konstantin Turitsyn
Title: Associate Professor, Department of Mechanical Engineering

This page left intentionally blank

Acknowledgments

I would like to thank the team of vegetation management experts that assisted in this work. Your unwavering support and attention to this project was noticeable and meaningful. The task that you all take on is an important one, and I'm glad that you took the time to consider my contributions in making that work a safer, more reliable, and better for all of the residents in the service territory.

I'd like to thank my thesis advisors Georgia Perakis, and Kostya Turitsyn for their support of this work. Additionally, I'd also like to thank the Leaders for Global Operations program for their guidance and support throughout the program.

I'd like to thank my parents Dave and Kim McElroy who have shaped me into the person that I have become today. Without your guidance, support, lessons, humor, and spirt I would have never been able to accomplish anything that I have. And lastly, I would like to thank my girlfriend Nicole who supported me throughout this entire two year adventure. Your kind words, subtle motivation, helpful and lighthearted distractions, endless ability to listen, incredible and unparalleled outlook on the world, and outright excellence in all that you do drove me to be just a bit better. I could not have asked for a better companion to have by my side on this journey.

This page left intentionally blank

Contents

List of Figures

List of Tables

List of Equations

This page left intentionally blank

1 Introduction and Background

1.1 Utility Vegetation Management Overview

Utility vegetation management (VM) programs provide preventative maintenance for asset management; for this thesis the focus will be on trees and electrical power conductors. Vegetation surrounding assets are regularly trimmed or removed to prevent growth that would otherwise damage the assets. The work is completed in accordance with both state public utility commission and federal regulatory requirements. Examples of regulatory requirements are depicted in Figure 1, and show that the risk to assets is predominantly focused on physical clearance from the conductor. VM work is critical to a utility safely and reliably providing service in its territory, but can also present large challenges. One notable challenge is that scale associated with ensuring that the entire system is within compliance, where a system network or grid can span over 100,000 miles of overhead conductor and in excess of 100 million trees within reasonable proximity.

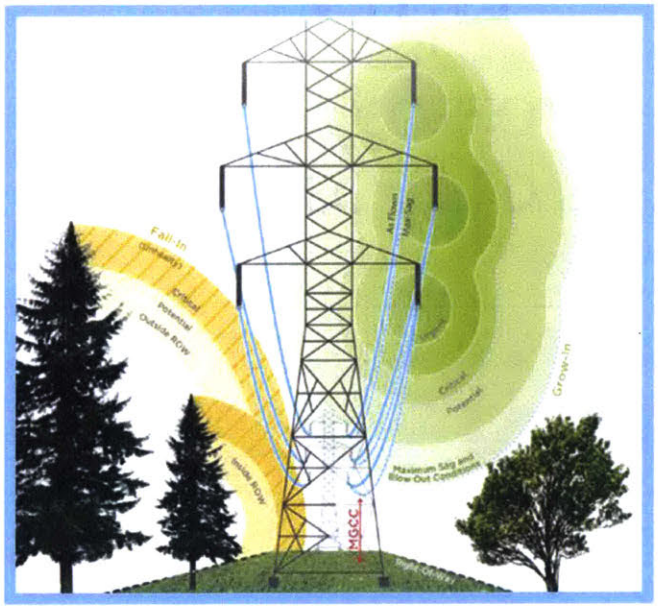

Figure 1: Electrical Compliance Example[1]

In response to the issues of large scale with high regulatory compliance metrics, VM groups will deploy a program similar to the one shown in Figure 2 to provide routine tree work. The major facets of VM programs are:

- **Planning and scheduling of work:** when covering the entire service territory it is critical to create a plan for the expected workload, it is customary for utility planners, patrollers, and tree workers to participate in this planning session.

[1] https://quantumspatial.com/our-solutions/electric

- **Circuit patrols:** consist of visual inspection, typically by walking, of an entire circuit and identifying specific trees that have the potential to become out of compliance in the given year - presumably all of the trees visited will be in compliance and will only need to be worked for growth in the coming year.
- **Performing tree work**[2]**:** where the trees that have been prescribed work from the patrols are physically treated. The tree crews will apply best practices for preforming the trims as to not harm the long-term health of the tree, but also strive for a trim that will not require revisiting the tree for multiple years[3].
- **Quality assurance and control:** quality steps occur following a patrol and/or a tree working step to ensure that standard procedures and regulatory requirements are followed. Examples include, verifying that trees that needed to be prescribed were, and conversely trees that did not require work were not.

The two most significant areas of work revolve around patrols and tree work. Patrols are performed for the entire service territory by trained specialists that prescribe work to ensure that both the utility assets are protected and to ensure the trees remain clear of electric facilities. The tree work prescription is then provided to an independent group or company that carries up the defined tree service. Lastly, another level of inspection, quality assurance and control, is conducted on both the patrol and tree work to confirm the general process does not have any issues. Utilities can select to perform this process over any time frame they deem suitable to stay within regulatory compliance; a typical selection is annually.

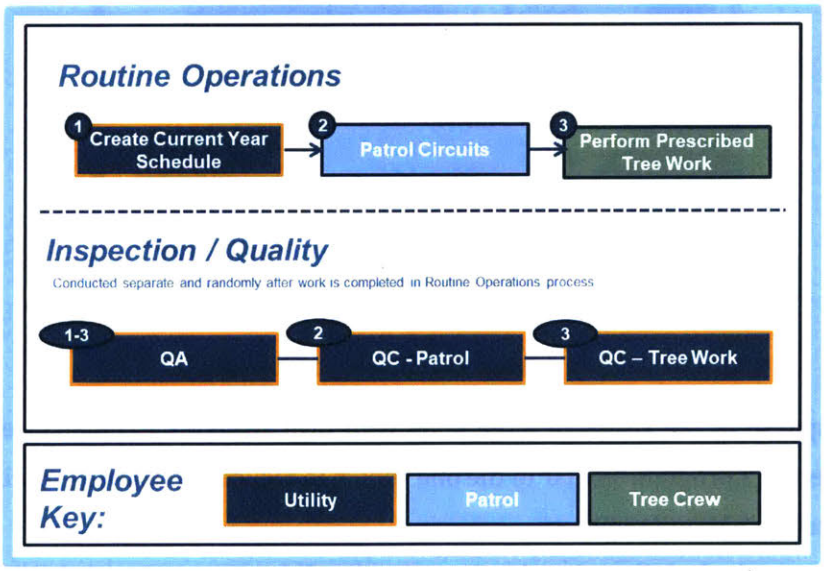

Figure 2: Vegetation Management Program Process Flow Diagram

[2] *Tree work* is a general phrase in vegetation management to imply the trimming and/or removing of trees to provide adequate clearance.

[3] A best practice would be a minimum of three years

1.2 Technology Adoption in Vegetation Management

Despite the possibility to implement VM programs with a relatively manual process, new technology has and will continue to play a key role in VM. The linkages in the routine operations of Figure 2 between the utility, patrol, and tree crews in the field have been aided by the implementation of a digital records system over the last 20 years. This digital system has built up to include instantaneous generation of electronic work instructions, tracking work progress, and quick and reliable access to previous work conducted. In recent years, VM programs have been investigating and testing the possibility for the following technologies to improve their operations reliability and efficiency.

1.2.1 Remote Sensing

One major area of assistance for VM programs comes via remote sensing technologies, or electronic observation from a distance. Within this broad definition there exist multiple sensor packages and even further platforms to deploy these sensors. Each of the sensors can span wide ranges of spectra from infrared to ultra violet, and can be installed on platforms that include: unmanned aircraft systems (UASs, or more commonly referred to as drones), rotary aircraft (helicopters), fixed-wing aircraft and orbital satellites. Although all of these potential systems can provide meaningful information to vegetation managers, Light Detection and Ranging (LiDAR) on fixed wing aircraft has provide the greatest promise.

LiDAR is a technology that uses the reflection of visible light to measure distance. The system collects numerous distance measurements (on the order to 2 to 100's of points per square meter) by performing continuous swaths of the system and finally creating what is referred to as a point cloud. Figure 3 shows a representative point cloud that contains millions of induvial points that make up multiple trees, houses, and utility lines. The color of the individual trees in Figure 3 signifies unique tree species, which is determined through the process of hyperspectral[4] imagery analysis. This information is then fused with the LiDAR point cloud to produce a multivariate, single data set.

[4] *Hyperspectral* refers to additional sensors information that is acquired in the infrared, visible, or ultraviolet wavelengths that is then mapped to the individual LiDAR point cloud point. In essence appending additional information to that point in 3D space.

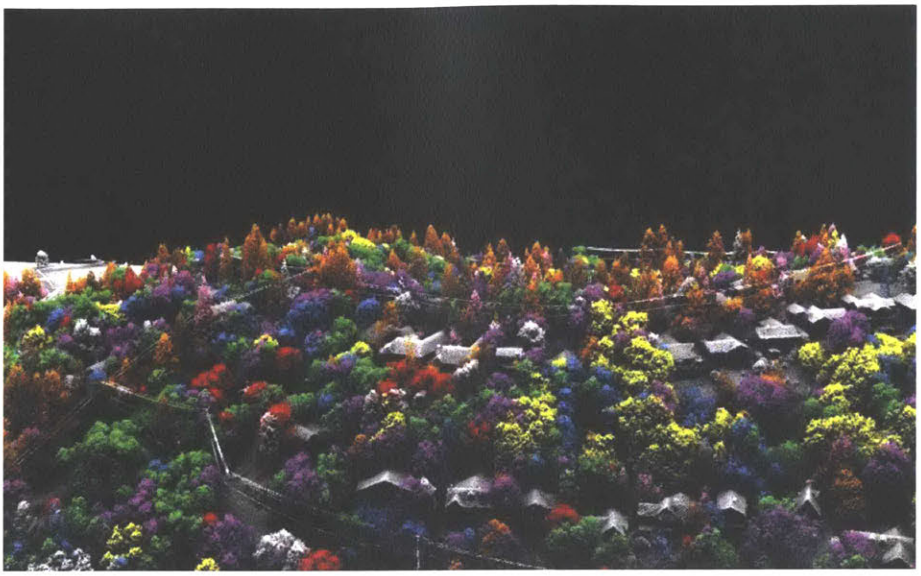

Figure 3: Representative LiDAR point cloud[5]

The key advantage of LiDAR to vegetation mangers is the ability to quickly and accurately acquire information on the clearance between vegetation and conductors over thousands of miles of overhead electrical lines, while also producing a data source that can be revisited and reused. Under the current system, people are required to manually patrol the entire network to make visual inspections that cannot be re-examined outside of their recommended prescriptions without physically revisiting the site.

1.2.2 Tree Growth Modeling

An additional advantage of LiDAR is extremely granular information about individual trees. Currently VM programs focus at the parcel level and the number of trees located at that parcel. However, the continuity of an individual tree record is not guaranteed. With the implementation of LiDAR, tree level records will be available, if not required. In practice the information gathered could even be one level lower at the individual branch levels, as this is fundamentally what is measured in a LiDAR scan and trimmed by the tree crews. With this added information it would be possible for VM programs to investigate the use of tree growth models. At its fullest capability, this would allow for a process flow where patrolling of a circuit may not be required, as the system would have a statistical prediction of whether each tree needs work or not. A future state system like this would be extremely challenging, if not infeasible, without the measurements provided by LiDAR.

1.3 Problem Statement and Goals

Despite all of the advantages of implementing a LiDAR program, an issues arises with technology adoption, where VM programs encounter the dilemma of wanting to pursue LiDAR as

[5] https://quantumspatial.com/our-solutions/electric

a technology to improve their operations, but still cannot immediately trust the output data. Subsequently, the VM groups find it prudent, especially in the high risk and critical regulatory environment, to test the technology. The nature of the test would be comparing the number of detections[6] to an expected value provided externally or by direct comparison to a human patroller. However, during the testing process, a problem arises associated with how to anticipate the number of detections given the degree of variability in the number of units worked annually.

The goal of this thesis is to address the problem of unknown circuit-level tree work counts by applying modern statistical methods to predict annual trees worked, and variables that are influential to those deviations. The result of applying this model will be a decreased timeframe for full scale adoption of LiDAR technology for utility VM programs.

1.4 Focus of Project

Within VM programs there are details, such as groupings of work, which are not fully covered in Figure 2. In order to properly scope the predictive model, focus areas were required to make the model both general enough to be comprehensive, yet detailed enough to provide accuracy and usefulness. The result of this scoping is a model concentrating on the *distribution* system for *routine trims* on an *annual* basis for *circuits*. Below is further explanation of the decisions made in the scoping effort.

1.4.1 Distribution

The scope will be on the electrical distribution system. Although a model could be created separately for a transmission system, the differences in regulatory structure made it more suitable to only provide a model for distribution circuits.

1.4.2 Circuit Level

The scale of the model will be at the circuit level. Utilities, and VM programs by association, will customarily divide required work within the service territory by a geographic hierarchy from the full system to geographic divisions to circuits. Where circuits are the general name given for a contiguous stretch of conductor from the distribution substation to the termination of the line, the model will attempt to predict at this level. This selection was made due to the applicability to LiDAR detections and furthermore to the general work planning structure used in VM programs. Divisions and systems were not selected, or even analyzed, because data inconstancies and nuances were found at the individual circuit levels. Furthermore, individual trees in a logistic model (trim this year or not) were not pursued because the data was not yet available with any confidence.

---

[6] *Detections* is a term used to imply a set of criteria, mostly distance, are met for an induvial tree as determined through LiDAR data analysis. For example, if using Figure 1 for Distribution a LiDAR detection would be when a tree is found to be within 10 inches of a conductor. Although this is inside of the regulatory compliance range, the detection distance and logic can potentially be greater.

Lastly it is critical to distinguish that predictions are made for each circuit on an annual basis which, for ease, will be referred to as a *circuit-year*. This grouping is reasonable as the patrol and tree work for an individual circuit is traditionally carried out on an annual basis over a couple week span.

1.4.3 Routine Trims

The unit being predicted will be tree trims. This base unit is applied mainly because it is the most commonly performed VM program operation, as well as the clearest to predict. However VM programs conduct other work, a few of additional services that are performed but not included in the prediction model are:

- **Brush:** not all trees have clear distinct branches that can be trimmed and tracked, instead brush units can be applied to wispy branches that are measured on an approximate volume removed compared to a single tree.
- **Removals:** at times patrollers may prescribe the removal a tree, for example there are some fast growing species that may have grown to a point where compliance may not be guaranteed and removing the tree is the best course of action. Note, this value is still included in the model because a tree that was worked over many prior years that will not be worked in the future due to removal could be helpful for prediction.
- **Hazard Trees:** in the case of a dead and dying tree, patrollers will prescribe a unique removal associated with unhealthy trees.
- **Proactive Reliability Tree Work:** this group of tree working, both trims and removals, is related to additional operations that are conducted in a truly preventative fashion. The trees in this category are healthy and not out of compliance, but are a risk to safely and reliably providing electrical service. This work can be conducted for example on specific species that might be subject to shedding branches, or geographically based on historical weather issues that have resulted in downed power lines.

1.5 Thesis Hypothesis

The hypothesis tested in this thesis is that with statistical modeling it is possible to predict the number of number of trees trimmed in a given circuit-year at a higher accuracy than a naïve model or simple running average. The increase in model accuracy is attributable to a collection of additional variables that help indicate the number of trees worked, in addition to clustering similar circuits before to model creation.

1.6 Thesis Contributions and Outline

This thesis describes the process of creation, rationale, and results for numerous predictive models from a representative vegetation management program. In Chapter 2, the assumptions, sources, and variables implemented, and more importantly variables excluded, from the data

frames have are discussed. Chapter 3 conducts data exploration consisting of plotting of the variables presented in Chapter 2 to inform the model creation process, with a specific emphasis on avoid selecting collinear variables. Chapter 4 describes the statistical models implemented, the assumptions made with each of the models input variables, and lastly describes the process to be implemented for variable and model selection. Chapter 5 presents the results of the modeling process and short comings observed from the multiple model combinations that could be implemented. The final result is a model that marginally improves the volume weighted Mean Absolute Percentage Error (vwMAPE), by 1.1%, and symmetric Mean Absolute Percentage Error (sMAPE), by 0.3%, over a simple running average prediction. Although these seem to be less than stellar results form a predictive sense, the results provide guidance for LiDAR adoption, meaningful variables, as well as overall vegetation management operations procedures. Finally Chapter 6 provides summarized findings, recommendations of future work, and major contributions. Below is a list of the notable contributions:

- The most useful predictive variables are the US Drought Monitor Index values and VM program process improvements metrics
- Continuous weather variables, not discrete variables that contain for example the number of days a criteria is met, are superior for tree trim prediction
- Average weather variables, not extremes for a time period, are superior for tree trim prediction
- Clustering prior to modeling created marginally better performance, with clustering by the running average number of units being superior to other more sophisticated (i.e. K-Means) clustering methods
- Linear regression returns superior models than CART and Random Forest models

1.7 Literature Review

Previous research in the area of vegetation growth predictions have focused in two major areas: large scale growth patterns and local growth of pruned high value vegetation; such as, olive trees [1], and pears [2], where pruning is used as a way of increasing growth and subsequently the per plant yield.

Large Scale Growth Modeling - The large scale growth research has developed multiple growth models, with seminal analytical publications starting around 1933 [3]; however, when applied to separate regions and/or problems numerous modeling updates have been evaluated in [4] [5], and [6]. [6] focused on vegetation management as whole, and presents a four-tiered system of models that consider temporal aspects of modeling from strategic planning (50-100 years), long-term planning (10-15years), multiyear planning (3-5 years), and operation planning (up to 1 year). Similar to the range in years of models, the variables considered also have wide range is the physical size and diversity – from individual tree growth mechanism at the cellular

level, to tree nutrient completion and shading at the forest level. Some attempts have been made in vegetation management groups to predict individual species growth in small samples [7], but the largest research focus has been related to predicting vegetation outages [8], [9], [10].

Local Growth Modeling - More recent studies [1], [2], [9] have begun investigations into local growth via the use LiDAR based data as a variable, with specific research traction in the applications of pruning volume perditions for biomass energy use [1]. Although research into tree growth modeling has been conducted across a wide array of topics and variables, the regime of vegetation growth pertaining to utility trim predictions on an annual level is a limited.

However, the research in the area of predictive modeling at utilities has become more common. For example, prior work from [8], [11], and [12] have shown promising results when applying statistical modeling to utility based questions. Specifically, [8] highlights the value of vegetation-related data.

2 Data Sources

2.1 Overview

There are six main data sources that were used for the creation of the data frame analyzed: Work Management System (WMS), Project Management Database (PMD), Tree Growth Data, Circuit Location Information, Weather Data, and Drought Data. In total these databases combined to create a data frame of 14 project years, comprised of 35,960 circuit-years, and 100+ variables that could be utilized for prediction.

