

Predicting Rejection Rates of Electric Distribution Wood Pole Assets

by
Boyan Lyubomirov Kelchev
B.A. Computer Science, University of San Diego, 2009

Submitted to the Department of Electrical Engineering and Computer Science and the MIT Sloan School of Management in partial fulfillment of the Requirements for the Degrees of

Master of Science in Electrical Engineering and Computer Science
and
Master of Business Administration

In conjunction with the Leaders for Global Operations Program at the
Massachusetts Institute of Technology
June 2017

©2017 Boyan Lyubomirov Kelchev. All rights reserved.

The author hereby grants MIT permission to reproduce and to distribute publicly copies of this thesis document in whole or in part in any medium now know or hereafter created.

Signature of Author **Signature redacted**
Boyan Lyubomirov Kelchev
MIT Sloan School of Management
Department of Electrical Engineering and Computer Science

Certified by **Signature redacted** May 12, 2017
~~Georgia Perakis~~, Thesis Supervisor
William F. Pounds Professor of Management Science
MIT Sloan School of Management

Certified by **Signature redacted**
~~Patrick V. Met~~, Thesis Supervisor
Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science

Accepted by **Signature redacted**
Leslie Kolodziejski, Chair of the Committee on Graduate Students
Department of Electrical Engineering and Computer Science

Accepted by **Signature redacted**
Maura Heron, Director of MIT Sloan MBA Program
MIT Sloan School of Management



THIS PAGE INTENTIONALLY LEFT BLANK

Predicting Rejection Rates of Electric Distribution Wood Pole Assets

by

Boyan Lyubomirov Kelchev

Submitted to the Department of Electrical Engineering and Computer Science and the MIT Sloan School of Management on May 12, 2017 in partial Fulfillment of the Requirements for the Degrees of Master of Science in Electrical Engineering and Computer Science and Master of Business Administration

Abstract

Pacific Gas & Electric Company's (PG&E) electric distribution system includes approximately 2.4 million wood utility poles. The Pole Test & Treat (PTT) program at PG&E is responsible for inspecting these poles, prolonging their service life through the use of chemical treatments or structural reinforcements, and identifying poles that need to be replaced. Following industry best practices and taking advantage of the vast knowledge and experience of the PTT team, PG&E inspects poles every 10 years. The company believes that the next step in improving the performance of the PTT program is to leverage the data collected since the inception of the program and utilize modern statistical methods to better understand and predict decay in their wood pole assets.

In this thesis, we describe the possibilities and limitations of using PG&E's current data to predict the results of future inspections. We study both the possibility of making predictions at the individual pole level, predicting whether a pole will be rejected during the next inspection cycle, and at the aggregate level, predicting what the overall rejection rate in a subpopulation of poles will be in the future.

In order to accomplish this, we first studied the available data sources and performed exploratory analysis to understand the characteristics of the different variables and form hypotheses about the main drivers of rejections during pole inspections. Next, we attempted to build a classification model to predict the results of future inspections. This showed us that our current data cannot be used to yield an accurate prediction at the individual pole level. Then, we developed a model to estimate the overall rejection rates of subpopulations of poles. The result was a prediction with a Mean Absolute Percentage Error of about 30%. While not ideal, this model gives PG&E the ability to budget and plan for future work better.

Finally, we leveraged the results of the prediction model to simulate the evolution of rejection rates in the future. The simulation highlighted a well-known problem in the utility industry - the problem of aging infrastructure. The relatively low average age of poles and the low replacement rates observed in the past few inspection cycles mean that PG&E will likely experience a drastic increase in rejection rates as the average age of its pole population grows. Planning for the accompanying increase in manpower and work hours required will be of great importance to PG&E in the next few decades.

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

Thesis Supervisor: Patrick Jaillet
Title: Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I would like to take a moment to thank all those who provided guidance and support throughout my research project and the preparation of this thesis. First, I would like to thank all the people I had the pleasure to work with at PG&E. Without an exception, all of them were willing to take time out of their busy days to help me get up to speed with the subject matter and provide invaluable feedback on the analyses we performed. I would specifically like to thank all the folks from the Pole Test & Treat Program, especially Mike Pallatroni, Mike Koffman, and Pavel Chovanec; my supervisor, Karen O'Connor; as well as Jay Singh and Eric Back. I would also like to thank David Feliciano and Arvind Simhadri, both MIT LGO 2015 alumni, for helping me get integrated into the PG&E family. Thank you also to MIT LGO 2016 alumni Gregory Eschelbach and Lillian Meyer for providing insightful information and recommendations about my project and working at PG&E.