2.1.1 Work Management System (WMS)

The Work Management System (WMS) is the database that contains all of the tree work that has been conducted in the service territory for the last 15+ years. For the purposes of the data frame creation, information was evaluated from 2001 to 2017. The variables provided from this database include:

- **Total Trims:** the prediction variable for the models; it depicts the number of routine trims that occurred in a given circuit-year.
- **Project Year:** the organizational year in which tree work is conducted. Assumed to begins on October 1 and ends on Sept 30 of the following year. For example, work conducted on circuit ABC between October 1, 2016 and September 30, 2017 is assigned to the 2017 project year.
- **R1 Removals:** are a specific group of removals for trees that have a diameter at breast height (DBH) of less than 12 inches but greater than or equal to 4 inches. All removals less than 4 inches are considered Brush and not counted in this metric.
- **Division[7]:** the regional grouping created by the VM group to divide work evenly. This categorical variable was included but not used in the data frame as a potential clustering variable where historical groupings could have nuanced meanings not fully described by the other variables.
- **Species Trim Data:** the Total Trims variable is an aggregation of individual tree trim records. For instance, these records will indicate that on 12/23/2007 a Palm Tree with a DBH of 17 inches was trimmed at 123 Main Street on Circuit ABC. Although plenty of information is contained in this type of record, the purpose was to obtain the specific species information and combine all of the identical species records up to the circuit level. The final result being circuit-year ABC-2007 had X Palm Trees, Y Black Oaks, Z Orange Trees, etc. The result of this system is combined with information from Tree Growth Data in Section 2.2.3.

[7] Not implemented in final models

2.1.2 Project Management Database (PMD)

The Project Management Database (PMD) is similar to WMS in that it provides insight into the trims conducted in a circuit-year; however, the PMD contains additional information about the project in that year. Some of the unique project level information found in PMD is:

- **Project Quarter:** organizationally VM programs might elect to group circuits by seasonal quarters for planning purposed. Similar to the Divisions variable it is included as a way of exploring potential grouping of circuits made by VM experts.
- **Planned Units:**[8] at the beginning of a project VM mangers are required to insert a planned number of trims. Presumably this variable is a culmination of knowledge and negotiation with patrollers and tree trimmers in the field. Although this is included in the data frame, it was excluded from model creation due to a high degree of variability for the actual units. This variability is assumed to be by mangers not applying a consistent timing and rationale for updating this variable.
- **Tree Contractor:**[9] is a categorical variable for the group that provided tree work for that circuit-year, as it is common that a single tree crew to cover an entire circuit every year. The rationale behind including this variable is gaining potential insight into the influence that an individual tree work group might provide. That is, a certain group may typically trim more trees than the average over last year's units, where another contractor might do the opposite. This variable was omitted from model creation due to challenges in data cleanliness.

2.1.3 Tree Growth Data

The Tree Growth Data was provided by a vegetation patroller as their stratified (slow, medium, fast, etc.) assessment document for how quickly a tree might grow on average in the service territory. The document maps over 150+ species to the following categories: SF (Super-Fast), FSF (Fast-Super-Fast), F (Fast), MF (Medium-Fast), M (Medium), SM (Slow-Medium), S (Slow), and NA (Not Applicable). These variables were added to the data frame as a way of potentially characterizing similar circuits.

2.1.4 Circuit Location Information

The location of the circuit was provided by the utilities GIS team. For ease, the circuit locations were taken to be at the geometric midpoint of the circuit. No additional variables and / or weightings were applied in an attempt to compensate for circuits that were longer or shorter. The implications associated with this location method are most impactful in the linking of weather information.

[8] Not implemented in final models
[9] Not implemented in final models

2.1.5 Weather Data

One of the most crucial groups of variables to include is weather data. The weather data utilized in this model was obtained from the U.S. National Oceanic and Atmospheric Administration (NOAA) "Global Summary of the Month"[10] dataset. This specific weather data source was pursued due to successful implementations in [8] and [11], an overarching reliability of NOAA weather stations as a whole, and a diversity of locations. Monthly averages were selected as a way to still identify extreme conditions, while not becoming too granular (such as daily values) and large to become a hindrance during modeling. Additional modifications are discussed in Section 2.2.1. The following variable were added to the data frame:

- **Monthly Average Minimum Temperature (TMIN):** Average of daily minimum temperature
- **Monthly Average Maximum Temperature (TMAX):** Average of daily maximum temperature
- **Monthly Average Temperature (TAVG):** Computed by adding the unrounded monthly/annual maximum and minimum temperatures and dividing by 2.
- **Extreme maximum temperature for month (EMXT)**
- **Extreme minimum temperature for month (EMNT)**
- **Number of days with maximum temperature >= 90 degrees Fahrenheit (DX90)**
- **Number of days with minimum temperature <= 32 degrees Fahrenheit (DT32)**
- **Total Monthly Precipitation (PRCP)**
- **Highest daily total of precipitation in the month/year (EMXP)**
- **Number of days in month with >= 0.01 inch (DP01)**
- **Number of days in month with >= 1.00 inch (DP10)**

These variables were selected to investigate if there is stronger predictive power associated with average weather patterns, more extreme weather conditions, or the frequency of being above or below certain thresholder vales. Excluded from the data frame are variables related to: wind, snow, and soil conditions. These were all excluded due to their sparse nature in the data frame, in essence there are very few stations that record these parameters regularly. Although snow could be predictive by providing mountain snows that support long-term water aiding vegetation growth, it was excluded as relating snow fall at the closest weather station would not necessarily imply that the given circuit would benefit from that snowfall.

2.1.6 Drought Data

In addition to the weather data explored, US drought information via the Palmer Drought Severity Index was added to the data set. Monthly data by climate division was obtained from the NOAA National Climate Data Center[11]. Interestingly, the divisional basis of the data is far coarser

[10] https://www.ncdc.noaa.gov/cdo-web/datasets
[11] https://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp#

than the individual weather station data, as shown in Figure 4. The result is there will be limited distinction among the individual circuits but potential power within the overall variable on an annual basis – meaning, years of greater or lesser drought will be discernible in the models. Additionally the Modified Palmer Drought Severity Index was included due to a high correlation with the other Palmer Index[12] and is considered the operational variant of the index since 1991 replacing the original Palmer Index that was developed in 1965 [13]. Lastly, the Vegetation Drought Response Index (VegDRI) was not included because the current VegDRI model is currently being developed for data back to 1989 [14].

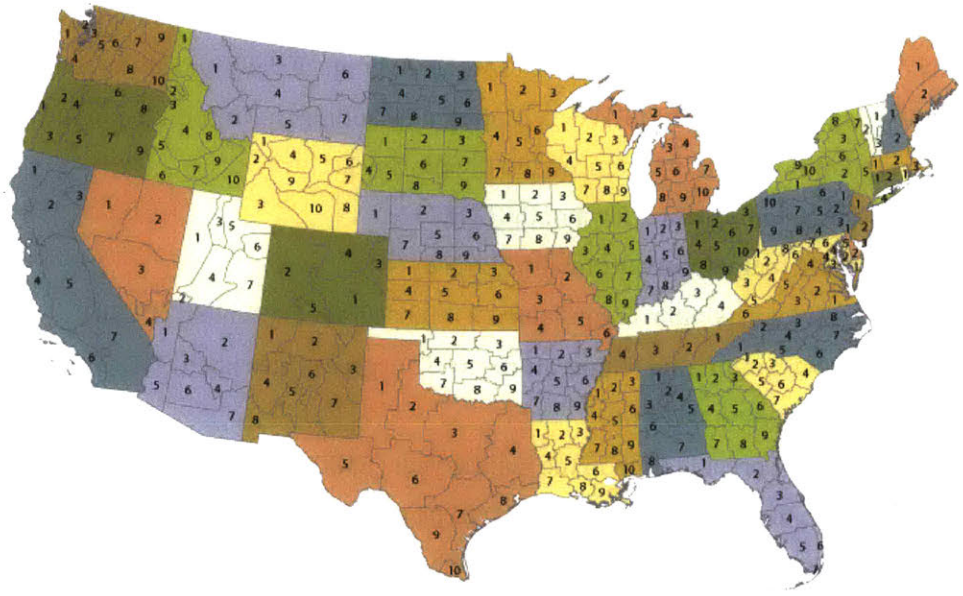

Figure 4: NOAA U.S. Climatological Divisions[13]

2.1.7 Practices Variable

Over the range of years in the dataset there have been numerous subtle changes to overall vegetation management process and procedures – with changes seeking to install process improvements to better carry out the annual work. To capture these alterations a comprehensive variable was provided by vegetation management experts to adjust for these changes. The variable is applied to all circuits in a given project year, no specific weighting were provided for specific circuits, or circuit-years[14].

[12] Other drought indexes include: Palmer Z Index and the Palmer Hydrological Drought Index, and the Vegetation Drought Response Index (VegDRI)
[13] https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-divisions.php
[14] This could be a meaningful implementation to consider for future models as a form of outlier suppression.

2.2 Preparing the Data Set

The original data sources from Section above2.1 all needed to be linked to a given circuit-year, along with additional processing conducted on a few the raw values. The process and assumptions made are listed in the follow sections.

2.2.1 Weather Data

Among all of the data sources, the weather data was the most involved to link to a desired circuit-year. The start the linking process, the closest weather station for the desired given circuit-year was identified. Next, all available weather information from that weather station for the *previous* project year was assigned. Figure 5 shows a typical operational year, noting that the weather data is obtained from the previous project year, not calendar year preceding it. For example, Fall weather data for a circuit-year of 2017 is obtained from 2015. This approach assumes that the predictions for the upcoming project year are made at the beginning of that project year. The implication of this method is that for a circuit that is inspected and trimmed late in a project year, for instances August, the data associated will not contain any of the most recent weather that has been encountered since September of the prior year, up to an 11 month gap in the August example. This scheme, despite the late project year issue discussed, was utilized because the model will also be used for planning purposes that are currently solidified before the start of new project year, and would require future weather inputs to be valid.

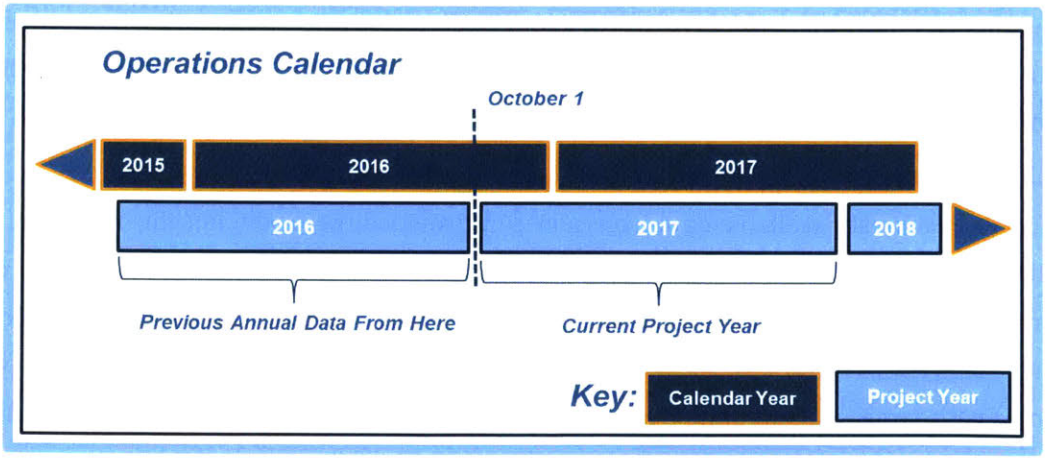

Figure 5: Operational Calendar for System Analyzed

Once the weather information for the closest circuit is applied, a check is completed to ensure that all weather parameters were assigned, as it is highly possible that a given weather station may not have all of the weather variables desired – for example, there may be a station that only records temperature and not precipitation. If weather variables were not all filled, the next closest weather station is found and any missing values were assigned to the circuit-year. This

process is repeated until a full weather data frame is obtained. Figure 6 shows the distances for all of the weather stations, note that most weather stations are reasonably close to the circuits, with a mean of 6.4 miles and max of 26.4 miles.

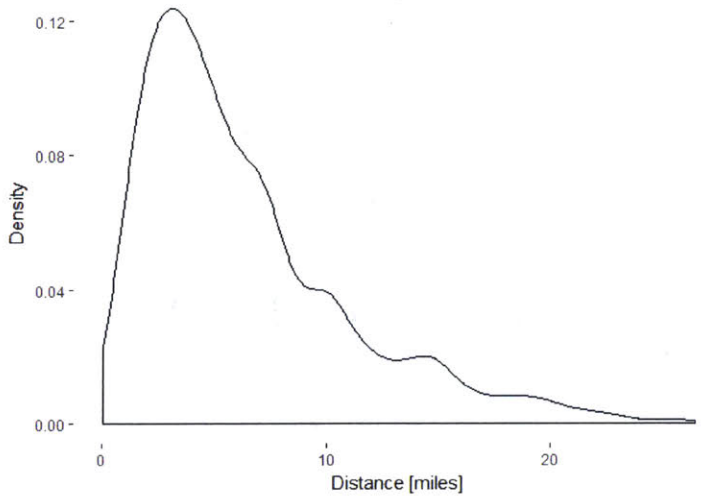

Figure 6: Weather Station Distance Density Plot for 2017 Circuits

Finally, weather data is grouped by season (Fall (FAL): October to December, Winter (WIN): January to March, Spring (SPR): April to June, Summer (SUM): July to September) to limit the number of variables considered in the model. This step was taken to limit the likelihood of collinearities between similar months and additionally shrink the data frame for decreased computational time. This operation decreased the number weather variables from 156 to 39.

2.2.2 Creating Running Average and Standard Deviation Parameters

Previously available data from a circuit provided the model insight into the size of the circuit. Two variables were initially considered: last year's total trims and a running average of prior year's total trims. For modeling, the running average was implemented in favor of the single previous year. Within a single year, there was a significant lag term that presented a source of error with large shifts in units from year to year, which significantly impacted predictions. For the running average a maximum of five previous year total trims was used. When there were less than five years of data available, the running average would take only the available data, for example, for circuit-year of 2004 the running average used only 2002 and 2003 trim information. Although this created circuit years with less than five years running average, the method was implemented with the assumption that including the earlier years in the data set with smaller running averages was better than creating a smaller training set. This assumption was not verified in the modeling conducted.

The same method discussed is used for calculating the standard deviation; however to ensure that the standard deviation was reasonably calculated, three values must have been present. This selection omitted all project years for 2001 and 2002, as 2003 was the earliest that three years' worth of information was available to perform the standard deviation calculation. This variable was not used in general prediction, and is only applied to the model presented in Section 5.4.2 where the normalization is performed on the total trims.

2.2.3 Species Growth Data

The final manipulated group of parameters is related to species growth data. This process consisted of taking the WMS generated species data from Section 2.1.1 and relating them to the growth rate (i.e. Slow, Medium, etc.) of the individual species. First, all of the species were assigned to a growth rate from their individual species. Next, the percentage of that growth rate making up the circuit was assigned to each growth rate[15]. Figure 7 shows the distribution of growth rates for 18 of the geographic divisions studied. This method was pursued due to a recommendation in [8] that stated species concentrations was a better predictive variable than land coverage information.

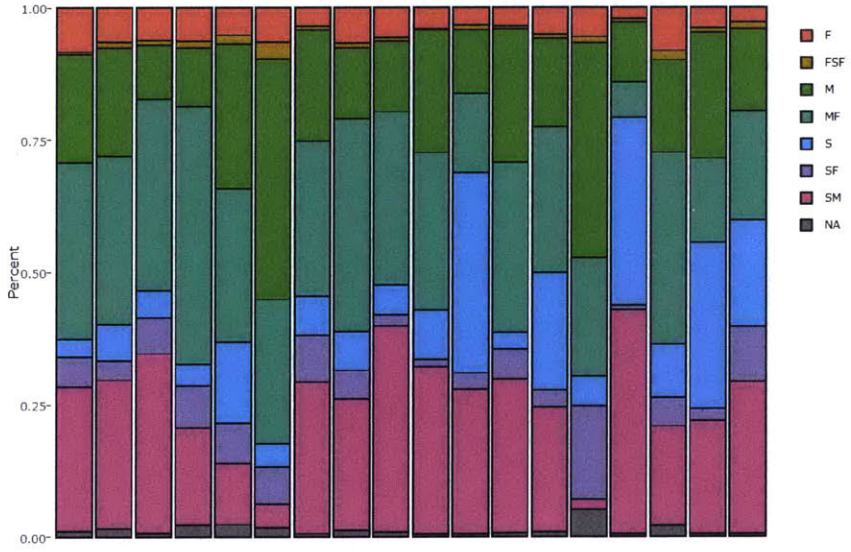

Figure 7: Divisional Breakdown of Tree Growth Rate Types

Additionally, the top three species percentage of the total circuit unit count[16], calculated in a similar set of steps with growth rate, were also added to the data frame. Percentages were utilized for these variables as a way of clustering similar circuits, but were not included as raw

[15] This implies a circuit-year could have, for example, F = 0.625, M = 0.125, and S = 0.250, with the rest equal to zero.

[16] An example of the species percentage variable is a circuit that has 25 oak trees, 12 palm trees, 10 birch trees, and three other species with 4 trees in a given circuit-year. The highest tree species percentage of the entire circuit, in this case 42%, 20%, and 17% would be reported.

unit values to avoid high correlation with the running average values that are assumed basis for circuit-year scaling.

The rationale behind creating these species variables was related to clustering, where it could be advantageous to group circuits with high concentrations of similar species together. For example, a cluster of *fast* circuits could grow at a rate that is independent of weather variables; that is, despite the best trim practices a circuit full of Palm Trees, a Super-Fast species, would need to be revisited annually regardless of the weather conditions. On the other end of spectrum, a circuit compromised of slow growing species might have an increased trim cycle and subsequently might react to long term weather patterns in a unique way. The inclusion of these variables allows for the creation of these type of clusters.

3 Data Exploration

The exploratory data analysis is dedicated to discovering any unique features of the data frame and evaluate any potential useful and/or harmful relationships found within the data. The areas explored include the prediction variable, total annual trims, and the weather data.

3.1 Distribution of Trims

Figure 8 and Figure 9 depict the probability density function for each project years' total trims. Figure 8, a plot of the raw total trims count, shows that many of the circuits in the data frame are skewed to the lower trim count. Figure 9 preformed a log transformation[17] and shows that this transformation creates a nearly normalized circuit distribution. This transformation was not pursued for further analysis due to the fact that the higher unit circuits, which also have higher variances, were challenging to predict in the log form. This was amplified by the face that small log errors were amplified to meaningful raw unit prediction errors that ultimately hinder accuracy metrics. This form of transformation also expands the lower range units further, giving a wider range to values that are practically helpful for accuracy metrics. Lastly, the observed difference in the 2003 distribution can be attributed to only having 82 observations, versus the typical 2,000+ circuits per year.