I would also like to thank my MIT advisors Professor Georgia Perakis and Professor Patrick Jaillet for the guidance and feedback they have given me throughout the past year.

Last but not least, I would like to extend gratitude to my family in Bulgaria as well as my LGO classmates and other MIT friends for your support in the past two years and especially during this research project. Stress and anxiety are but tiny inconveniences when I have all of you around me.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

- 1 Introduction and Background 12**
 - 1.1 Company Overview 12
 - 1.2 Pole Test & Treat Program Overview 12
 - 1.3 Problem Statement and Goals 15
 - 1.4 Thesis Contributions 15
 - 1.5 Thesis Outline 15
 - 1.6 Literature Review 16

- 2 Data Sources 17**
 - 2.1 Overview 17
 - 2.1.1 Pole asset and pole inspection data 17
 - 2.1.2 Land cover and soil classes 18
 - 2.1.3 Weather data 19
 - 2.2 Preparing the data set 20
 - 2.2.1 Past inspection data as model features 20
 - 2.2.2 Categorizing the values of some nominal variables 21
 - 2.2.3 Incorporating environmental data 22
 - 2.3 Final data set 23

- 3 Exploratory Analysis 24**
 - 3.1 Age of pole population 24
 - 3.2 Rejection rates for different pole species and chemical treatments 25
 - 3.3 Rejection rates in different climate zones 26
 - 3.4 Rejection rates for different land cover and soil types 27

- 4 Rejection Classification Model 29**
 - 4.1 Assessing classifier performance using all features 29
 - 4.1.1 Motivation 29
 - 4.1.2 Classification methods used 29
 - 4.1.3 Methodology 30
 - 4.1.4 Results 31
 - 4.2 Assessing the importance of predictors 33
 - 4.2.1 Motivation 33
 - 4.2.2 Methodology 33
 - 4.2.3 Results for unstubbed poles 33
 - 4.2.4 Results for stubbed poles 35
 - 4.2.5 Conclusion 37

- 5 Predicting Rejection Rates 38**
 - 5.1 Motivation and hypothesis 38
 - 5.2 Methodology 38
 - 5.3 Results 38

6 Pole Population Aging Simulation Model	41
6.1 Motivation	41
6.2 Methodology	42
6.3 Results	43
7 Conclusions and Future Work	47
7.1 Improving data quality to facilitate future analyses	47
7.2 Using results of simulation model to inform operational decisions	48
A Appendix: Acronyms Used	49
B Appendix: Variables in Original Data Set	50
B.1 Pole Asset Information	50
B.2 Pole Inspection Records	51
B.3 Appendix: Land cover classes	52
B.4 Appendix: Soil Group Types	54
C Appendix: Weather Stations	55
D Appendix: Rejections by Age for Different Weather Profiles	56
E Appendix: Rejections by Age for Different Species and Original Treatments	57
F Appendix: Assessing Multicollinearity using Variance Inflation Factor	59

List of Figures

1	Pole inspection process	12
2	Pole inspection measurements	13
3	Pole reinforcement (i.e. "stubbing")	14
4	Percentage of poles in each land cover class	18
5	Percentage of poles in each land cover class	19
6	Age profile of pole population	25
7	Distribution of poles by species and original treatment	26
8	Rejections by Mean Temperature	27
9	Rejections by age for different land cover types	28
10	Rejections by age for different soil group type	28
11	Accuracy of Bagging classifier for different subsets of features	34
12	Top 20 features selected in analysis of poles that have not been reinforced	35
13	Accuracy of Bagging classifier for different subsets of features	36
14	Top 20 features selected in analysis of reinforced poles	37
15	Model accuracy comparison between simplified model and full model	39
16	Model predictions vs actuals for a single run of the prediction model	40
17	Simulation of rejection rates as pole population is aged	43
18	Simulation of rejection rates as pole population is aged	44
19	Simulation of rejection rates forcing a 10% replacement rate	45
20	Simulation of rejection rates forcing a 11% replacement rate	46
21	Simulation of rejection rates forcing a 12% replacement rate	46
22	Location of weather stations used in analysis	55
23	Rejections by Age and Average Occurrence of Fog	56
24	Rejections by Age and Average Precipitation Amount	56
25	Douglas Fir rejection rates by age	57
26	Western Cedar rejection rates by age	57
27	Western Pine rejection rates by age	58
28	Variance inflation factor of variables in data set	59