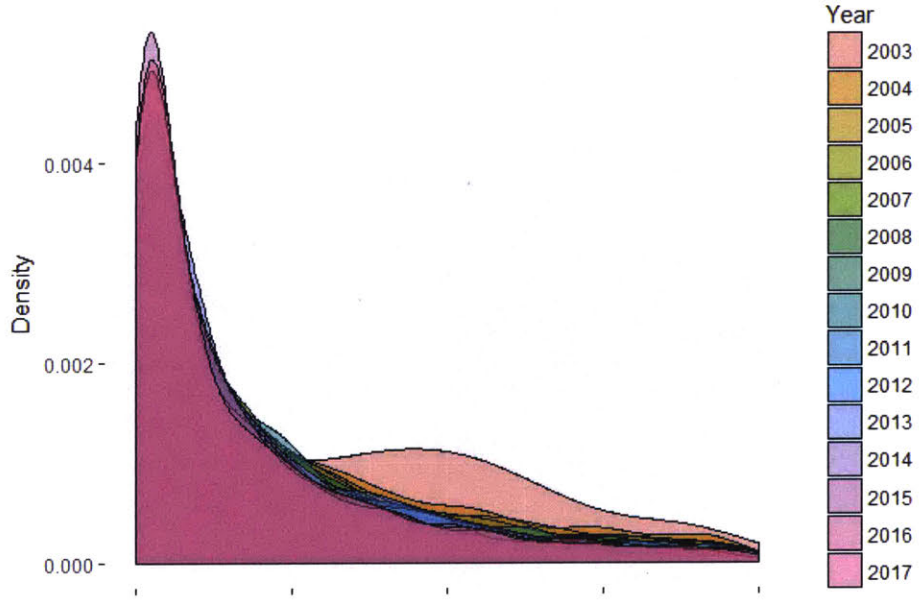

Figure 8: Total Trims Probability Density by Year

[17] A transformation of: $log_{10}(TotalTrims + 1)$ was applied. The addition of 1 was added to avoid issue with the circuit-years that reported 0 trims.

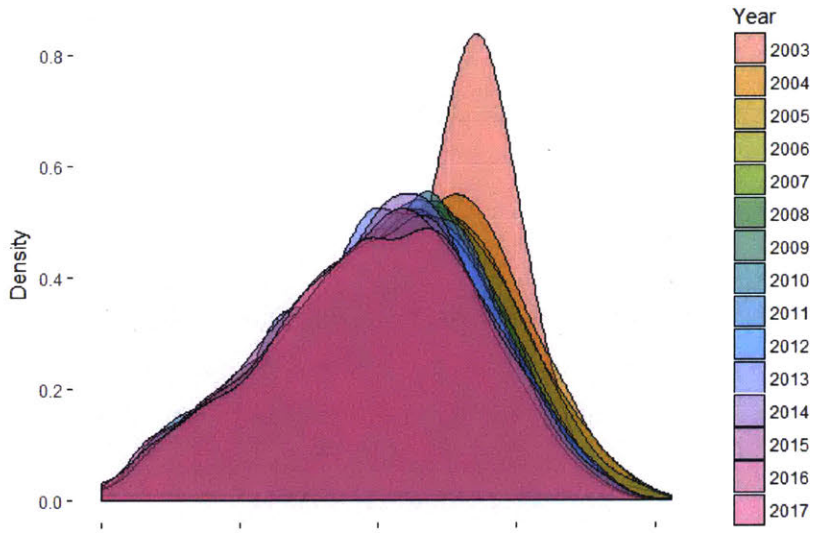

Figure 9: Log Transformation of Total Trims Probability Density by Year

Another transformation investigated was normalization, using the running average and standard deviation. This transformation is shown in Figure 10. This transformation is of interest as it will provide insight into what makes a circuit-year have a number of units above or below a mean year, and thus will evaluate the error from the mean. An interesting observation when the total trims are presented in this fashion is that the center of the distributions are all slightly below 0. For a perfectly normal distribution the errors would be expected to be centered at 0, this tends to imply that there have been a general decrease in the number of units over the years – as the mean value applied is taken from the prior year's averages.

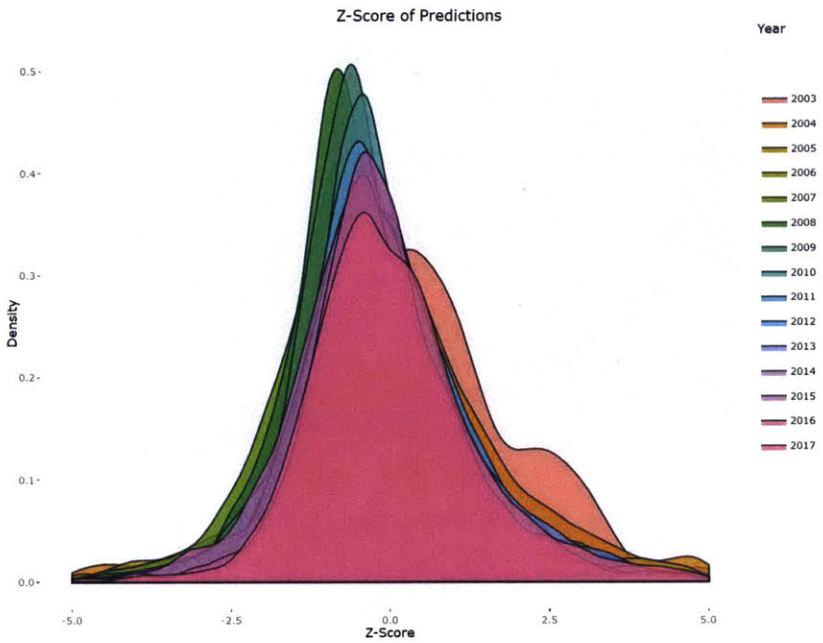

Figure 10: Transformation of Total Trims by Normalization

Figure 11Figure 11, an annual boxplot of circuit-year trim, is an attempt to access the prevalence of outliers on an annual basis. The resulting data shows that a large percentage of the circuit-years are classified as outliers. The maximum circuit in a year also appears to be decreasing over the range of years, where 2004 and 2005 contain circuit-years that are roughly two times that 2016 or 2017. Figure 12, a zoomed version of Figure 10, shows a similar trend with the medians and third quartiles deceasing over of the project years.

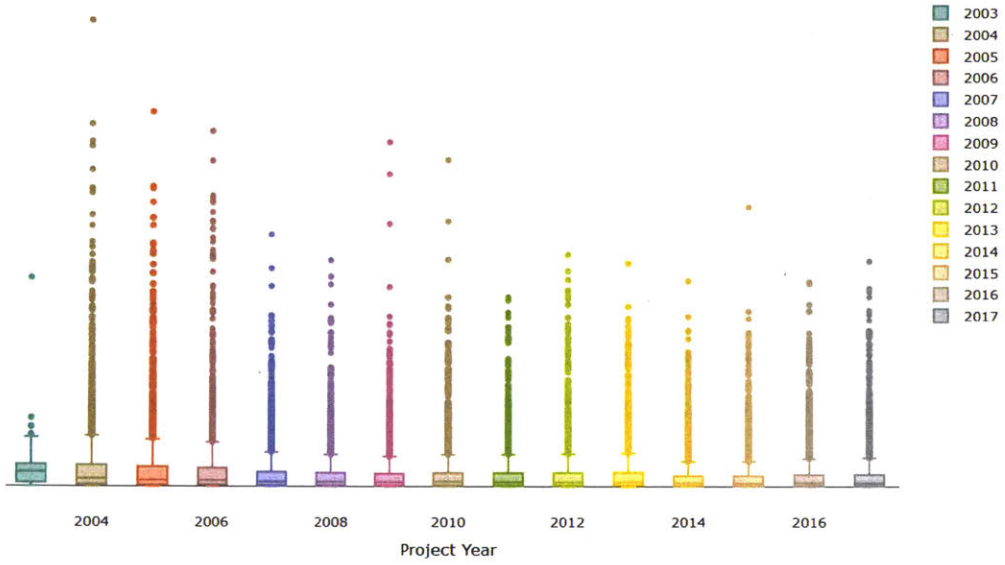

Figure 11: Total Trims Box Plot by Year

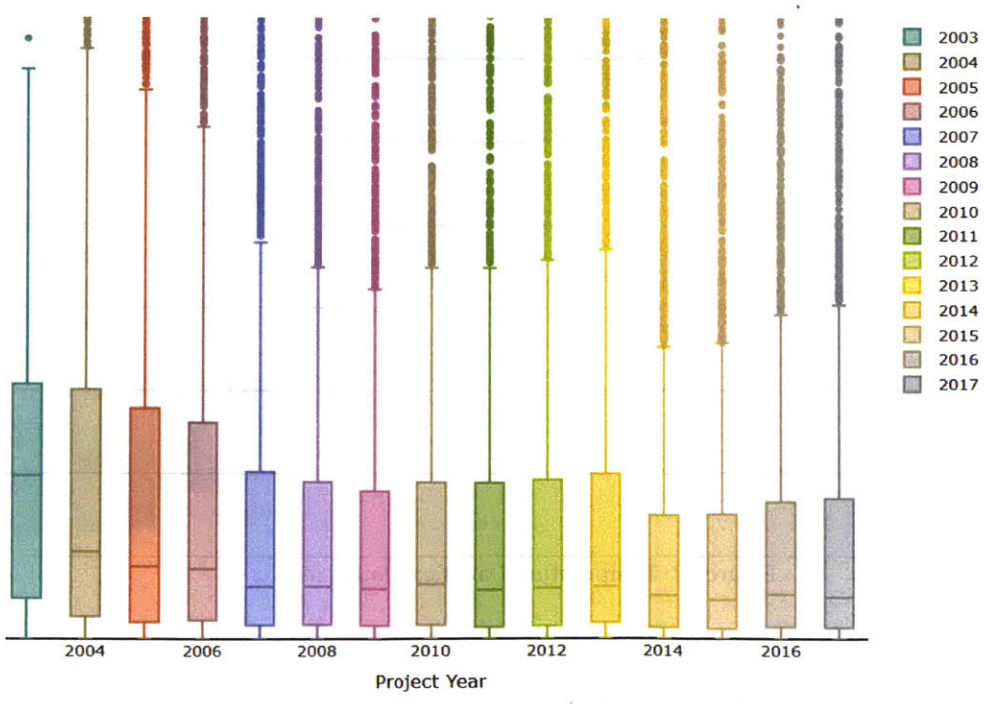

Figure 12: Total Trims Box plot by Year without outliers

3.2 Weather Data

The investigation of the weather parameters contained three immediate groupings of interest: temperature, precipitation, and drought. Within each of these groups there are focuses to analyze any overarching trends and any collinearities that might exist that would impact variable selection.

3.2.1 Temperature Data

The temperature data contains three general metrics: averages (AVG,MIN,MAX), extremes (EMXT,EMNT), and days (DX90,DT32) in which a weather event has occurred (for example number of days above 90 degrees F). As each of these variables are derived from the same physical phenomena it is expected that collinearity will be present, and as such variable selection prior to model fitting will be required.

Figure 13 takes an initial look at all of the average and extreme temperature parameters via a correlation matrix. Many of the variables appear to be highly correlated, specifically weather variables from the Spring and Summer seem to have the highest correlations. This implies that temperatures from April to June are highly related to temperatures from July to September.

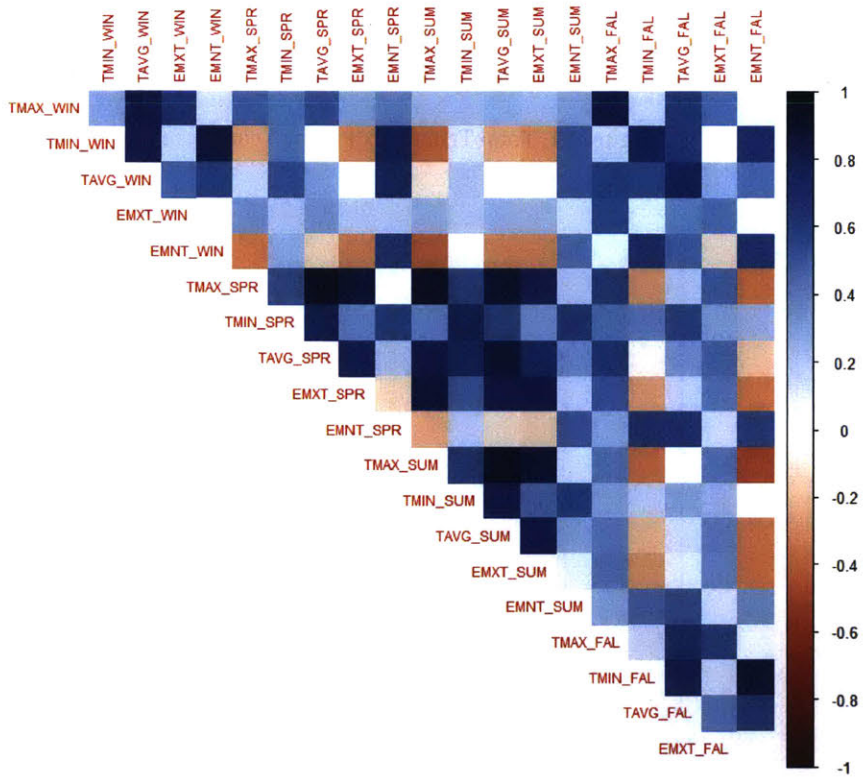

Figure 13: Temperature Data Correlation Matrix[18]

[18] Figure 13 is a correlation matrix, where the boxes colors are related to the correlation value between the two variables from dark red with a correlation of -1 to dark blue with a correlation of +1

Figure 14 explores only the Winter variables, and the same high correlations are seen between the continuous average and extreme variables. Additionally, the days' variables are included and tend to have a lower correlation; however, this is likely an impact of the nature of the variable. For example, in the case of Winter variables, there are probably only a few circuits, if any, that were above 90 degrees F. As a result, it is unlikely that the days variables should be correlated, but for the same reason potentially less insightful for prediction.

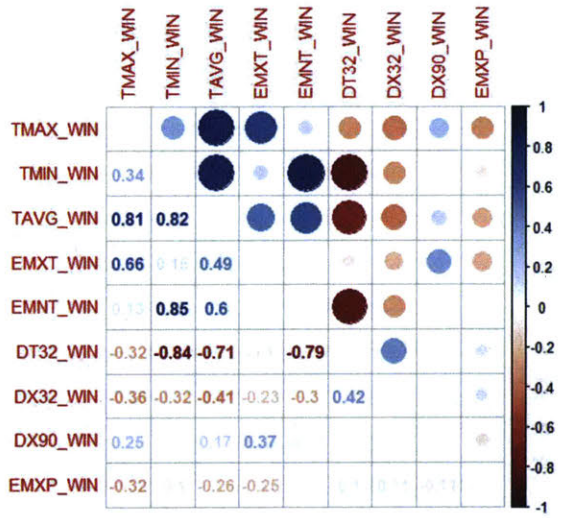

Figure 14: Correlation Plot for Winter Temperature Variable[19]

Boxplots in Figure 15: Boxplot of Minimum Average Temperature by Season and YearFigure 15, Figure 16, and Figure 17, look at few of trends and predominant features of the weather data. Notably in Figure 15 and Figure 16 we observe that many of the box plots for the Spring and Summer seasons appear quite similar, which supports the multiple correlation matrix findings. However, one notable difference between the two different variables in the size of the interquartile range (IQR). In the case of the minimum average variables in Figure 15 the IQR is smaller, especially in the Spring and Summer seasons. Furthermore, Figure 16 does not have as many outliers, once again highlighted in the Spring and Summer seasons. For variable selection this would tend to imply that there is higher variability seen in the Winter and Fall seasons for temperature, which could be helpful for later clustering and modeling predictions. Figure 17 creates a similar boxplot, but for a variable that is preforming a daily count of days with a minimum temperature below 32F within the season. In this variable we notice highly skewed distributions, where in this case Spring and Summer, with exception of one year for the Spring,

[19] Figure 14 is a correlation matrix, where the bottom left is the listed correlation value and colored in the same fashion as Figure 13 (dark red with a correlation of -1 to dark blue with a correlation of +1). The upper right displays the same information where the circles are the size of the absolute value of the correlation (-1 and +1 the largest) and the fill following the same coloring scheme as the bottom left

the IQR range is 0 with a median of 0. Although the shape of distributions are slightly different, the correlations can still be high. Combining with Figure 14 we observe that the DT32_WIN has a -0.84, -0.79, and -0.71 correlation coefficient with the average minimum temperature, the average-average temperature, and the extreme minimum temperature respectfully.