List of Tables

1	Original pole asset information	20
2	Supplier categorization	21
3	Original treatment categorization	22
4	Land cover categorization	22
5	Summary statistics of distance from poles to closest weather station	23
6	Sample confusion matrix calculated from classifier results	31
7	Results of classifiers for all poles that have not been reinforced	32
8	Results of classifiers for poles that have not been reinforced and are located 20 miles or less from a weather station	32
9	Results of classifiers for all reinforced poles	32
10	Results of classifiers for reinforced poles located 20 miles or less from a weather station	33
11	Mean Absolute Percent Error (MAPE)	41
12	Abbreviations used	49

13	Original pole asset information	50
14	Original pole inspection information	51
15	Land cover classification	52
16	Land cover classification	53
17	Soil classification	54

THIS PAGE INTENTIONALLY LEFT BLANK

1 Introduction and Background

1.1 Company Overview

Pacific Gas & Electric Company (PG&E) is one of the largest combined electric and gas utilities in the nation. The company operates in most of central and all of northern California, spanning an area from Bakersfield to the south to the border with Oregon to the North. PG&E owns and operates assets in all parts of the electric system: generation, transmission, and distribution. In particular, PG&E's electric distribution system includes approximately 2.45 million utility poles. Approximately 98% of all poles are wood poles (Figure 1a) (other poles can be made from steel, concrete, fiberglass). Wood poles are an essential part of the electric distribution network, which connects the electric transmission system with most customers. PG&E inspects these poles on a regular basis to ensure their structural integrity. The company is overseen by the California Public Utilities Commission (CPUC), which also regulates the pole inspection process and stipulates how often poles should be inspected.



(a) Utility pole in San Francisco Bay Area (b) Structural reinforcement (i.e. stub) (c) Chemical treatment applied to poles

Figure 1: Pole inspection process

1.2 Pole Test & Treat Program Overview

In the mid-1990s, PG&E institutes the Pole Test & Treat (PTT) Program, which is responsible for inspecting utility wood poles, prolonging their service life through the use of chemical treatments (Figure 1c) or structural reinforcements (Figure 1b), and identifying poles that need to be replaced.

In other words, PG&E's PTT program looks to cost-effectively maintain the safety and reliability of wood pole assets. Central to PTT's operations is its pole inspection process.

Following industry best practices and taking advantage of the vast knowledge and experience of the PTT team, PG&E currently inspects poles every 10 years. During each inspection, several checks are made and measurements taken. Most importantly, poles are checked for internal decay at different levels below and above ground. To accomplish this, a field specialist excavates around the pole, bores holes in the pole at different heights, and uses a metal instrument (Figure 2b) to determine whether there is internal decay. If decay is present, the thickness of the remaining wood shell is measured.



(a) Measuring circumference of pole (b) Instrument for measuring shell thickness and internal decay

Figure 2: Pole inspection measurements

Similarly, the circumference of each pole is measured (Figure 2a) and compared to the pole's original circumference in order to determine whether the pole has suffered any external decay. Any external damage that may have been caused by a vehicle, a human, or birds, such as woodpeckers, is also noted. Remaining shell and remaining effective circumference are the two main criteria used to determine whether a pole passes an inspection or not. Additionally, a pole may not pass an inspection because of other factors such as excessive external damage in the non-serviceable part of a pole (e.g. top part of the pole) or a significant increase in the loading of a pole (e.g. when additional equipment is to be installed on it).

When a pole does not pass an inspection, it is rejected, which means that it must be reinforced or replaced. If the pole is in a condition that allows prolonging its lifetime with a structural reinforce-

ment, the pole is reinforced or "stubbed" (Figure 3). Stubbing a pole consists of affixing a steel reinforcement to the side of the pole with a portion of it being drilled into the ground. The size of the reinforcement depends on the size of the pole. If a pole is in a condition that does not allow for it to be reinforced, the pole is replaced.



(a) Excavation for placement of steel stub (b) Placement of steel stub at rejected pole (c) Placement of bands to affix stub

Figure 3: Pole reinforcement (i.e. "stubbing")

Apart from measuring the shell thickness and effective circumference of a pole, the field specialist assesses the condition of the bottom and top parts of the pole. In some cases depending on the age and characteristics of the pole (e.g. wood species type, original chemical treatment performed when the pole was manufactured), preservatives may be applied externally in order to protect the bottom section of the pole from fungi and insects. Similarly, poles may be treated internally through fumigation during an inspection.

The pole inspection process and criteria for rejection change after a pole has been reinforced. The bottom of the pole is still visually inspected for external damage that may compromise the protection provided by the stub. At the same time, because the bottom of the pole is protected by the reinforcement, the below ground shell thickness is no longer measured. Instead, the shell thickness close to the top of the reinforcement is measured to ensure that the pole is still strong enough and will not snap above at the level or above the stub.