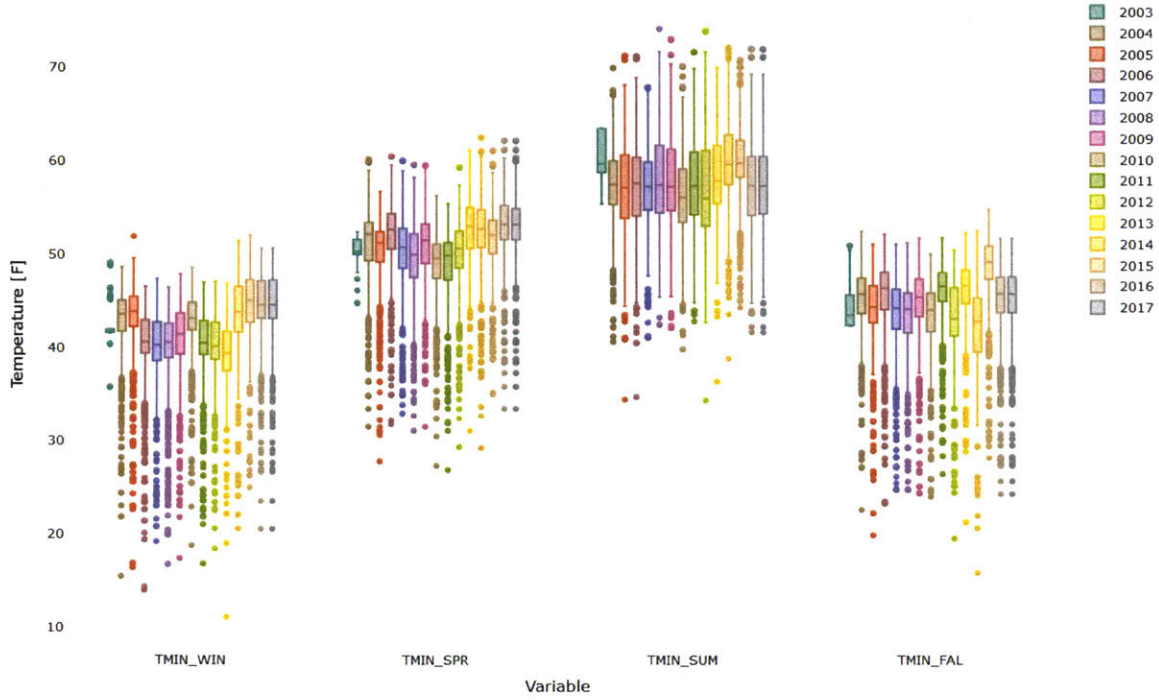

Figure 15: Boxplot of Minimum Average Temperature by Season and Year

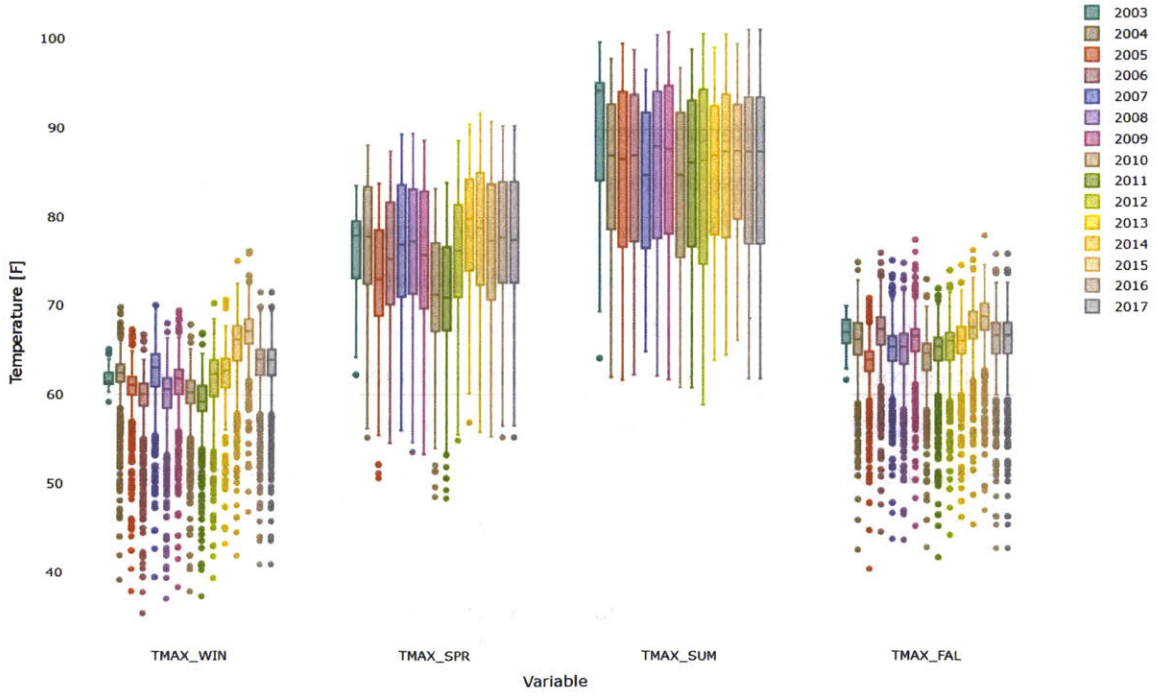

Figure 16: Boxplot of Average-Average Temperature by Season and Year

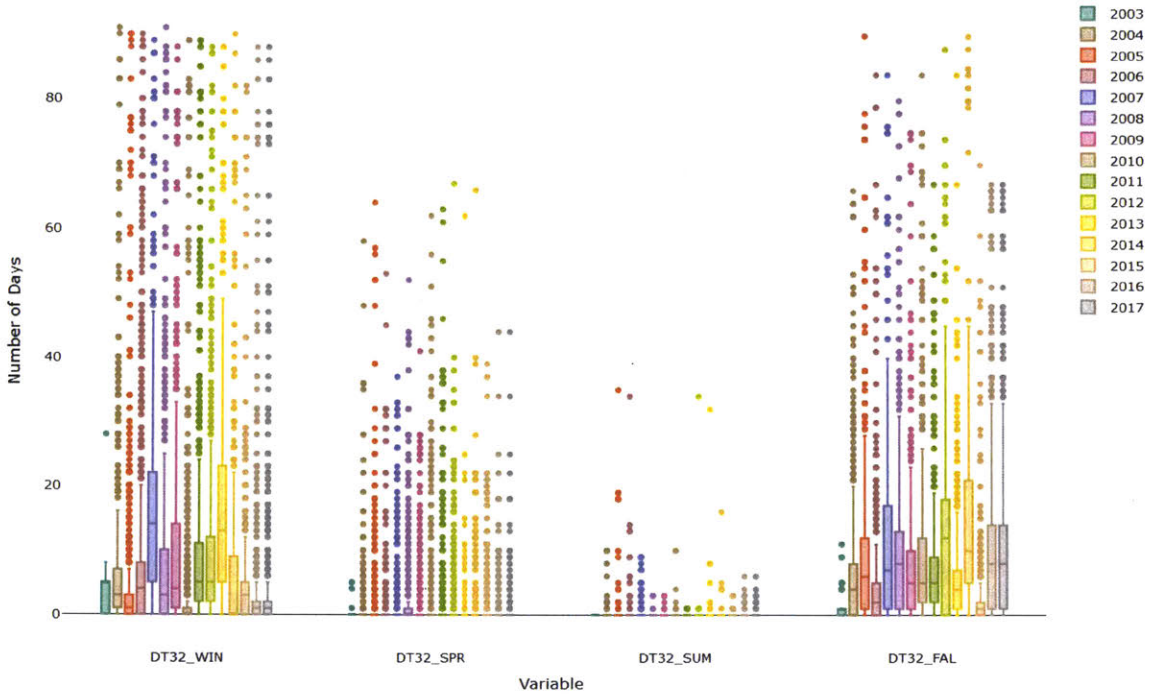

Figure 17: Boxplot of Days Below 32 degrees F by Season and Year

3.2.2 Precipitation Information

The precipitation data, much like the temperature data, contains three groupings of average, extremes, and daily counts. As noted in the temperature section it is expected that collinearity exists within the data. However, unlike the temperature data the precipitation data is potentially more prone to outlier years of high rains, or extended low rain years.

Figure 18 shows a high correlation among of many of the variables, with the only low correlations found with Summer variables related to the Winter and Fall. This finding indicates that only a single family of precipitation data should be evaluated. Figure 19 looks at the total precipitation for the multiple seasons over the 14 years of the data frame. Within it we see that the low correlations for the Summer season in Figure 18 are attributable to considerably low rainfall in that season on average. Interestingly, the Spring season on average has low median precipitation compared to the Fall and Winter seasons. Lastly in Figure 18, the annual trends within a given season have large variability year to year, a feature not seen in the temperature information. For example, in the Winter season precipitation in 2014 and 2016 are well below and outside of the IQR of the all other years.

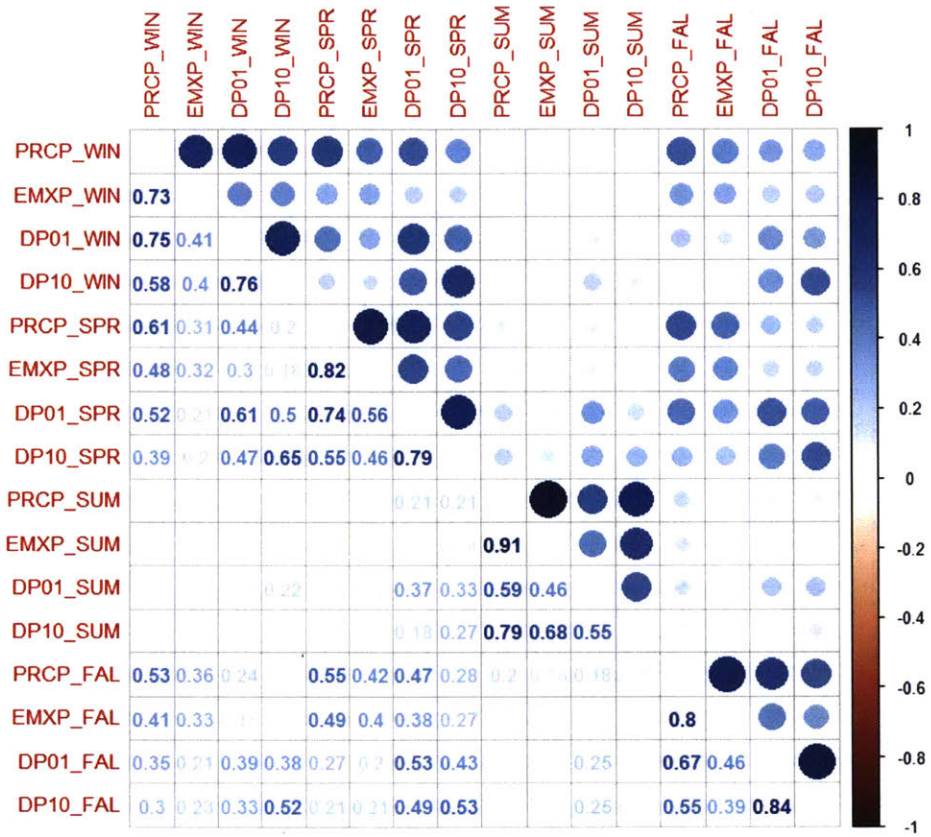

Figure 18: Precipitation Data Correlation Plot

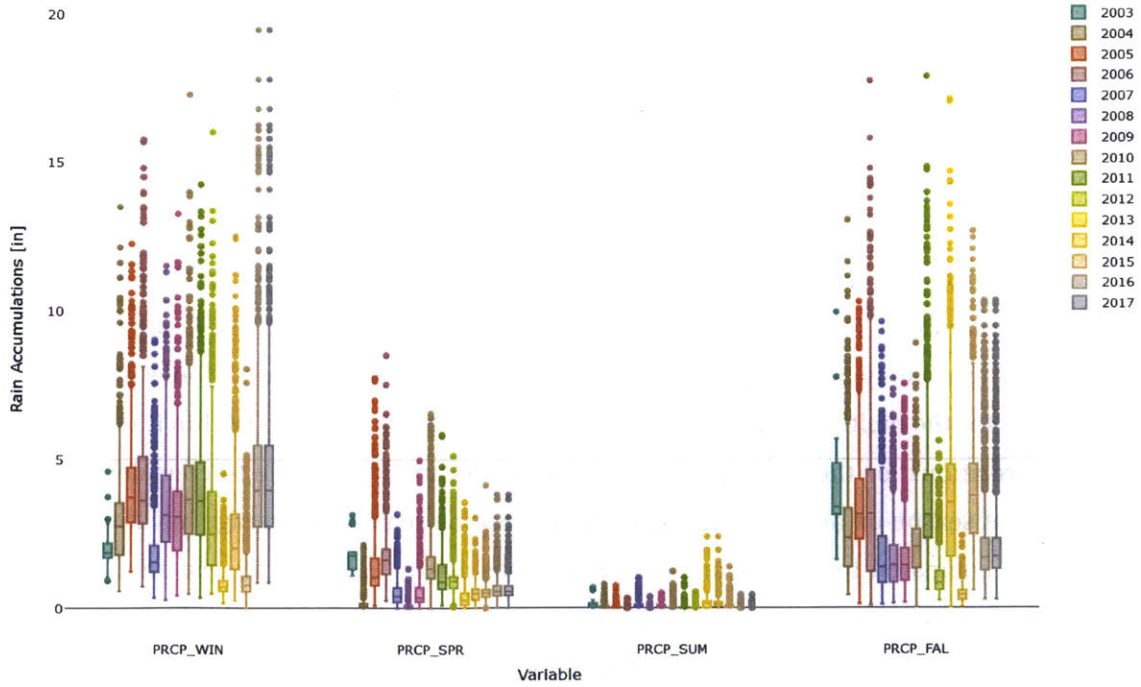

Figure 19: Boxplot of Total Precipitation by Season and Year

3.2.3 Drought Data

Drought data differs from the preceding variables as it is an externally calculated index, as compared to a distinct measurement. Additionally, the drought indexes have an averaging component that accounts for prior weather information to create the index. Consequently a single extreme precipitation event or warming season, although influential in the other parameters, may not be reflected as strongly in this index. The analysis of the drought indices included: any correlations that might exist between the variables, as there are four potential indices over four seasons to select from, and overall trend information, as it might differ slightly from the previous spot measurement data.

Figure 20 provides the correlation matrix for all drought indices investigated, and shows high correlation among all of the variables. This is interesting as even within the same index the seasonal component does not appear to be unique enough to drive a lower correlation. This finding is potentially related to the nature of the index that takes in historical data and not change drastically seasons by season, resulting in variables that are fairly well correlated.

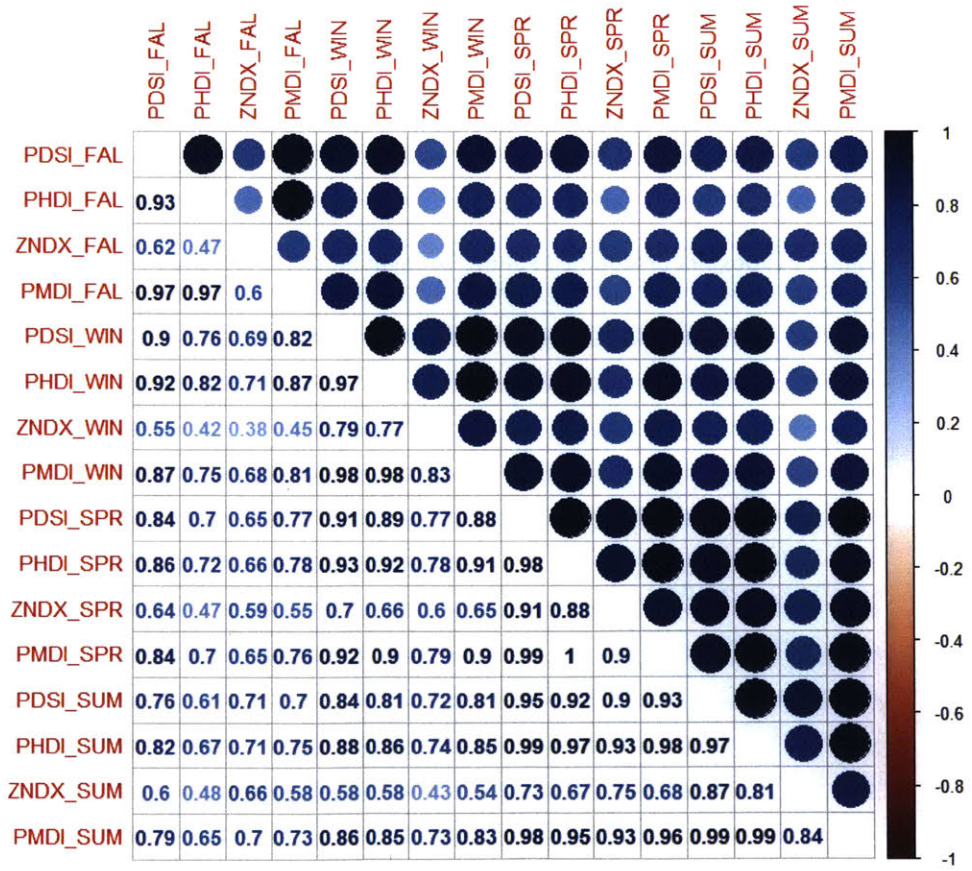

Figure 20: Drought Index Correlation Plot

Figure 21 evaluates the change in Palmer Modified Drought Index over the years of 2003 to 2017. As noted in the precipitation data there were arid conditions in 2014 and 2016, especially in the Winter season. This is seen in the time series of Figure 21where the 2015 and 2016 project years are associated with extreme drought conditions.

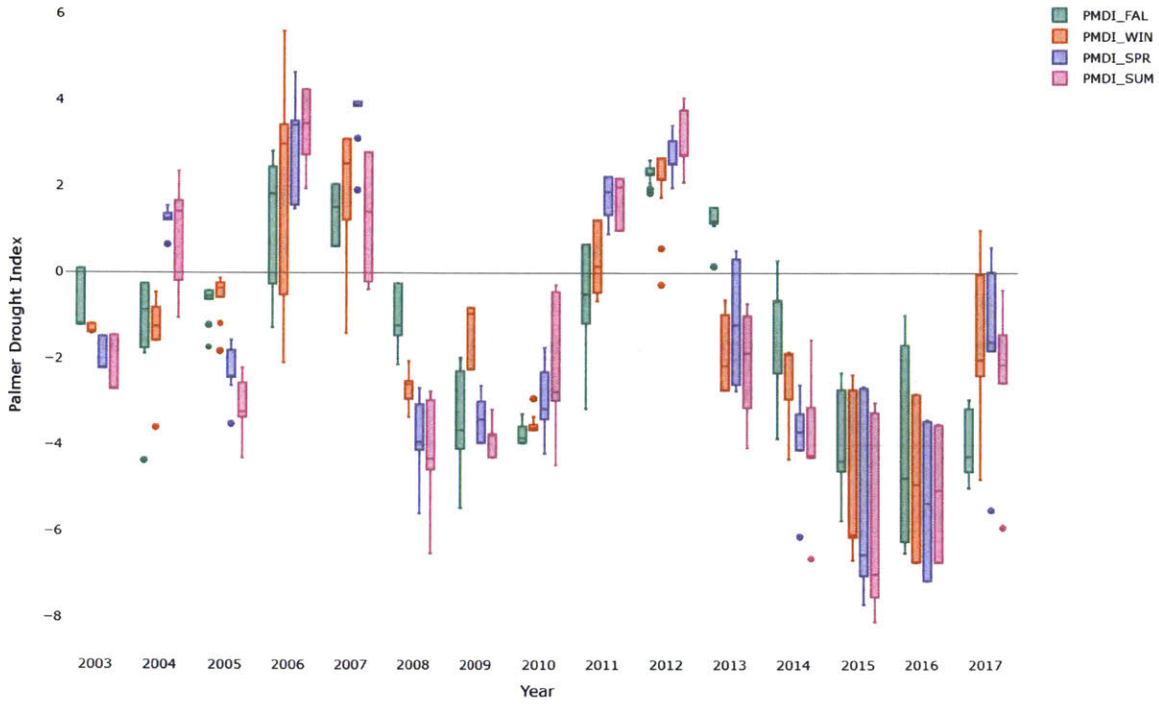

Figure 21: Boxplot of Modified Palmer Drought Index (PMDI) over Years

4 Modeling Approach

4.1 Overview

The approach used to create the final prediction model consisted of five major areas: data frame creation (variable selection), data splitting into test and training, clustering, modeling for each cluster, and accuracy metrics of the model. The flow of these steps is presented in Figure 22.

The procedure for this can be repeated for numerous different variable combinations, data splitting concepts, clustering methods, modeling schemes, and accuracy metrics. This thesis will present a fairly extensive set of these possibilities, but it is by no means exhaustive.