It is important to note that the PTT program has undergone a number of changes throughout the years. Most importantly for our analysis, the pole inspection process and the criteria used to

reject a pole have evolved. For example, until about a few years ago, the average shell thickness determined whether a pole should pass an inspection, be reinforced, or be replaced. Since 2014 a proprietary formula, unknown to PG&E and developed by a private company and an industry leader, has been used to calculate wood strength taking into consideration the remaining shell thickness and circumference along with other factors such as the position of the decay pockets in a pole.

1.3 Problem Statement and Goals

PG&E believes that the next step in improving the performance of the PTT program is to leverage the data collected since the inception of the program and utilize modern statistical methods to better understand the risks posed by the company's distribution wood pole assets, attempt to predict decay in poles, and potentially identify efficiency opportunities, related, for example, to the inspection procedures and the frequency of inspections.

The goal of this thesis is to describe the possibilities and limitations of using modern statistical methods on PG&E's current pole data to better understand the drivers of wood decay in poles, analyze the relationships between environmental conditions and wood pole decay, and ultimately predict the results of future inspections both at the individual pole level and at the aggregate level (i.e. for varying subpopulations of poles).

1.4 Thesis Contributions

The current thesis describes the results of our research project, which constituted PG&E's first attempt in utilizing machine learning methods to analyze its utility pole asset and inspection data. We present the challenges and limitations in using such statistical methods to make predictions about pole inspection results despite the large amounts of available data. We further describe a method for estimating rejection rates for subpopulations of poles. While not ideal, the resultant prediction with a Mean Absolute Percentage Error (MAPE) of approximately 30% gives PG&E the ability to budget and plan for future work better. Last but not least, we propose a simulation method, which provides a rough order of magnitude estimate of the rejection rates the company can expect in the next several decades.

1.5 Thesis Outline

In Chapter 2, we describe the data that was available for our analysis, including both PG&E data on poles and pole inspections as well as publicly available environmental data. In Chapter 3, we discuss the results of our preliminary exploratory analysis, which helped us understand the characteristics of the different variables in the data as well as form hypotheses about the main drivers of rejections during pole inspections. In Chapter 4, we describe the use of several different classification algorithms in an attempt to predict the results of future inspections on a per-pole basis. In Chapter 5, we describe a model for estimating the overall rejection rate of subpopulations of poles (e.g. within a district in Northern California). In Chapter 6, we discuss the use of the rejection rate estimation model to simulate how rejection rates may evolve in the future. The simulation highlights a well-known problem in the utility industry - the problem of aging infrastructure. The

results of the simulation provide an approximate but concrete picture of the pole rejection rates PG&E may face in the next several decades. Finally, in Chapter 7 we summarize the results of this research project and present a case for why the likely increase in rejection, and especially replacement, rates in the future will lead to a significant increase in manpower and work hours required for pole replacements in the next few decades.

1.6 Literature Review

While there are many private technology companies that are using machine learning algorithms to analyze large amounts of data and are targeting utility companies as potential customers, the use of data analytics at utilities until recently seems to have been confined mostly to analyzing past data for the purposes of reporting. In the past several years, however, PG&E and other utilities have made great strides towards utilizing historical data for asset health prediction and predictive maintenance. In two recent master theses, for example, MIT graduate students, along with teams from PG&E, studied the possibility of predicting corrosion on buried natural gas distribution pipelines [5] and predicting wires-down events [3]. Modern statistical methods were also used in another PG&E-sponsored master's research project for anomaly detection in natural gas regulator stations [2].

Before this project, however, the company had not attempted to apply similar methods to analyzing utility pole data. During our literature review, we were unable to find public information about many others who have performed similar studies either. [7], for example, discusses a Bayesian statistical model for estimating the structural reliability of utility poles during hurricanes, but not under normal operating conditions and during periodic inspections.

That being said, other topics relevant to our project, such as the study of decay in utility wood poles, is documented well in literature. Underground and above-ground external and internal decay are common failure modes in utility poles. The rate and type of decay depend on a multitude of factors, including the type of wood species the pole is made of, the chemical treatment used during manufacturing, the environmental conditions the pole experiences throughout its service life, and others. [6] discusses different types of preservatives utilized for the purpose of preventing wood decay. The report also gives information on a number of agents of decay, including fungi, insects, woodpeckers, and marine borers. [4] applies survival analysis to more than 17,000 utility pole records from the Pacific Northwest and concludes that poles with a larger circumference decay sooner, that Douglas fir poles last shorter than Western Red Cedar poles, and that the presence of a transformer makes decay more likely. The author acknowledges that the results may not apply generally to all geographies, but the method could be applied to utility poles in different regions of the country.

In this thesis, we take these findings into consideration as we form hypotheses about the variables that are important in our model, but we focus primarily on applying machine learning methods in an attempt to predict the results of inspections in the future.

2 Data Sources

2.1 Overview

There are 5 main data sources that we included in our statistical analysis: Pole asset information, pole inspection records, weather logs, land cover types, and soil group types.

2.1.1 Pole asset and pole inspection data

Pole asset information includes pole location data and characteristics known from the time of installation. A full listing of the data columns can be found in Appendix B.1 but here is a list of some of the most important ones:

- **Species** - Wood species type (e.g. Douglas Fir, Western Red Cedar, Western Pine, Lodgepole Pine)
- **Orig Treatment** - Original chemical treatment performed during manufacture to ensure the durability of the pole (e.g. Pentachlorophenol, Cellon Gas, Creosote)
- **Ins Year** - Year when the pole was installed
- **Original Circumference** - Original circumference of the pole as specified at the time of manufacture

Pole inspection records include information and measurements recorded during a PTT inspection or a stubbing project. We had about 3.5M inspection records from the inception of the PTT program in the mid 1990s through March 2016. These included all records to-date from cycles 2 and 3, the two most recent inspection cycles, and a subset of the records from cycle 1. A full listing of all data columns is available in Appendix B.2. The following is a list of some of the most important ones:

- **Visit Date** - Date of the inspection
- **Result Status** - Result of the inspection (Pass, Reinforce, Replace)
- **Excavation** - Whether an excavation was performed and if so what type - partial or full
- **External Treatment** - Whether below ground preservative was applied to the pole
- **Ground line Shell Avg** - Average shell thickness at ground line
- **Pole Bottom Condition** - Qualitative assessment (bad, fair, good) of the condition of the bottom part of the pole
- **Effective Circ** - Current circumference accounting for any external damage

2.1.2 Land cover and soil classes

The land cover and soil group types data was provided by PG&E's GIS group, but both data sets are publicly available. The land cover information is from the National Landcover Dataset 2011¹. The soil data is from the Natural Resources Conservation Service's soils dataset STATSGO². Descriptions of the land cover and soil classes can be found in appendices B.3 and B.4. Figures 4 and 5 below describe the distribution of poles by land cover and soil groups.