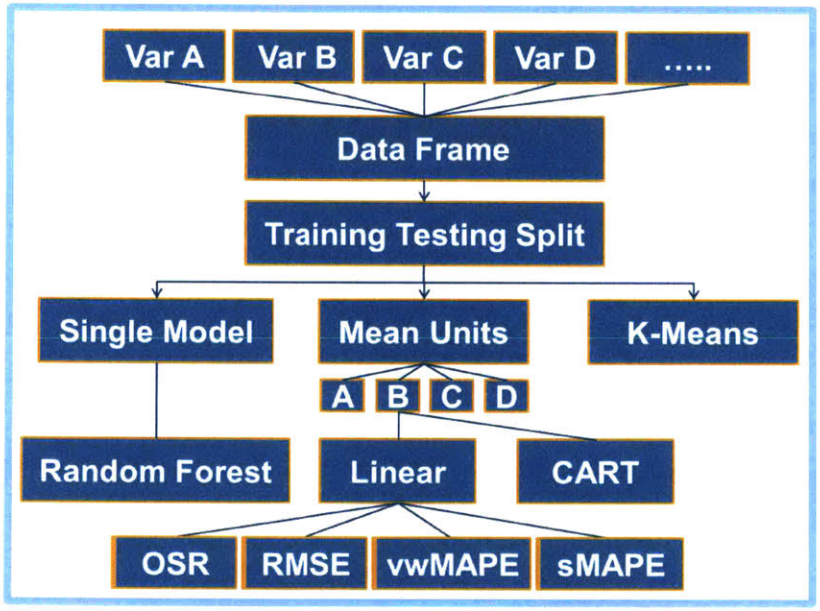

Figure 22: Modeling Approach Flow Chart

4.2 Data Frame and Variable Selections

As noted in Section 3, there are multiple variables that are a highly correlated and variable selection will have to be performed prior to modeling. Additionally, the amount of data from the various sources, such as the species information discussed in Section 2.2.3, is overwhelming and thus needs to be down-selected. In lieu of running every combination of variables through the model, a broader method of selecting an individual set was conducted. The five areas down-selected were: temperature, precipitation, drought, best practices, and PMD species information. The focus of narrowing for temperature and precipitation was on a variable over all seasons; drought was on what season to include for the Modified Palmer Drought Index; and best practice and PMD species was on if those variables should be included in the data frame as a whole. Section 4.7 expands on the exact steps and parameters considered.

4.3 Test and Training Split

To access the predictive capability of the models, the data was split into training and testing sets. The division was made based on sequential years –the earlier data is the training set and the testing is performed on the most recent years. The desire for pursuing this approach is that if the weather in the testing set is abnormal, which from Figure 21 of the drought information appears to indicate, and unseen by the models training set there could be a challenge in predicting. Despite this potential advantage, the method that will be applied throughout the analysis will be via method one, an annual split. The result is a split with a testing set that contains information for project years from 2003 to 2014 of 27,876 circuit years (77.5% of all circuit-years), and a testing set from 2015 to 2017 projects year of 8,084 circuit-years (22.5% of all circuit years).

4.4 Clustering Methods

Clustering was pursued to assess if aggregation assists in the prediction modeling. This is a plausible pursuit as a single model would have to return a value ranging from 0 to the maximum number of trims. Conversely, it is not practical to create a single model for each circuit, which would total in 2,000+ models, as there are certainly some circuits that behave similarly, and as such can increase prediction accuracy. To embrace this, three forms of clustering will be utilized, no clustering, clustering by units and K-Means clustering; as illustrated in Figure 22.

The first method is no clustering, which sets a pseudo-control for the effectiveness of any other clustering method. Second, a clustering based on the number of units from the running average of the circuits, referred to as the units clustering method from here forward. Functionally, this bracketed the expected answers, as well as evaluated if circuit size alone has unique characteristics. Lastly, a K-Mean clustering method was used to group similar circuits. The clustering took all continuous variables that are not trim-related of the given data frame. The trims variables (running average, standard deviation, and lag of removals) were not included as a way of differentiating the K-Means clustering method from the unit method. Instead, clustering included weather parameters, species information, and raw location (Latitude and Longitude). Within the K-Mean clustering the optimal grouping was evaluated, using both minimized within groups sum of squares and outright model accuracy, with results shown in Section 5.3. For both the circuit-year size and K-Mean clustering methods all initial assessments will use a grouping of five clusters.

4.5 Models

For prediction, three general modeling approaches were used: Stepwise Linear Regression, Classification and Regression Tree (CART), and Random Forest. These models were pursed for their overall ease of implementation, as well as, interpretability of the output - as buy-in and

understanding of the models is a key component of the model accelerating technology acceptance. Note, all models were implemented in the R Programming language, specific details of the models are addressed below:

4.5.1 Stepwise Linear Regression

Background - Stepwise linear regression is a unique form of linear regression where variables are systematically added or removed from a base model by increasing the adjusted R^2 of the given model [15], [16]. For this analysis both backward[20], initially including all variables from the input data frame and removing variables to maximize adjusted R^2, and forward[21], starting only with an intercept term and adding variables to maximize adjusted R^2, were used. Following variable selection the Variable Inflation Factor (VIF) is calculated to determine if highly correlated variables were utilized in the final model(s) – these results are reported for the final model in Section 5.5.2.

Advantages – With stepwise linear regression the number of variables introduced into the model are limited. In a qualitative sense the variables that are included in the model indicate a predictive strength; but furthermore, the model will also provide a quantitative measure of significance form the coefficients t-value – the results of which are presented in Section 5.5.1. Lastly, a continuous output is provided, which could prove useful in getting improvements in the model, in contrast to the CART model where a stratified base value will be the output.

Disadvantages – The major disadvantage of the linear model is that the models can't be augmented to provide a prediction that is greater than zero – that is the models can, and likely will, make predictions for negative units. This issue is of greatest concern in the lower circuit-year predictions where predictions are expected to be close to zero. Lastly, due to the potential interactions of the variables determining the change in the output (number of trims) is only correlated, not necessarily caused, by the change of an in an input. For example, it is tempting but not possible to conclude that increased rainfall will linearly change the output based on its coefficient as interactions with temperature and the drought index will also change with rainfall.

4.5.2 CART

Background and Tuning – The CART model will create a decision tree leading a single stratified value. In order to determine the number of branches to include on the decision tree a few additional methods were used. First, a 5-fold cross validation assessed the accuracy of the model [16]. The records in the data set were split into 5 equally-sized groups, 4 of which were used to train the model and the 5th-used to test the model against the actual results. This was performed five times changing the fold used for testing each time until predictions had been made for all

[20] Referred to as *lm-step* in the Section 5 results tables
[21] Referred to as *lm-forward* in the Section 5 results tables

records in the data set. Additionally, within the cross-validation a complexity parameter, cp, was varied from 0.002 to 0.1 in 0.002 steps [16]. The cp value determines the depth of the classification tree – a tree with great depth, and thus complexity, would have enough branches to predict every circuit-year uniquely. However, when cross-validated it would likely be over fit and increase errors. The balance is the best model adds in additional complexity, when it increases accuracy in cross-validation. The CART model selected contains the cp value with the highest cross-validated accuracy based on the vwMAPE accuracy metric, Equation 3 below.

Advantages – In terms of technology adoption, the CART model clearly indicates how predictions were made, and as such would be an advantageous model to pursue. Additionally, it is unlikely that it will predict circuit-year trims less than zero.

Disadvantages – the CART model will be prone to issues of selecting stratified outputs, functionally binning circuits that should not necessarily be group with one another. On average this should be minimized in the model creation; however, the lack of granularity could present challenges in the prediction.

4.5.3 Random Forest

Background and Tuning – Random Forest follows a similar approach to the CART model; however, rather than having a single tree, multiple trees are created and the final prediction is determined from an ensemble of these multiple CART models [16]. In this analysis the number of trees used was 500. As opposed to the CART model, first the random forest for an individual tree will create a training set consisting of a random selection of circuit-years, where some circuit-years will be selected numerous times, while others are not selected at all. Second, in place of the complexity parameter, a random forest will attempt to capture high order effects via the *mtry* parameter. The *mtry* parameter selects the number of variables that can be used at each tree split. To determine the proper selection for *mtry* the value is test of a range of 10 to 20 for this thesis [16]. The last departure from the CART model is cross-validation was not implemented. Instead out-of-bag sampling was used, a process that creates the testing set from the unselected parameters from the initial random selection. Due to the random input selection and deep trees created by the random forest that functionally create multiple clusters as a whole, this method was only applied to the no split clustering approach.

Advantages – The advantage of the Random Forest method is primarily found in the increased prediction strength that is possible, but not guaranteed, due to its natural ensemble of models. The procedure of randomly selecting both circuit-years for input and variables at each split allow for higher order, non-linear effects to be represented in the model. For example, it is possible to have highly influential circuit-years and variables that are omitted for a specific tree, and as such second order variables importance are expressed.

Disadvantages – The ability to interpret the model output is challenging, if not impossible, to precisely discuss. As such, for the application of LiDAR adoption via accurate predictions, if the modeling method lacks clarity it may encounter issues with later acceptance.

4.6 Accuracy Measures

Four accuracy metrics were used to make final model selection: out of sample R^2 (OSR), root mean square error (RMSE), volume weighted mean absolute parentage error (vwMAPE), and symmetric mean absolute percentage error (sMAPE). OSR and RMSE were included as they are fundamental to the variable selection methods used in the models. Additionally forecasting, especially with predictions close to zero, present unique challenges in error metric selection and required the introduction of vwMAPE and sMAPE. The use of these metrics allowed for the percentage errors on larger circuit-years to be weighted more heavily than that of smaller circuit-years, where a larger percentage error would be expected. Below are the equations for the metrics implemented:

$$OSR^2 = 1 - \frac{\sum_{t=1}^{t=N}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{t=N}(y_t - \bar{y})^2}$$

Equation 1: Out-of-sample R^2

$$RMSE = \sqrt{\frac{\sum_{t=1}^{t=N}(y_t - \hat{y}_t)^2}{N}}$$

Equation 2: Root Mean Square Error

$$vwMAPE = \frac{\sum_{t=1}^{t=N}|y_t - \hat{y}_t|}{\sum_{t=1}^{t=N} y_t}$$

Equation 3: Volume Weighted Mean Absolute Percentage Error

$$sMAPE = \frac{\sum_{t=1}^{t=N}|y_t - \hat{y}_t|}{\sum_{t=1}^{t=N}(y_t - \hat{y}_t)}$$

Equation 4: Symmetric Mean Absolute Percentage Error

In all of the equations, t in an index ranging where 1 to N number of circuit-years within the testing set, y_t is the observed number of trims in a given circuit-year, and $\hat{y}_t$ is the model prediction for a given circuit-year. In terms of metric output, OSR^2 is improves with higher values, whereas, RMSE, vwMAPE, and sMAPE are more favourable with lower values. Additionally, the vwMAPE and sMAPE variables are listed and later reported as percentages. For model selection the primary metric, not the sole metric, used in evaluation is vwMAPE. Although a focus was placed on vwMAPE, and to a lesser degree on the sMAPE, later results indicate that the other parameters had similar, if not identical, rank outputs for the models.

4.7 Final Approach

Through all the different method discussed above the final variable and model selection method consisted of seven steps, they are described below. The minimal model included the five-year running average, the project year, and previous year removals. The units clustering method, with five separate clusters, is used for Steps 1 through 5.

1. Starting from the minimal model and only including seasonal total precipitation (PRCP) data. Determine the best seasonal[22] temperature variable to include (TMIN, TMAX, TAVG, EMXT, EXNT, DX90, DT00).
2. Hold the best temperature variable from Step 1. Then select the best seasonal precipitation variable (PRCP, EMXP, DO01, DP10).
3. Holding the results from Step 2, select the best individual season of the Modified Palmer Drought Index (PMDI).
4. Holding the result from Step 3, add in best practices variable. Include variable if it improves accuracy.
5. Holding the results from Step 4, add in species data and other PMD variables. Include variables if they improves accuracy.
6. Holding the results from Step 5, use K-Mean and no clustering and repeat analysis. Additionally for no clustering, create a random forest model. Note the most accurate clustering method.
7. Holding the results from Step 6, access the influence of outliers by removing units that are greater than five standard deviations from the mean. Pursue removing variable if accuracy improves.
8. Holding the results from Step 7, normalize the number of trims prediction variable. Re-scaled the output for accuracy metrics. Select the best model.
9. Holding the results from Step 8, access the best number of clusters for both the K-Mean and units clustering methods.

[22] *Seasonal* implies that the base variable in all four seasons is evaluated. For example, with TMIN

5 Model Results

The modeling results fall into a five general categories: baseline model creation, variable selection steps, clustering method analysis, input data treatments, and variable significance. The results are derived from implementing the Modeling Approach discussed in Section 4. Specifically, the results presented in the following tables will contain the variable types used in the creation of the model, the clustering method and number of clusters, the modeling method used, and the accuracy metrics. Recall, the modeling methods used are linear stepwise regression in the backward, *lm-step* in the results tables, and forward, referred to as *lm-forward* in the results tables, directions, as well as CART.

5.1 Baseline Models

Prior to executing the analysis method from Section 4, two baseline models were selected as comparison: a naïve model, and a running average model. The naïve model simply predicts the current years trim count to be the prior year's trim count, or:

$$y_t = y_{t-1}$$

Equation 5: Naïve Model Formulation

The running average predicts that the current years trim count as the average for x number of prior years, or:

$$y_t = \frac{\sum_{n=1}^{n=x} y_{t-n}}{n}$$

Equation 6: Running Average Model Formulation

Table 1 shows the accuracy measures for both of these models, indicating that the running average is a superior model from all metrics. These models will be used as a baseline of success for additional models. That is, before a more complicated model with additional variables is recommended it must, at a minimum, have superior accuracy.

Table 1: Baseline Model Performance

Model	OSR	RMSE	vwMAPE	sMAPE
Naïve	0.755	290.1	38.9%	19.9%
Mean	0.810	255.7	34.8%	17.1%

5.2 Variable Selection

Variable selection will look to include the most useful predictors from seasonal temperature variables, precipitation variables, drought season, species and WMS variables, and lastly the best practices variable. Each of these steps are taken as a way to either include variables without the issue of high correlations, or access the impact of including the variables.

5.2.1 Temperature Variable Selection

As discussed in Section 3.2.1, the temperature variables in the data frame are highly correlated with one another, and as such, are added independently to the models to access the best variable to select. Table 2 depicts the results for the multiple temperature variables evaluated.

Table 2: Temperature Variable Selection Model Performance

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	23	23	23
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	1	1	1
Units	5	AVG	PRCP	-	-	-	-	-	lm-step	0.785	271.9	36.0%	18.7%	10	15	13
Units	5	AVG	PRCP	-	-	-	-	-	lm-forward	0.779	275.2	36.3%	18.9%	20	21	16
Units	5	AVG	PRCP	-	-	-	-	-	CART	0.790	268.4	35.5%	18.2%	5	5	4
Units	5	EMNT	PRCP	-	-	-	-	-	lm-step	0.780	274.6	36.1%	19.2%	17	17	21
Units	5	EMNT	PRCP	-	-	-	-	-	lm-forward	0.780	274.6	36.1%	19.2%	17	17	21
Units	5	EMNT	PRCP	-	-	-	-	-	CART	0.788	269.8	35.5%	18.3%	8	7	7
Units	5	MAX	PRCP	-	-	-	-	-	lm-step	0.774	278.8	36.4%	19.3%	22	22	22
Units	5	MAX	PRCP	-	-	-	-	-	lm-forward	0.774	278.4	36.2%	19.2%	21	18	19
Units	5	MAX	PRCP	-	-	-	-	-	CART	0.788	269.7	35.5%	18.3%	7	6	6
Units	5	MIN	PRCP	-	-	-	-	-	lm-step	0.795	265.3	35.2%	18.2%	2	2	2
Units	5	MIN	PRCP	-	-	-	-	-	lm-forward	0.795	265.3	35.2%	18.2%	3	3	3
Units	5	MIN	PRCP	-	-	-	-	-	CART	0.792	267.6	35.5%	18.3%	4	8	8
Units	5	DX	PRCP	-	-	-	-	-	lm-step	0.781	274.3	36.3%	18.8%	14	19	14
Units	5	DX	PRCP	-	-	-	-	-	lm-forward	0.781	274.3	36.3%	18.8%	15	20	15
Units	5	DX	PRCP	-	-	-	-	-	CART	0.788	269.9	35.7%	18.3%	9	10	9
Units	5	EMXT	PRCP	-	-	-	-	-	lm-step	0.782	273.3	35.7%	18.6%	13	11	12
Units	5	EMXT	PRCP	-	-	-	-	-	lm-forward	0.783	273.3	35.8%	18.5%	12	12	11
Units	5	EMXT	PRCP	-	-	-	-	-	CART	0.788	269.7	35.4%	18.3%	6	4	5
Units	5	DT	PRCP	-	-	-	-	-	lm-step	0.780	274.6	35.8%	18.9%	19	14	17
Units	5	DT	PRCP	-	-	-	-	-	lm-forward	0.780	274.6	35.8%	18.9%	18	13	18
Units	5	DT	PRCP	-	-	-	-	-	CART	0.784	272.3	35.6%	18.3%	11	9	10

From Table 2, the best performing models use the seasonal TMIN variable, average minimum temperature. Additionally, it is clear that for the TMIN variable there is better model performance found with the linear model method then the CART models. Evaluating only the CART models, we observe similar accuracy statistics for models, which is related to the similarity in the trees created – specifically, it is not uncommon for the only variable used in the tree to be the running average. Practically, this is simply further dividing the clusters, which were already split by number of units to begin with, into smaller groupings without the aid of the additional variables. This signals that the differences in the temperature variables are not as impactful to the CART models. The CART models also show an interesting attribute in that the fourth through ninth best models are all CART models, so even though the linear model preformed best in an individual sense, using many of the other variables resulted in a less accurate model overall. Although the model preforms best with TMIN relative to the other temperature variables available, it does not however out preform the running average method of Equation 6. From these results, the TMIN seasonal temperature variables will be used in all subsequent models.

5.2.2 Precipitation Variable Selection

Table 3 evaluates the various seasonal precipitation variables. The preferred precipitation variable is the gross rainfall variable, PRCP. Interestingly, and similar to the temperature variable selections, the best performing model is one in which the variables are both continuous (not discrete day counts) and aggregate (not single point extremes). This would tend

to indicate that single point weather events are not uniquely critical to trim predictions. Although previous work by [8] has shown extreme weather events to be useful in wires down prediction, the opposite is observed here where sustained weather differences are preferred. From these results, the PRCP seasonal precipitation variables will be used in all subsequent models.

Table 3: Precipitation Variable Selection Model Performance

Input Model Parameters										Accuracy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naive	0.755	290.1	38.9%	19.9%	11	11	9
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	1	1	1
Units	5	MIN	PRCP	-	-	-	-	-	lm-step	0.795	265.3	35.2%	18.2%	2	2	3
Units	5	MIN	PRCP	-	-	-	-	-	lm-forward	0.795	265.3	35.2%	18.2%	3	3	4
Units	5	MIN	PRCP	-	-	-	-	-	CART	0.792	267.6	35.5%	18.3%	5	8	6
Units	5	MIN	EMXP	-	-	-	-	-	lm-step	0.790	268.6	35.4%	18.5%	8	6	8
Units	5	MIN	EMXP	-	-	-	-	-	lm-forward	0.790	268.6	35.4%	18.5%	7	5	7
Units	5	MIN	EMXP	-	-	-	-	-	CART	0.791	267.9	35.5%	18.3%	6	7	5
Units	5	MIN	DP	-	-	-	-	-	lm-step	0.770	281.1	38.5%	19.9%	10	10	11
Units	5	MIN	DP	-	-	-	-	-	lm-forward	0.770	281.0	38.5%	19.9%	9	9	10
Units	5	MIN	DP	-	-	-	-	-	CART	0.792	267.3	35.4%	18.2%	4	4	2

5.2.3 Drought Variable Selection

As noted in the previous two sections, there appears to be increased predictive strength from weather parameters associated with long-term conditions. As such, we anticipate that the addition of the drought variable(s) would provide additional strength. Table 4 displays the model results with all of the seasonal Modified Palmer Drought Indexes.