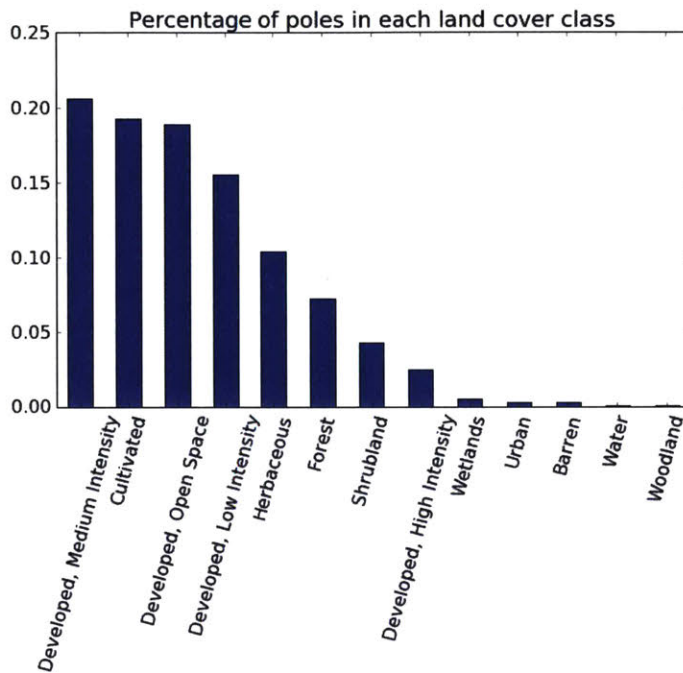


Figure 4: Percentage of poles in each land cover class

¹http://www.mrlc.gov/nlcd11_leg.php

²http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053629

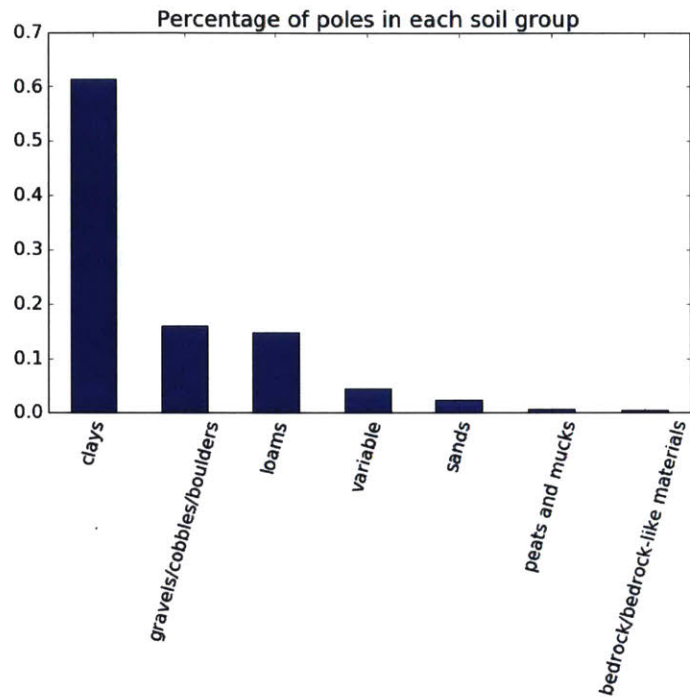


Figure 5: Percentage of poles in each land cover class

2.1.3 Weather data

We used daily weather summary data from the National Climatic Data Center's (NCDC) website³. The following table describes the weather elements available in the data NCDC's data set.

³<http://www7.ncdc.noaa.gov/CDO/cdoselect.cmd>

Weather element	Description
TEMP	Mean temperature (.1 Fahrenheit)
DEWP	Mean dew point (.1 Fahrenheit). A measure of atmospheric moisture. It is the temperature to which air must be cooled in order to reach saturation (assuming air pressure and moisture content are constant). A higher dew point indicates more moisture present in the air. It is sometimes referred to as Dew Point Temperature, and sometimes written as one word (Dewpoint) ⁴ .
WDSP	Mean wind speed (.1 knots)
MXSPD	Maximum sustained wind speed (.1 knots)
GUST	Maximum wind gust (.1 knots)
MAX	Maximum temperature (.1 Fahrenheit)
MIN	Minimum temperature (.1 Fahrenheit)
PRCP	Precipitation amount (.01 inches)
FRSHTT	6-digit Indicator (e.g. 010010) for occurrence of: Fog, Rain or Drizzle, Snow or Ice Pellets, Hail, Thunder, Tornado/Funnel Cloud. This element was split into 6 binary variables indicating the presence of the respective weather condition on the day of the measurement.

Table 1: Original pole asset information

2.2 Preparing the data set

In this section, we discuss some of the most important preprocessing steps that were performed to prepare the data for statistical analysis. Most of the preprocessing of the data was performed using the Anaconda python distribution and the pandas and numpy libraries in particular. The code was run on an 8-processor Virtual Machine with 16GB of RAM provided by PG&E’s Business Technology group. Without the availability of the powerful virtual server it would have been difficult to process the data and perform the analysis described in the following sections.

2.2.1 Past inspection data as model features

First, we selected only those poles that we had two or more inspection records for. This way, we could use the result of the very last inspection as the dependent variable and keep the measurements taken from the prior inspection along with the environmental data (weather, land cover, and soil) as features in the model. PG&E is currently in its third 10-year inspection cycle. Consequently, there were poles in our data set with inspection records from cycle 1 and cycle 2 or from cycle 2 and cycle 3. There were no poles with records from all three inspection cycles due to the fact that a portion of the inspection data from cycle 1 was kept in a separate database. Designing our data set this way meant that we could train and test a classification model that used the results of prior inspections along with environmental information in order to attempt to predict the results of future inspections.

Second, we used the installation date and the date of the most recent inspection to calculate the age of the pole as of the most recent inspection. We excluded any pole records with data that

seemed erroneous. For example, there were about 65,000 poles with an installation year of 1900, which had been used as a placeholder value during a database migration for pole records with unknown installation dates. Additionally, we excluded about 100,000 poles with missing installation dates and about 63,000 poles with installation dates prior to the pole’s recorded manufacture year. This last case may have been due to the same equipment ID being used for two different poles.

Lastly, we excluded all poles that had at one point in their lifetime been marked for replacement, but had additional inspection records after that.

2.2.2 Categorizing the values of some nominal variables

During the exploratory analysis and preliminary phases of our statistical modeling efforts, we noticed that some of the values of variables such as Supplier, Original Treatment, Surface, and Land cover could be consolidated into fewer categories. This resulted into a simpler model because it reduced the number of categories of the respective nominal variable thereby reducing the number of binary variables in the final data set. Additionally, it made the model more generalizable because it reduced the number of values that had very low variance and may not be present in the input data sets during future executions of the model.