Table 4: Drought Index Selection Model Performance

Input Model Parameters										Accuracy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naive	0.755	290.1	38.9%	19.9%	20	20	18
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	1	3	1
Units	5	MIN	PRCP	-	-	-	-	-	lm-step	0.795	265.3	35.2%	18.2%	6	9	7
Units	5	MIN	PRCP	-	-	-	-	-	lm-forward	0.795	265.3	35.2%	18.2%	7	10	8
Units	5	MIN	PRCP	-	-	-	-	-	CART	0.792	267.6	35.5%	18.3%	10	15	11
Units	5	MIN	PRCP	ALL	-	-	-	-	lm-step	0.778	276.1	36.7%	20.0%	19	19	20
Units	5	MIN	PRCP	ALL	-	-	-	-	lm-forward	0.778	276.1	36.7%	20.0%	19	19	20
Units	5	MIN	PRCP	ALL	-	-	-	-	CART	0.783	273.2	35.4%	18.4%	17	12	13
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-step	0.802	260.5	34.7%	18.0%	3	2	3
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-forward	0.802	260.5	34.7%	18.0%	2	1	2
Units	5	MIN	PRCP	SPR	-	-	-	-	CART	0.785	271.9	35.0%	18.3%	14	6	9
Units	5	MIN	PRCP	FAL	-	-	-	-	lm-step	0.783	273.1	35.9%	19.3%	16	17	17
Units	5	MIN	PRCP	FAL	-	-	-	-	lm-forward	0.783	273.0	35.9%	19.3%	15	16	16
Units	5	MIN	PRCP	FAL	-	-	-	-	CART	0.789	268.9	35.5%	18.4%	12	14	12
Units	5	MIN	PRCP	SUM	-	-	-	-	lm-step	0.794	266.0	34.9%	18.5%	9	4	14
Units	5	MIN	PRCP	SUM	-	-	-	-	lm-forward	0.794	266.0	35.0%	18.5%	8	5	15
Units	5	MIN	PRCP	SUM	-	-	-	-	CART	0.789	269.0	35.3%	18.2%	13	11	6
Units	5	MIN	PRCP	WIN	-	-	-	-	lm-step	0.798	263.2	35.0%	18.2%	5	7	5
Units	5	MIN	PRCP	WIN	-	-	-	-	lm-forward	0.801	261.3	35.1%	18.1%	4	8	4
Units	5	MIN	PRCP	WIN	-	-	-	-	CART	0.790	268.4	35.5%	18.3%	11	13	10

The introduction of the Drought Index helped nearly all models, regardless of the season. The only approach that clearly did not, were the models that included all drought variables, which is expected as high correlation among all of the variables will hinder higher accuracy. Subsequently, the selection of a single drought variable is required, with the Spring drought variable leading to the best performing model. As compared to the baseline model, the presence of a drought variable produces a model with a lower vwMAPE metric. This result is not entirely expected, as Figure 21 shows that the drought index is most severe in 2015 and 2016, and not encountered in the training set. From these results, the Spring PMDI variable will be used in all subsequent models.

5.2.4 Practices Variable

Table 5[23] presents the results of including the vegetation management expert's variable for updates in annual practices. From Table 5 we observe that the inclusion of this variable is meaningful, with improvements seen in the linear models accuracy that surpass the running average model, and are an added improvement, by 0.8% in the sMAPE metric, from the model selected in Section 5.2.3. This result indicates that expert opinions on the human factors associated with carrying out the tree work are critical for unit count prediction. From these results, the practices variable will be used in all subsequent models.

Table 5: Practices Variable Model Performance

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	11	11	11
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	3	5	1
Units	5	MIN	PRCP	-	-	-	-	-	lm-step	0.795	265.3	35.2%	18.2%	6	7	6
Units	5	MIN	PRCP	-	-	-	-	-	lm-forward	0.795	265.3	35.2%	18.2%	7	8	7
Units	5	MIN	PRCP	-	-	-	-	-	CART	0.792	267.6	35.5%	18.3%	8	10	9
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-step	0.802	260.5	34.7%	18.0%	5	4	5
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-forward	0.802	260.5	34.7%	18.0%	4	3	4
Units	5	MIN	PRCP	SPR	-	-	-	-	CART	0.785	271.9	35.0%	18.3%	9	6	8
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-step	0.810	255.1	34.5%	17.2%	1	2	2
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-forward	0.810	255.1	34.5%	17.2%	2	1	3
Units	5	MIN	PRCP	SPR	X	-	-	-	CART	0.780	274.7	35.4%	18.4%	10	9	10

5.2.5 Species and PMD Variable Selection

The introduction of Species and WMS variables include: species concentrations (details in Section 2.2.3), percentage of highest individual species (details in Section 2.2.3), quarter in which the circuit was inspected and trimmed, and circuit location data via latitude and longitude. Table 6[24] presents the results for including the Species and PMD variables. The inclusion of these variables provides an increase in accuracy; however, to a smaller degree than the previous variables. In the case of the forward linear stepwise regression model, the improvements are on the order of 0.1% for vwMAPE and sMAPE. Despite the small change, from these results the species and PMD variables will be used in all subsequent models as there is a at least some observed improvement.

Table 6: Species and PMD Variable Model Performance

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	14	14	14
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	5	7	1
Units	5	MIN	PRCP	-	-	-	-	-	lm-step	0.795	265.3	35.2%	18.2%	8	9	9
Units	5	MIN	PRCP	-	-	-	-	-	lm-forward	0.795	265.3	35.2%	18.2%	9	10	10
Units	5	MIN	PRCP	-	-	-	-	-	CART	0.792	267.6	35.5%	18.3%	11	13	12
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-step	0.802	260.5	34.7%	18.0%	7	6	7
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-forward	0.802	260.5	34.7%	18.0%	6	5	6
Units	5	MIN	PRCP	SPR	-	-	-	-	CART	0.785	271.9	35.0%	18.3%	12	8	11
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-step	0.810	255.1	34.5%	17.2%	3	3	4
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-forward	0.810	255.1	34.5%	17.2%	4	2	5
Units	5	MIN	PRCP	SPR	X	-	-	-	CART	0.780	274.7	35.4%	18.4%	13	12	13
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-step	0.816	251.3	34.5%	17.1%	1	4	3
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.815	252.1	34.5%	17.1%	2	1	2
Units	5	MIN	PRCP	SPR	X	X	-	-	CART	0.792	267.1	35.3%	18.2%	10	11	8

[23] Previously selected models for Section 5.2.2, and Section 5.2.3 are included for reference. Additionally an "X" denotes that the variable was included in that modeling method

[24] Previously selected models for Section 5.2.2, Section 5.2.3, and Section 5.2.4 are included for reference. Additionally an and an "X" denotes that the variable was included in that modeling method

5.3 Clustering Method Results

Section 5.2 perused a modeling approach with only a single clustering method – five clusters with the unit method. However, there are additional clustering methods that could provide additional insight, namely, no clustering and K-Means –these results are discussed Section 5.3.1. The original model also only investigated a single clustering size, Section 5.3.2 explores the results of altering the cluster size for optimal selection. Finally, Section 5.3.3 investigates the accuracy of the individual models, where all previous errors have been reported on the entire testing set.

5.3.1 Clustering Methods

Table 7 evaluates the effectiveness of the clustering methods, comparing units, K-Means, and no clustering[25]. The results indicate that for the cluster sizes considered, the best model is clustering by units. Interestingly, the K-Means clustering method is outperformed by the single clustering method for a few of the metrics. One potential explanation for this observation is that the K-Mean clustering is in essence multiple single models due to the fact that the K-Mean clustering implemented omits any circuit trim variables. Lastly, the use of the random forest for the single method appears to be a fair predictor from a ranking standpoint; however, the gross accuracy metrics are still below the running average method.

Table 7: Clustering Method Model Performance

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	12	11	11
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	3	3	1
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-step	0.816	251.3	34.5%	17.1%	1	2	3
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.815	252.1	34.5%	17.1%	2	1	2
Units	5	MIN	PRCP	SPR	X	X	-	-	CART	0.792	267.1	35.3%	18.2%	9	4	5
K-Means	5	MIN	PRCP	SPR	X	X	-	-	lm-step	0.805	259.0	38.4%	19.0%	5	9	6
K-Means	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.804	259.7	38.8%	19.1%	6	10	7
K-Means	5	MIN	PRCP	SPR	X	X	-	-	CART	0.782	273.6	37.4%	19.4%	11	8	10
Single	1	MIN	PRCP	SPR	X	X	-	-	lm-step	0.794	266.2	36.6%	19.3%	8	7	9
Single	1	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.794	266.2	36.6%	19.3%	8	7	9
Single	1	MIN	PRCP	SPR	X	X	-	-	CART	0.782	273.5	41.1%	21.2%	10	12	12
Single	1	MIN	PRCP	SPR	X	X	-	-	RF	0.805	258.8	35.3%	17.8%	4	5	4

5.3.2 Clustering Size Verification

In the above results the only clustering methods with five clusters have been considered. To address the validity of this prior method both the K-Means and unit clustering methods were cycled over cluster sizes from 3 to 20 clusters. Figure 23 shows the model selected in Section 5.2.3 using K-Means over the range of cluster sizes. The RMSE error plot on the left does not appear to have any meaningful trend information, whereas the vwMAPE plot appears to show that smaller clusters are preferred, with a minimum for the linear models at about five clusters. Figure 24 repeats this method with the final model tested, from Section 5.4.2, using the units clustering method. The plots on the left report both stepwise linear models and the CART model results, while the plots on the right exclude the CART model, and focus the y-axis to access any closer

[25] No clustering is referred to as *Single* in the results tables

potential trends in the vwMAPE and sMAPE metrics. The focused plots on the right of Figure 24 are most strongly characterized by fairly sizable variance in the results, with a slight upward trend and a minimum around five clusters. There are low accuracy metrics around 10 clusters for the vwMAPE metric and 16 clusters for RMSE metric, with no coherent explanation or trend for added accuracy. Although there are not any concrete recommendation from these figures, it would appear that the assumption of five clusters in the previous models is adequate for modeling, with only marginal gains seen by updating to another cluster size.

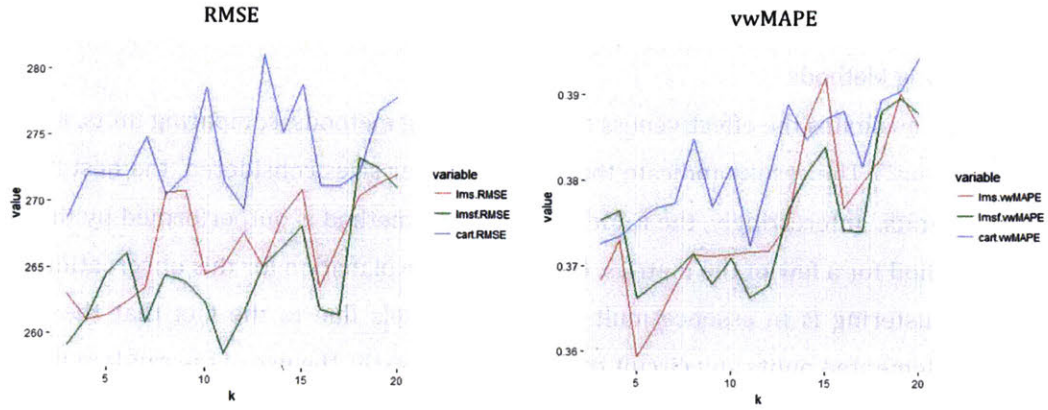

Figure 23: K-Mean Clustering Size Accuracy Plots

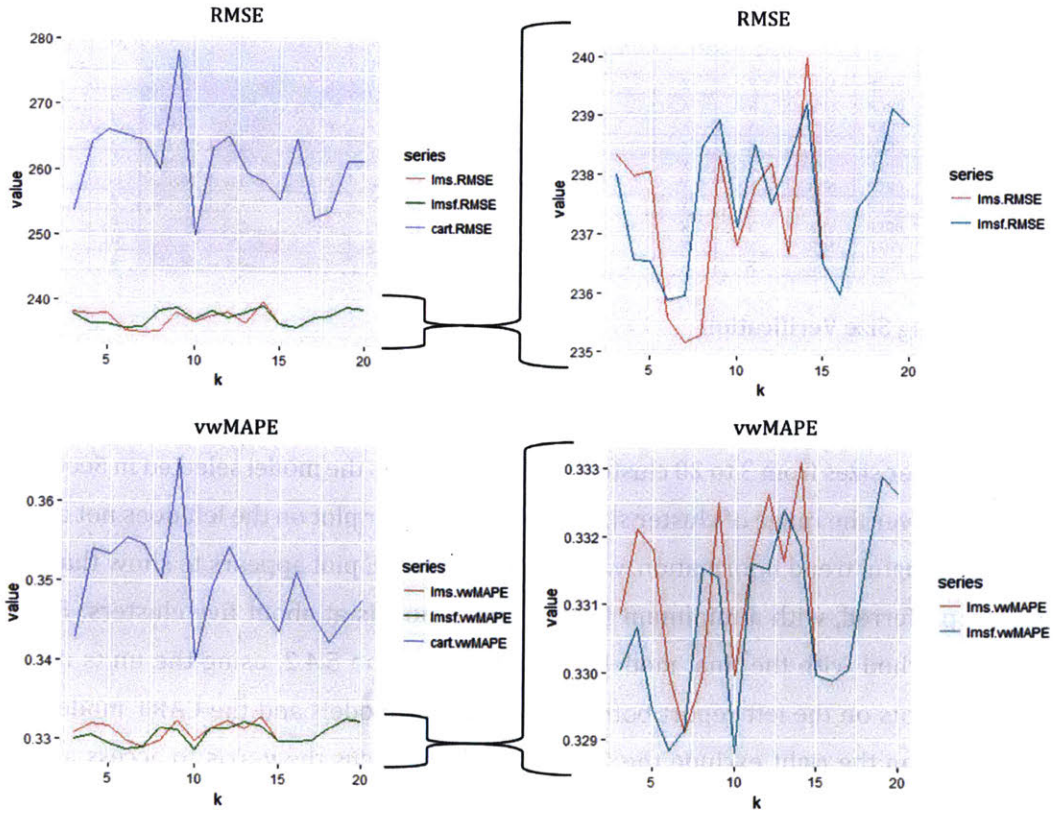

Figure 24: Units Clustering Size Accuracy Plots

5.3.3 Inter-Model Errors

The final consideration for the clustering methods is to investigate the accuracy of individual cluster models. The model evaluated is that with the highest accuracy model from Table 7. It is comprised of: five cluster based on unit size, all potential variables included, a forward stepwise linear regression model, and no consideration of outliers and/or transformations. The results for the individual models are shown in Table 8, where the individual models are also compared against the running average method for the same data. Note, for the individual clusters results, the *Number* column depicts the cluster number. Where the clusters with a lower cluster number are associated with a lower unit count; that is, Cluster 1 will have the smallest number of trims, and Cluster 5 will have the largest number of trims. The upper three models are for the entire testing set and are included for general reference.

Table 8: Individual Cluster Model Performance

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	-	-	-
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	-	-	-
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.815	252.1	34.5%	17.1%	-	-	-
Units	1	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.059	32.9	57.0%	30.8%	2	10	10
-	1	-	-	-	-	-	-	-	Mean	0.065	32.8	55.9%	29.6%	1	9	9
Units	2	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.162	44.3	46.9%	23.5%	4	7	8
-	2	-	-	-	-	-	-	-	Mean	0.174	44.0	47.0%	23.0%	3	8	7
Units	3	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.132	101.2	40.1%	19.3%	5	5	5
-	3	-	-	-	-	-	-	-	Mean	0.118	102.0	40.3%	19.5%	6	6	6
Units	4	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.248	171.7	32.6%	16.2%	7	1	2
-	4	-	-	-	-	-	-	-	Mean	0.212	175.7	33.6%	16.5%	8	4	4
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.591	626.6	32.8%	16.4%	9	2	3
-	5	-	-	-	-	-	-	-	Mean	0.591	626.7	32.9%	16.2%	10	3	1

The first observation, is that the accuracy of the models tends to increase with the higher circuit count, for all metrics besides RMSE – a slightly expected result as the RMSE is scale dependent. However, when looking at the smaller unit clusters the OSR^2 is very low, at about 0.059. Figure 25 expands on this observation by showing that the gross errors are more tightly packed for clusters with less units.

Figure 25: Scatter Plot of Model Error by Unit Cluster

Secondly, when comparing the individual models to the running average model, the results for lower unit circuit clusters appear to favor the running average whereas the larger circuit clusters appear to favor the linear model. This trend supports the concept that with larger circuit counts there is a possibility of prediction by additional variables outside of the basic mean. Additionally, for the cumulative models, accuracy metrics tend to benefit from this facet as there is added prediction accuracy for the variables that have higher weighting, most notably in the case of the vwMAPE metric. In short, the analysis of individual cluster models shows that the lower the unit count, the better it is to predict with a simple mean, where larger circuits benefit from the additional variables included in the model. No action was taken, such as using the running average for Cluster 1 and 2, in order to mitigate the lower accuracy of the linear models.

5.4 Input Data Treatments

The above modeling approaches have not yet addressed additional treatments to the data, with specific focus on outliers and variable normalization methods. In Section 5.4.1 the results of implementing an outlier method will be presented and in Section 5.4.2, with the outlier removed data frame, the results of a normalization scheme will be discussed.

5.4.1 Outlier Refinement

From Figure 25, we observe a few circuit-years with exceedingly high error; in an effort to mitigate the impact of these errors on model creation, an outlier removal method was pursued. Although statistical methods could be applied for variable removal, such as Cook's Distance, a more general approach was taken in removing all circuit-years that were over five standard deviations (using the five point running average and standard deviation) from the mean. The rationale behind this method is that any circuit-year that was above this value was caused by a notable human factor that the model could not take into account. Examples include: increasing a circuit length, reallocating circuits to a new project, and a large addition of trees to a specific parcel.