Care was taken to categorize the values in logical ways that would not compromise the accuracy of the model. For example, original treatment values were categorized if they were known to refer to the same treatment or to related treatments with similar performance. Similarly, suppliers were categorized if they were known to be the same company or have the same parent company. Land cover types with low representation in the data were categorized based on the classification put forth by the Multi-Resolution Land Characteristics (MRLC) Consortium.

The tables below describe the categorizations that were performed:

Category	Values
Carney	B J Carney, Carney Co
Baxco	J H Baxter Co, Baxco
McFarland Cascade Co	McFarland Cascade Co, L D McFarland

Table 2: Supplier categorization

Note that some of the treatments listed in the table below are no longer used but are included here for completeness. There may be old records of poles treated with the respective original treatments that have already been replaced.

Category	Values
Penta	Penta, Penta in Petroleum, Pentachlorophenal
Cellon	Cellon Gas, Penta in Liquid Petroleum Gas - Cellon, Butane
Methylene Chloride	Penta in Chlorinated Hydrocarbon Solvent, Methane Propane
Creosote	Creosote, Penta in Creosote
Metal	Ammoniacal Copper Zinc Arsenate, Chromated Copper Arsenate, Zinc Metal Arsenate
Copper Napthenate	Copper Napthenate, CNI, Copper Napthenate ('CNI')
Other	Non-Pentachlorophenal, Other, Unknown, OTH, U
No treatment	No treatment

Table 3: Original treatment categorization

Note that the values that were not grouped together with other values are not included in the table below. For example, land cover categorizations such as Urban and the different levels of the classification Developed were not grouped with other categories because of their higher representation in the data set.

Category	Values
Barren	Barren/Other, Barren Land
Forest	Deciduous Forest, Evergreen Forest, Mixed Forest, Hardwood Forest, Conifer Forest
Shrubland	Dwarf Scrub, Shrub/Scrub, Desert Shrub, Shrub
Herbaceous	Grassland/Herbaceous, Sedge/Herbaceous, Lichens, Moss, Herbaceous, Herbaceous
Cultivated	Agriculture, Pasture/Hay, Hay/Pasture, Cultivated Crops
Wetlands	Woody Wetlands, Emergent Herbaceous Wetlands, Emergent Herbaceous Wetlands, Wetland
Water	Open Water, Perennial Ice / Snow, Water

Table 4: Land cover categorization

2.2.3 Incorporating environmental data

The land cover and soil data provided by PG&E's GIS group was incorporated into the pole data set easily because it was already associated with pole equipment identification numbers. Adding the weather data to the data set was more challenging.

The National Climatic Data Center's (NCDC) provides an interface that allows for daily summary weather data to be downloaded from stations across the globe. We downloaded weather data for the past twenty years for all weather stations in California because we wanted to summarize the weather conditions that each pole had experienced in the time between the two most recent inspections that had been performed on it. This way we could use the weather summary values as features in our statistical model.

The first challenge we came across was that there were weather stations with different identification codes that had operated in the same geographical location or within a few miles from one another at different periods of time. To address this, we wrote a procedure to consolidate the weather data for such stations under a single station.

We associated each pole with the closest weather station based on the Vincenty distance formula using the python geopy library. Unfortunately, the number and density of weather stations is not sufficient to provide accurate weather summaries for many of the poles in PG&E’s area of service. The table below lists the summary statistics for the distance from poles in our data set to the closest weather station. In districts such as Willow Creek, North Bay, Solano, Kern, and areas in the Sierra Nevada Mountains 70-90% of the poles are further than 20 miles away from the closest weather station. In districts such as Garberville and Ukiah that percentage is close to 100%. A map of the weather stations that we used is available in Appendix C.

Mean	Standard Deviation	Minimum	25 th per-centile	50 per-centile	75 th per-centile	Maximum
56.08	31.03	1.00	33.00	54.00	74.00	126.00

Table 5: Summary statistics of distance from poles to closest weather station

To summarize the weather conditions reported in the vicinity of a pole, we decided to calculate averages and standard deviations of the daily weather summary values reported at the weather station associated with that pole. The average would give us a general idea of the respective weather element (e.g. temperature, precipitation, etc) while the standard deviation would capture the variation of that element. Because wood decay is a process that develops over long periods of time and in the interest of simplicity, we chose to use this over other methods such as attempting to analyze seasonal weather patterns or extreme weather conditions. Lastly, we excluded any pole records for which the calculated averages or standard deviations were not calculated correctly due to missing weather values.