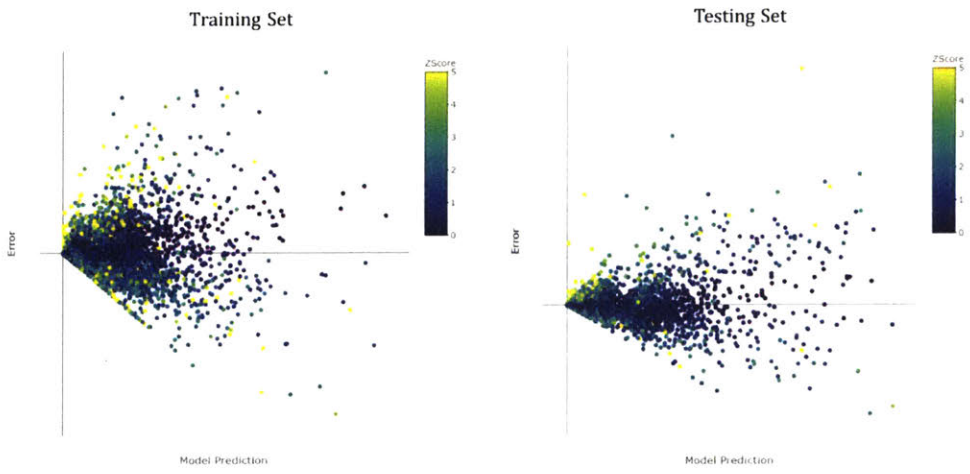

Figure 26: Scatter Plot of Model Error by Z-Score for Training and Testing Datasets

The result of removing circuit-years above five standard deviations was a decrease of 1131 circuit-years or 3.24% of the total data frame. Of these removals, 64 were due to a standard deviation of 0. Although this could be an accurate metric for low count circuits, especially those that have had multiple years of 0 units, no specific effort was made to reintroduce these circuit-years. This high outlier removal percentage is unexpected as a standard deviation event should occur far less than with number of circuit year's remove. Figure 26 shows the error signals observed versus the model predictions and colored by the Z-Score, all purely yellow (5 and above) circuit-years were removed. Table 9 displays the accuracy metrics for this updated data frame.

Table 9: Outlier Removal Model Performance

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	13	13	13
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	8	7	4
-	-	-	-	-	-	-	5	-	Naïve	0.764	278.7	38.2%	19.5%	12	11	12
-	-	-	-	-	-	-	5	-	Mean	0.818	244.9	34.0%	16.7%	3	3	3
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-step	0.816	251.3	34.5%	17.1%	6	6	6
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.815	252.1	34.5%	17.1%	7	5	5
Units	5	MIN	PRCP	SPR	X	X	-	-	CART	0.792	267.1	35.3%	18.2%	11	8	8
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-step	0.828	238.3	33.7%	16.6%	1	2	2
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-forward	0.827	239.1	33.6%	16.6%	2	1	1
Units	5	MIN	PRCP	SPR	X	X	5	-	CART	0.801	256.4	34.4%	17.9%	9	4	7
K-Means	5	MIN	PRCP	SPR	X	X	5	-	lm-step	0.818	245.2	37.8%	18.9%	4	10	9
K-Means	5	MIN	PRCP	SPR	X	X	5	-	lm-forward	0.815	247.0	38.5%	19.1%	5	12	10
K-Means	5	MIN	PRCP	SPR	X	X	5	-	CART	0.795	259.8	37.3%	19.3%	10	9	11

With the removal of variables, the base models, both naïve and running average, had to be readdressed. Table 9 indicates that the removal of outliers for the data frame helps in a relatively substantial manner, with the vwMAPE metric decreasing from 34.5% to 33.6% for the forward stepwise linear regression model. However the baseline model also improved, decreasing vwMAPE by 0.8%; thus, the overall gain in this method is not as immediately impactful. Additionally, assessed in Table 9, is that the removal of outliers did not impact the effectiveness of the K-Means model. From these results, the outlier approach will be used in all subsequent models.

5.4.2 Normalization and Final Model

Normalizing the circuit-years by project year allows for the potential to transform the trims data as an evaluation against the average. After the previous results shown in Figure 10, a normalization was performed only on the total trims data to access the validity of this observation. With this standardization, the models no longer included the running average, as it was functionally accounted for in the normalization routines, and prior year's removals were scaled by running average, to avoid scaling the prediction. As a result, the modeling effort predicted if an above or below average number of units were going to be worked that year. The final results were rescaled back into trims for continuity with in the accuracy metrics for the other models. Table 10 displays the results for the transformation. K-Means and the no clustering method were also evaluated for completeness.

Table 10: Normalization Model Performance

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	5	-	Naïve	0.764	278.7	38.2%	19.5%	18	18	18
-	-	-	-	-	-	-	5	-	Mean	0.818	244.9	34.0%	16.7%	10	11	9
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-step	0.816	251.3	34.5%	17.1%	12	14	13
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.815	252.1	34.5%	17.1%	13	13	12
Units	5	MIN	PRCP	SPR	X	X	-	-	CART	0.792	267.1	35.3%	18.2%	16	15	16
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-step	0.828	238.3	33.7%	16.6%	3	7	8
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-forward	0.827	239.1	33.6%	16.6%	4	6	7
Units	5	MIN	PRCP	SPR	X	X	5	-	CART	0.801	256.4	34.4%	17.9%	14	12	15
Units	5	MIN	PRCP	SPR	X	X	5	x	lm-step	0.828	238.1	33.2%	16.4%	2	2	4
Units	5	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.830	236.5	32.9%	16.4%	1	1	3
Units	5	MIN	PRCP	SPR	X	X	5	x	CART	0.785	266.1	35.3%	18.5%	15	16	17
K-Means	5	MIN	PRCP	SPR	X	X	5	x	lm-step	0.820	244.0	33.9%	16.3%	7	10	1
K-Means	5	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.820	244.0	33.9%	16.3%	8	9	2
K-Means	5	MIN	PRCP	SPR	X	X	5	x	CART	0.780	269.5	36.2%	17.7%	17	17	14
Single	1	MIN	PRCP	SPR	X	X	5	x	lm-step	0.821	243.1	33.6%	16.4%	6	4	6
Single	1	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.821	243.1	33.6%	16.4%	6	4	6
Single	1	MIN	PRCP	SPR	X	X	5	x	CART	0.819	244.0	33.2%	17.0%	9	3	10
Single	1	MIN	PRCP	SPR	X	X	5	x	RF	0.817	245.6	33.7%	17.1%	11	8	11

There is a clear advantage to normalization, with the forward stepwise linear regression model improving vwMAPE from 33.6% (best result from Table 9) to 32.9%. The 0.7% improvement, relates to a 1.1% lower percentage error compared to the baseline model. One interesting observation found in Table 10 is that the performance of the no clustering method is high, with the third best accuracy metrics coming from a no clustering CART model. Furthermore, the linear models perform well, with identical forward and backward models resulting in a tie for fourth, and even performing better than the random forest model. Due to these results, the normalization approach will be applied. The final model that returns the best accuracy metrics, is one that includes:

- Temperature data using the seasonal average minimum temperature
- Precipitation data using the seasonal gross precipitation
- The Spring Modified Palmer Drought Index
- Vegetation Management practices input variable
- All species and WMS data
- Removing circuit-years with greater than five standard deviations
- Normalization
- Clustering by the five year running average into five clusters
- Modeling using a forward stepwise linear regression

5.5 Variable Significance

The primary use of this model is for future LiDAR implementations, and a significant component of that acceptance is an understanding of the prediction model as whole. Using the final model from Section 5.4.2, individual variable significance was determined for the linear model using two methods: standardized coefficients and variable inflation factors. While this analysis will indicate which variables are significant for prediction, it is important to note that their presence in the models does not imply that they are the cause for the number of units trimmed to change. As highlighted previously, a data frame containing relatively high

correlations, in this case with weather variables, leads to loose casual effects. Nevertheless with those caveats known, ranking of the variables significance can provide insight to future users understanding and later acceptance of the models.

5.5.1 Linear Model Variables

Table 11 contains the standardized coefficient for the variables present[26] in each of the models. The standardized coefficient is the normalized Student's t-score for each variable[27] in the model with higher values indicating that they are more significant. The values in Table 11 are absolute values, and are colored by magnitude with green being the most significant.

Table 11: Linear Model Cluster Standardized Coefficients

	Cluster				
	1	2	3	4	5
Practices	8.93	10.65	11.11	10.73	15.58
iProjectYr	6.43	9.14	10.46	9.65	12.68
(Intercept)	6.46	9.15	10.07	9.71	12.78
PMDI_SPR	-	1.96	2.27	2.00	5.76
TMIN_WIN	-	2.15	3.65	2.09	3.65
PRCP_SUM	2.67	2.98	-	-	5.39
PRCP_WIN	4.32	3.71	-	-	2.49
PRCP_FAL	3.06	-	2.18	2.68	1.67
TMIN_SPR	-	-	3.49	2.35	2.53
TMIN_FAL	1.55	3.53	1.43	-	-
LagR1	-	-	1.45	2.37	1.97
Prec1	3.12	-	-	1.68	-
MF_Per	-	-	-	2.63	1.69
S_Per	-	-	-	-	4.00
LON	-	-	3.81	-	-
LAT	-	-	3.62	-	-
UNK_Per	1.53	-	-	-	1.57
PRCP_SPR	-	-	2.45	-	-
Qtr	-	-	-	-	2.16
Prec2	-	-	-	-	2.01
FSF_Per	-	-	-	1.74	-

From Table 11 three general tiers of variables are observed: highly significant variables present in all models, moderately significant variables present in many models, and marginally significant variable present in only a few models. The most significant variables are the Practice and Project Year. The Practices variable presence with such high significance implies that expert input and updates made to the process are meaningful predictors. Additionally, the Project Year variable insinuates that an annual trend is present, which agrees with the original observations from Figure 10, 11 and 12. The majority of the next tier of variables are weather variables with the most meaningful being Drought, a sign that long term variables are preferred for trim predictions. Next in this tier is Winter Temperature, and Precipitation for Summer, Winter and Fall. The combination of this variable shows that precipitation tends to more critical than temperature; however, in referencing Figure 15 and Figure 19 this could potentially be attributed to the higher variance seen in the precipitation data. Lastly, the bottom tier variables exclusively contained PMD-related variables. Interestingly, some of these variables have T-Score below a 95%

[26] "-" imply the variable was not included in the model

[27] The null hypothesis tested with is that the coefficient is 0. The t-score is an indication of if the coefficient is statically different than 0, as the errors within the model will create a statistical range for each coefficient.

confidence interval (what would customarily be considered for variable selection); however, due to the variable selection method in the stepwise regression they were systematical included in the model due to their maximization of the adjusted R^2 [15].

5.5.2 Variable Inflation Factors

Despite variables being significant it is prudent to check the impact that correlation with others variables in the model. One analytical method of conducting this is with the Variable Inflation Factor (VIF), a method where each variable is removed from the model and change in variable coefficients is assessed. A high VIF metric entails that it related with another variable in the model and should be considered for removal, it is certainly recommended to remove variables with a VIF greater than 10 [17]; however, [16] recommends a VIF as low as 5. Table 12 presents the VIF for all models.

Table 12: Linear Model Variable Inflation Factor (VIF)

	Cluster				
	1	2	3	4	5
Practices	3.57	3.29	3.92	3.87	3.70
iProjectYr	3.39	3.14	3.56	3.54	3.29
PMDI_SPR	-	1.30	1.41	1.16	1.18
TMIN_WIN	-	2.79	3.63	1.37	1.55
PRCP_SUM	1.18	1.09	-	-	1.17
PRCP_WIN	1.42	1.09	-	-	1.75
PRCP_FAL	1.74	-	2.62	1.39	1.93
TMIN_SPR	-	-	3.57	1.44	1.56
TMIN_FAL	1.14	2.78	3.17	-	-
LagR1	-	-	1.11	1.02	1.17
Prec1	1.01	-	-	1.13	-
MF_Per	-	-	-	1.24	1.49
S_Per	-	-	-	-	1.39
LON	-	-	6.00	-	-
LAT	-	-	4.46	-	-
UNK_Per	1.02	-	-	-	1.07
PRCP_SPR	-	-	2.11	-	-
Qtr	-	-	-	-	1.01
Prec2	-	-	-	-	1.08
FSF_Per	-	-	-	1.08	-

Within Table 12 it is clear that the Practice and Project Year variables have higher VIFs. These variables by their nature (process improvements were changed on an annual basis) are correlated and as such have high VIFs. However, when one of the variables is removed (models ran independently and results not reported) the prediction accuracy decreases, as to be expected with the high coefficients in Table 11. As such, both are included in the final model. Moreover, the highest VIFs are related to LAT and LON, circuit physical location parameters. These variables are likely high VIF variables due to their correlation with the more significant weather variables; that is, the location of a circuit is probably correlated with a specific weather region. Overall the results shown in Table 12 do not provide any major cause for concern that correlation among the variables is impacting model accuracy.

6 Conclusions and Improvements

6.1 Recommendations

The process utilized in this thesis created a statistical model that better predicted tree trims than both the naïve and running average models. The models all included additional variables and were compared with a variety of error metrics. On the surface, this supports the hypothesis of Section 1.5; however, the implementation on this model could defeat the underlying purpose of the thesis: LiDAR adoption. Despite the increased accuracy, the approximate 1% gain may not overcome the ease and overall general accuracy of the running average. As for LiDAR adoption and the tangential use case of annual planning in vegetation management programs, it is recommended that the far similar running average model be implemented in the short term. An implementation of the running average will be intuitive for all parties, and will garnish roughly the same accuracy. This approach could limit the time dedicated to planning and unit tracking within the utility and could free up time for additional challenges in the VM program. The analytical model should continue to be updated through a selection of the improvements and future work discussed below, as a more in-depth analysis into variability and a systemic acceptance of LiDAR detections is still necessary for VM programs going forward.

6.2 General Findings

In the previously discussed result and data frame creation sections, multiple observations were made. Below is a condensed version of the findings for VM programs that might consider embarking on a statistical model for their service territory:

- Continuous weather variables are superior to discrete measurements; for example, average minimum temperature is a better predictor than the number of days a temperature condition obtained.
- Average and gross parameters were better predictors than extreme values, which suggest that tree growth is more related to longer term weather patterns. This is further supported by the high significance of the drought variable. The implication of this is that individual weather events should not be considered on their own impactful to unit prediction. For example, a single monthly extreme weather event should not be assumed to directly impact unit count. Instead, a longer term view should be considered.
- Additional variables, such as the species information was of limited usefulness in the models. Even though the model accuracy was increased with their inclusion the challenges in data cleanliness associated with creating them are not immediately clear to have long term prediction benefit.

- The most useful predictor is associated with overall program process improvements. Unit prediction is susceptible to factors external to the environmental weather conditions, and are instead largely determined by human related processes within the VM program.
- The running average method, five year running average of units, is a fairly accurate model. It would stand to reason that it could be improved upon easily by making manual modifications that include known large deviations that would be encountered in a given year. On the other hand, using the naïve model is inferior, and should not be applied. In fact, from this analysis, it is a prudent measure for VM programs to focus on circuit unit counts exclusively in an average sense as compared to a prior year.
- In this study clustering created generally better performance. Tactically, clustering by units returned consistently the best results, and five clusters presented a fair starting point for analysis.
- Typical forecasting will recommend a weighted or symmetric error value such as variants of the vwMAPE or sMAPE; however, all error statistics used in this work returned very similar rank order results.

6.3 Future Work

The results presented in this project are far from exhaustive in respect to the statistical model. In this area a few improvements that would be recommended are:

- Include weather information from earlier years. In this model only a single year was used. The presence of average weather terms and the drought variable is a signal that longer term weather conditions could be critical for prediction.
- Weather data should be added based on a climate condition with highly reliable weather stations. The current method took all weather stations from NOAA and matched based upon the center of the circuit. Instead, it would be reasonable to include weather stations based on the weather conditions seen as an aggregate on that circuit; that is, apply more than a single weather station per circuit and ensure the weather station(s) selected accuracy represents the weather of that circuit. This is especially important for circuits that are in mountain regions that could encounter critical microclimates unseen by the closest weather station. Furthermore, longer circuits have the possibility of encountering wide degrees of weather and this should be considered as well.
- The drought data applied was geographically course, as seen in Figure 4. Instead it would be worthwhile to apply a more granular drought data set, for example applying the US Drought Survey graphical data that is on gird-based system. Additionally, the use of the Vegetation Drought Index, currently under refinement, would be an interesting variable

to include in the model. Finally, variables associated with land coverage, such as urban, suburban, rural, forest type, etc. could be useful as well.

- The focus on this project was at the circuit level; however, creating models at the division and system level utilizing a model ensemble method might provide higher accuracy. These models could also prove useful for overall planning efforts.
- With the high significance of the practice variable, it would be worth considering more human-related variables that were discussed but ultimately not included in this analysis. Additional variables could include: who conducted patrol and/or tree work and original planning estimates.
- Only four basic models were considered. Expanding to additional modeling methods (such as, neural networks, and learning machines) would be a prudent task to undertake.

Lastly with an eye to even later trends in VM programs, the growth of data and analytics in business over recent years indicates that statistical modeling will invariably be intertwined with the use of new technologies, such as LiDAR. The granularity of data provided by LiDAR, will take a natural next step to conduct analysis at the individual tree level. To support these LiDAR-based models, insights and methods obtained for higher level models, such as the one in this thesis, should prove useful for quickly and accurately deploying the new models. With these tree level models, current processes used in conducting utility VM will be capable of shifting in a way to improve the safety, quality, and timeliness of VM work.

Appendix A: List of Acronyms

Abbreviation	Meaning
CART	Classification and Regression Tree
DP01	Number of days with >= 0.01 inch
DP10	Number of days in month with >= 1.00 inch
DT32	Number of days with minimum temperature <= 32 degrees Fahrenheit
DX90	Number of days with maximum temperature >= 90 degrees Fahrenheit
EMNT	Extreme minimum temperature
EMXP	Highest daily total of precipitation
EMXT	Extreme maximum temperature
F	Fast (tree growth)
FAL	Fall: October to December
IQR	Inter Quartile Range
LiDAR	Light Detection and Ranging
M	Medium (tree growth)
MF	Medium Fast (tree growth)
NOAA	National Oceanic and Atmospheric Administration
OSR	Out of Sample R2
PMD	Project Management Database
PMDI	Palmer Modified Drought Index
PRCP	Total Precipitation
RMSE	Root Mean Square Error
S	Slow (tree growth)
SF	Super Fast (tree growth)
SFS	(Fast-Super Fast (tree growth)
SM	Slow Medium (tree growth)
sMAPE	Symmetric Mean Absolute Percentage Error
SPR	Spring: April to June
SUM	Summer: July to September
TAVG	Average Average Temperature
TMAX	Average Minimum Temperature
TMIN	Average Maximum Temperature
UAS	Unmanned Airborne Systems
VegDri	Vegetation Drought Response Index
VIF	Variable Inflation Factor
VM	Vegetation Management
vwMAPE	Volume Weighted Mean Absolute Parentage Error
WIN	Winter: January to March
WMS	Work Management System

Appendix B: All Models Accuracy Table

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	65	65	61
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	16	19	11
-	-	-	-	-	-	-	5	-	Naïve	0.764	278.7	38.2%	19.5%	61	62	60
-	-	-	-	-	-	-	5	-	Mean	0.818	244.9	34.0%	16.7%	10	11	9
Units	5	AVG	PRCP	-	-	-	-	-	lm-step	0.785	271.9	36.0%	18.7%	43	51	48
Units	5	AVG	PRCP	-	-	-	-	-	lm-forward	0.779	275.2	36.3%	18.9%	57	58	51
Units	5	AVG	PRCP	-	-	-	-	-	CART	0.790	268.4	35.5%	18.2%	31	37	28
Units	5	EMNT	PRCP	-	-	-	-	-	lm-step	0.780	274.6	36.1%	19.2%	53	53	56
Units	5	EMNT	PRCP	-	-	-	-	-	lm-forward	0.780	274.6	36.1%	19.2%	53	53	56
Units	5	EMNT	PRCP	-	-	-	-	-	CART	0.788	269.8	35.5%	18.3%	40	40	34
Units	5	MAX	PRCP	-	-	-	-	-	lm-step	0.774	278.8	36.4%	19.3%	62	59	59
Units	5	MAX	PRCP	-	-	-	-	-	lm-forward	0.774	278.4	36.2%	19.2%	60	55	54
Units	5	MAX	PRCP	-	-	-	-	-	CART	0.788	269.7	35.5%	18.3%	39	38	32
Units	5	MIN	PRCP	-	-	-	-	-	lm-step	0.795	265.3	35.2%	18.2%	22	25	26
Units	5	MIN	PRCP	-	-	-	-	-	lm-forward	0.795	265.3	35.2%	18.2%	23	26	27
Units	5	MIN	PRCP	-	-	-	-	-	CART	0.792	267.6	35.5%	18.3%	29	42	35
Units	5	DX	PRCP	-	-	-	-	-	lm-step	0.781	274.3	36.3%	18.8%	50	56	49
Units	5	DX	PRCP	-	-	-	-	-	lm-forward	0.781	274.3	36.3%	18.8%	51	57	50
Units	5	DX	PRCP	-	-	-	-	-	CART	0.788	269.9	35.7%	18.3%	41	44	36
Units	5	EMXT	PRCP	-	-	-	-	-	lm-step	0.782	273.3	35.7%	18.6%	49	45	47
Units	5	EMXT	PRCP	-	-	-	-	-	lm-forward	0.783	273.3	35.8%	18.5%	48	46	46
Units	5	EMXT	PRCP	-	-	-	-	-	CART	0.788	269.7	35.4%	18.3%	38	35	31
Units	5	DT	PRCP	-	-	-	-	-	lm-step	0.780	274.6	35.8%	18.9%	55	48	52
Units	5	DT	PRCP	-	-	-	-	-	lm-forward	0.780	274.6	35.8%	18.9%	54	47	53
Units	5	DT	PRCP	-	-	-	-	-	CART	0.784	272.3	35.6%	18.3%	44	43	37
Units	5	MIN	EMXP	-	-	-	-	-	lm-step	0.790	268.6	35.4%	18.5%	34	34	45
Units	5	MIN	EMXP	-	-	-	-	-	lm-forward	0.790	268.6	35.4%	18.5%	33	33	44
Units	5	MIN	EMXP	-	-	-	-	-	CART	0.791	267.9	35.5%	18.3%	30	41	30
Units	5	MIN	DP	-	-	-	-	-	lm-step	0.770	281.1	38.5%	19.9%	64	64	63
Units	5	MIN	DP	-	-	-	-	-	lm-forward	0.770	281.0	38.5%	19.9%	63	63	62
Units	5	MIN	DP	-	-	-	-	-	CART	0.792	267.3	35.4%	18.2%	28	31	24
Units	5	MIN	PRCP	ALL	-	-	-	-	lm-step	0.778	276.1	36.7%	20.0%	59	61	65
Units	5	MIN	PRCP	ALL	-	-	-	-	lm-forward	0.778	276.1	36.7%	20.0%	59	61	65
Units	5	MIN	PRCP	ALL	-	-	-	-	CART	0.783	273.2	35.4%	18.4%	47	30	40
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-step	0.802	260.5	34.7%	18.0%	19	18	20
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-forward	0.802	260.5	34.7%	18.0%	18	17	19
Units	5	MIN	PRCP	SPR	-	-	-	-	CART	0.785	271.9	35.0%	18.3%	42	22	29
Units	5	MIN	PRCP	FAL	-	-	-	-	lm-step	0.783	273.1	35.9%	19.3%	46	50	58
Units	5	MIN	PRCP	FAL	-	-	-	-	lm-forward	0.783	273.0	35.9%	19.3%	45	49	57
Units	5	MIN	PRCP	FAL	-	-	-	-	CART	0.789	268.9	35.5%	18.4%	35	39	38
Units	5	MIN	PRCP	SUM	-	-	-	-	lm-step	0.794	266.0	34.9%	18.5%	25	20	41
Units	5	MIN	PRCP	SUM	-	-	-	-	lm-forward	0.794	266.0	35.0%	18.5%	24	21	42
Units	5	MIN	PRCP	SUM	-	-	-	-	CART	0.789	269.0	35.3%	18.2%	36	28	23
Units	5	MIN	PRCP	WIN	-	-	-	-	lm-step	0.798	263.2	35.0%	18.2%	21	23	22
Units	5	MIN	PRCP	WIN	-	-	-	-	lm-forward	0.801	261.3	35.1%	18.1%	20	24	21
Units	5	MIN	PRCP	WIN	-	-	-	-	CART	0.790	268.4	35.5%	18.3%	32	36	33
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-step	0.810	255.1	34.5%	17.2%	14	15	15
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-forward	0.810	255.1	34.5%	17.2%	15	14	16
Units	5	MIN	PRCP	SPR	X	-	-	-	CART	0.780	274.7	35.4%	18.4%	56	32	39
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-step	0.816	251.3	34.5%	17.1%	12	16	14
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.815	252.1	34.5%	17.1%	13	13	13
Units	5	MIN	PRCP	SPR	X	X	-	-	CART	0.792	267.1	35.3%	18.2%	27	27	25
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-step	0.828	238.3	33.7%	16.6%	3	7	8
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-forward	0.827	239.1	33.6%	16.6%	4	6	7
Units	5	MIN	PRCP	SPR	X	X	5	-	CART	0.801	256.4	34.4%	17.9%	17	12	18
Units	5	MIN	PRCP	SPR	X	X	5	x	lm-step	0.828	238.1	33.2%	16.4%	2	2	4
Units	5	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.830	236.5	32.9%	16.4%	1	1	3
Units	5	MIN	PRCP	SPR	X	X	5	x	CART	0.785	266.1	35.3%	18.5%	26	29	43
K-Means	5	MIN	PRCP	SPR	X	X	5	x	lm-step	0.820	244.0	33.9%	16.3%	7	10	1
K-Means	5	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.820	244.0	33.9%	16.3%	8	9	2
K-Means	5	MIN	PRCP	SPR	X	X	5	x	CART	0.780	269.5	36.2%	17.7%	37	54	17
Single	1	MIN	PRCP	SPR	X	X	5	x	lm-step	0.821	243.1	33.6%	16.4%	6	5	6
Single	1	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.821	243.1	33.6%	16.4%	6	5	6
Single	1	MIN	PRCP	SPR	X	X	5	x	CART	0.819	244.0	33.2%	17.0%	9	3	10
Single	1	MIN	PRCP	SPR	X	X	5	x	RF	0.817	245.6	33.7%	17.1%	11	8	12

Appendix C: All Model Accuracy Table in Rank Order

Input Model Parameters										Accuruacy Metrics				Rankings		
Clustering	Number	Temp	Rainfall	Drought	Practices	Species	Outlier	Normalized	Model	OSR	RMSE	vwMAPE	sMAPE	RMSE	vwMAPE	sMAPE
Units	5	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.830	236.5	32.9%	16.4%	1	1	3
Units	5	MIN	PRCP	SPR	X	X	5	x	lm-step	0.828	238.1	33.2%	16.4%	2	2	4
Single	1	MIN	PRCP	SPR	X	X	5	x	CART	0.819	244.0	33.2%	17.0%	9	3	10
Single	1	MIN	PRCP	SPR	X	X	5	x	lm-step	0.821	243.1	33.6%	16.4%	6	5	6
Single	1	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.821	243.1	33.6%	16.4%	6	5	6
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-forward	0.827	239.1	33.6%	16.6%	4	6	7
Units	5	MIN	PRCP	SPR	X	X	5	-	lm-step	0.828	238.3	33.7%	16.6%	3	7	8
Single	1	MIN	PRCP	SPR	X	X	5	x	RF	0.817	245.6	33.7%	17.1%	11	8	12
K-Means	5	MIN	PRCP	SPR	X	X	5	x	lm-forward	0.820	244.0	33.9%	16.3%	8	9	2
K-Means	5	MIN	PRCP	SPR	X	X	5	x	lm-step	0.820	244.0	33.9%	16.3%	7	10	1
-	-	-	-	-	-	-	5	-	Mean	0.818	244.9	34.0%	16.7%	10	11	9
Units	5	MIN	PRCP	SPR	X	X	5	-	CART	0.801	256.4	34.4%	17.9%	17	12	18
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-forward	0.815	252.1	34.5%	17.1%	13	13	13
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-forward	0.810	255.1	34.5%	17.2%	15	14	16
Units	5	MIN	PRCP	SPR	X	-	-	-	lm-step	0.810	255.1	34.5%	17.2%	14	15	15
Units	5	MIN	PRCP	SPR	X	X	-	-	lm-step	0.816	251.3	34.5%	17.1%	12	16	14
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-forward	0.802	260.5	34.7%	18.0%	18	17	19
Units	5	MIN	PRCP	SPR	-	-	-	-	lm-step	0.802	260.5	34.7%	18.0%	19	18	20
-	-	-	-	-	-	-	-	-	Mean	0.810	255.7	34.8%	17.1%	16	19	11
Units	5	MIN	PRCP	SUM	-	-	-	-	lm-step	0.794	266.0	34.9%	18.5%	25	20	41
Units	5	MIN	PRCP	SUM	-	-	-	-	lm-forward	0.794	266.0	35.0%	18.5%	24	21	42
Units	5	MIN	PRCP	SPR	-	-	-	-	CART	0.785	271.9	35.0%	18.3%	42	22	29
Units	5	MIN	PRCP	WIN	-	-	-	-	lm-step	0.798	263.2	35.0%	18.2%	21	23	22
Units	5	MIN	PRCP	WIN	-	-	-	-	lm-forward	0.801	261.3	35.1%	18.1%	20	24	21
Units	5	MIN	PRCP	-	-	-	-	-	lm-step	0.795	265.3	35.2%	18.2%	22	25	26
Units	5	MIN	PRCP	-	-	-	-	-	lm-forward	0.795	265.3	35.2%	18.2%	23	26	27
Units	5	MIN	PRCP	SPR	X	X	-	-	CART	0.792	267.1	35.3%	18.2%	27	27	25
Units	5	MIN	PRCP	SUM	-	-	-	-	CART	0.789	269.0	35.3%	18.2%	36	28	23
Units	5	MIN	PRCP	SPR	X	X	5	x	CART	0.785	266.1	35.3%	18.5%	26	29	43
Units	5	MIN	PRCP	ALL	-	-	-	-	CART	0.783	273.2	35.4%	18.4%	47	30	40
Units	5	MIN	DP	-	-	-	-	-	CART	0.792	267.3	35.4%	18.2%	28	31	24
Units	5	MIN	PRCP	SPR	X	-	-	-	CART	0.780	274.7	35.4%	18.4%	56	32	39
Units	5	MIN	EMXP	-	-	-	-	-	lm-forward	0.790	268.6	35.4%	18.5%	33	33	44
Units	5	MIN	EMXP	-	-	-	-	-	lm-step	0.790	268.6	35.4%	18.5%	34	34	45
Units	5	EMXT	PRCP	-	-	-	-	-	CART	0.788	269.7	35.4%	18.3%	38	35	31
Units	5	MIN	PRCP	WIN	-	-	-	-	CART	0.790	268.4	35.5%	18.3%	32	36	33
Units	5	AVG	PRCP	-	-	-	-	-	CART	0.790	268.4	35.5%	18.2%	31	37	28
Units	5	MAX	PRCP	-	-	-	-	-	CART	0.788	269.7	35.5%	18.3%	39	38	32
Units	5	MIN	PRCP	FAL	-	-	-	-	CART	0.789	268.9	35.5%	18.4%	35	39	38
Units	5	EMNT	PRCP	-	-	-	-	-	CART	0.788	269.8	35.5%	18.3%	40	40	34
Units	5	MIN	EMXP	-	-	-	-	-	CART	0.791	267.9	35.5%	18.3%	30	41	30
Units	5	MIN	PRCP	-	-	-	-	-	CART	0.792	267.6	35.5%	18.3%	29	42	35
Units	5	DT	PRCP	-	-	-	-	-	CART	0.784	272.3	35.6%	18.3%	44	43	37
Units	5	DX	PRCP	-	-	-	-	-	CART	0.788	269.9	35.7%	18.3%	41	44	36
Units	5	EMXT	PRCP	-	-	-	-	-	lm-step	0.782	273.3	35.7%	18.6%	49	45	47
Units	5	EMXT	PRCP	-	-	-	-	-	lm-forward	0.783	273.3	35.8%	18.5%	48	46	46
Units	5	DT	PRCP	-	-	-	-	-	lm-forward	0.780	274.6	35.8%	18.9%	54	47	53
Units	5	DT	PRCP	-	-	-	-	-	lm-step	0.780	274.6	35.8%	18.9%	55	48	52
Units	5	MIN	PRCP	FAL	-	-	-	-	lm-forward	0.783	273.0	35.9%	19.3%	45	49	57
Units	5	MIN	PRCP	FAL	-	-	-	-	lm-step	0.783	273.1	35.9%	19.3%	46	50	58
Units	5	AVG	PRCP	-	-	-	-	-	lm-step	0.785	271.9	36.0%	18.7%	43	51	48
Units	5	EMNT	PRCP	-	-	-	-	-	lm-step	0.780	274.6	36.1%	19.2%	53	53	56
Units	5	EMNT	PRCP	-	-	-	-	-	lm-forward	0.780	274.6	36.1%	19.2%	53	53	56
K-Means	5	MIN	PRCP	SPR	X	X	5	x	CART	0.780	269.5	36.2%	17.7%	37	54	17
Units	5	MAX	PRCP	-	-	-	-	-	lm-forward	0.774	278.4	36.2%	19.2%	60	55	54
Units	5	DX	PRCP	-	-	-	-	-	lm-step	0.781	274.3	36.3%	18.8%	50	56	49
Units	5	DX	PRCP	-	-	-	-	-	lm-forward	0.781	274.3	36.3%	18.8%	51	57	50
Units	5	AVG	PRCP	-	-	-	-	-	lm-forward	0.779	275.2	36.3%	18.9%	57	58	51
Units	5	MAX	PRCP	-	-	-	-	-	lm-step	0.774	278.8	36.4%	19.3%	62	59	59
Units	5	MIN	PRCP	ALL	-	-	-	-	lm-step	0.778	276.1	36.7%	20.0%	59	61	65
Units	5	MIN	PRCP	ALL	-	-	-	-	lm-forward	0.778	276.1	36.7%	20.0%	59	61	65
-	-	-	-	-	-	-	5	-	Naïve	0.764	278.7	38.2%	19.5%	61	62	60
Units	5	MIN	DP	-	-	-	-	-	lm-forward	0.770	281.0	38.5%	19.9%	63	63	62
Units	5	MIN	DP	-	-	-	-	-	lm-step	0.770	281.1	38.5%	19.9%	64	64	63
-	-	-	-	-	-	-	-	-	Naïve	0.755	290.1	38.9%	19.9%	65	65	61

References

[1] J. Estornell, L. Ruiz, B. Velazquez-Martí and e. al., "Estimation of pruning biomass of olive trees using airborne discrete-return LiDAR data," *Biomass and Bioenergy,* vol. 81, pp. 315-321, 2015.

[2] J. B. Piccini, J. Magdalena and e. al., "Early yield prediction in pear based on canopy LIDAR scanning," in *Workshop on Information Processing and Control (RPIC)*, Mar del Plata, Argentina, 2017.

[3] L. Reineke, "Perfecting a stand-density index for even-aged," *Journal of Agricultural Research,* vol. 46, no. 6, pp. 627-638, 1933.

[4] B. Zeide, "How to Measure Stand Density," *Trees: Structure & Function,* vol. 19, no. 1, pp. 1-14, 2005.

[5] V. H. Dale, T. W. Doyal and H. H. Shugart, "A comparision of tree growth models," *Ecological Modelling,* vol. 29, pp. 145-169, 1985.

[6] E. G. Mason and H. Dzierzon, "Applications of modeling to vegetation management," *Canadian Journal of Forest Research,* vol. 36, no. 10, pp. 2505-2514, 2006.

[7] M. Follett, C. A. Nock, C. Buteau and C. Messier, "Testing a New Approach to Quantify Growth Responses to Pruning Among Three Temperate Tree Species," *Arboriculture & Urban Forestry,* vol. 42, no. 3, pp. 133-145, 2016.

[8] G. D. Eschelbach, Wires-Down Predictive Modeling and Preventative Measures Optimization, Masters thesis, Massachusetts Institute of Technology, 2016.

[9] D. Wanik, J. Parent, E. Anagnostou and B. Hartman, "Using vegetation management and LiDAR-derived tree height data to improve outage predictions for electric utilities," *Electric Power Systems Research,* vol. 146, pp. 236-245, 2017.

[10] D. Radmer, P. Kuntz, R. Christie, S. Venkata and R. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Transactions on Power Delivery,* vol. 17, no. 4, pp. 1170-1175, 2002.

[11] B. L. Kelchev, Predicting Rejection Rates of Electric Distribution Wood Pole Assets, Masters thesis, Massachusetts Institute of Technology, 2017.

[12] S. D. Whipple, Predictive Storm Damage Modeling and Optimizing Crew Response to Improve Storm Response Operations, Masters thesis, Massachusetts Institute of Technology, 2014.

[13] "Historical Palmer Drought Indices," National Oceanic and Atmospheric Administration, [Online]. Available: https://www.ncdc.noaa.gov/temp-and-precip/drought/historical-palmers/overview. [Accessed 20 June 2017].

[14] "Vegetation Drought Response Index," The National Drought Mitigation Center, [Online]. Available: http://vegdri.unl.edu/Home.aspx. [Accessed 20 June 2017].

[15] F. L. Ramsey and D. W. Schafer, The Statistical Sleuth : A Course in Methods of Data Analysis, Boston: Brooks/Cole, Cengage Learning, 2013.

[16] D. Bertsimas, . A. O'Hair and . W. Pulleyblank, The Analytics Edge, Belmont, Massachusetts: Dynamic Ideas LLC, 2016.

[17] P. F. Vellema and R. E. Welsch, "Efficient Computing of Regression Diagnostics," *American Statistician,* vol. 35, no. 4, pp. 234-242, 1981.