2.3 Final data set

After we completed all preprocessing of the data and incorporated the weather and other environmental variables we were left with a data set of about 850,000 records. Each record represented a unique pole with attributes that included:

- the result of the last inspection
- the measurements of the prior inspection
- the land cover and soil type of the area the pole was located in
- the average and standard deviation of several weather element values as recorded by the closest weather station that we had data for

3 Exploratory Analysis

In the exploratory analysis phase, we analyzed some of the characteristics of our data set and the relationship between various variables and the observed rejection rates. Below are some of the results that informed the subsequent analysis described in the next sections.

3.1 Age of pole population

Figure 6 below shows the age profiles of utility poles in PG&E's system at the time of their PTT inspection during cycles 1, 2 and 3. Note that the plotted data consists solely of those inspection records available in the current database system of record. In particular, the plotted data comprises 750,000 records from cycle 1, all 1,918,000 records from cycle 2 and the 265,000 records available to date from cycle 3. The poles inspected since cycle 3 started in 2015 are mostly poles from the San Francisco Bay Area, which is the area with some of the oldest poles in PG&E's system. This explains the much higher proportion of poles between 50 and 70 years of age in cycle 3 in Figure 6. In fact, the graph for cycle 3 is only included for illustrative purposes and should not be compared directly to those for cycles 1 and 2 as it does not represent the entirety of PG&E's pole population.

Of more interest, however, is the comparison between the age profiles in cycle 1 and cycle 2. We notice a significant increase in the average age of poles between the two cycles. This focuses our attention on the relatively low replacement rates observed since the inception of the PTT program, and, in turn, poses questions about what the maximum lifespan of a pole is and what will happen to the pole population as its average age rises in the next 20 to 40 years and beyond. We explore these questions in more detail in Section 6.

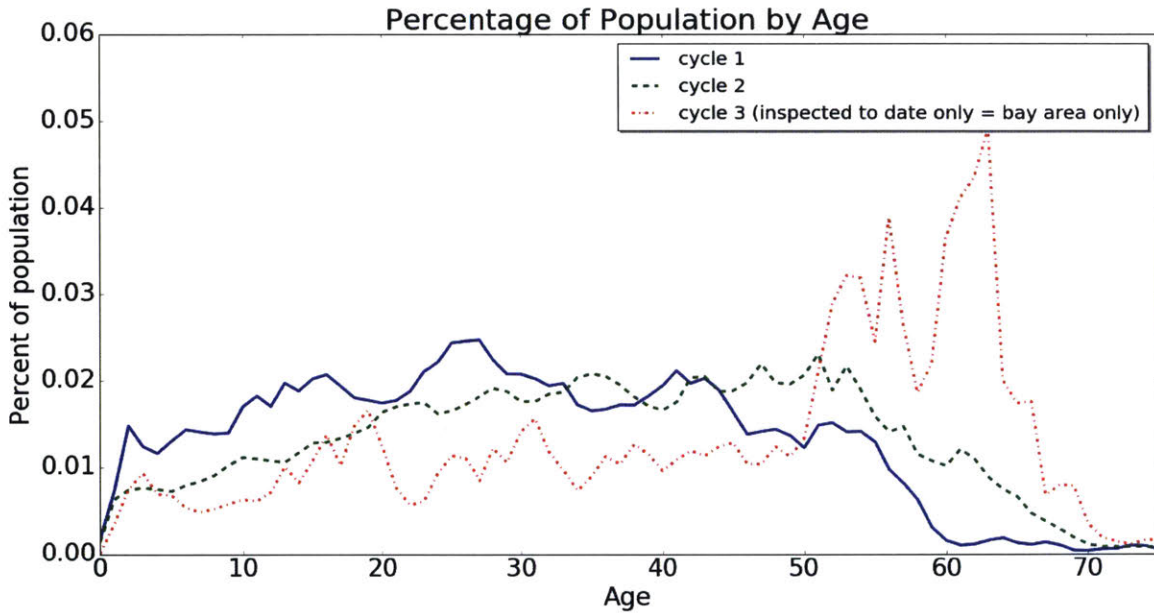


Figure 6: Age profile of pole population

3.2 Rejection rates for different pole species and chemical treatments

The graphs below illustrate the breakdown of the combination of type of species and original treatment in our data set. In Appendix E, we have included a separate graph of the rejection rates by age for each species type breaking down the percentages by original treatment in each graph. It can be seen that Douglas Fir poles have lower rejection rates than Western Red Cedar poles. Furthermore, the rejection rates of Douglas Fir poles treated with Pentachlorophenol seem to have a more clear linear relationship with age than other species-treatment combinations. A comparison between Douglas Fir and Western Pine poles also suggests that, in general, Cellon poles perform worse than both Creosote and Pentachlorophenol poles. This analysis seems to corroborate the hypothesis that Species and Original Treatment may be important variables in our model.

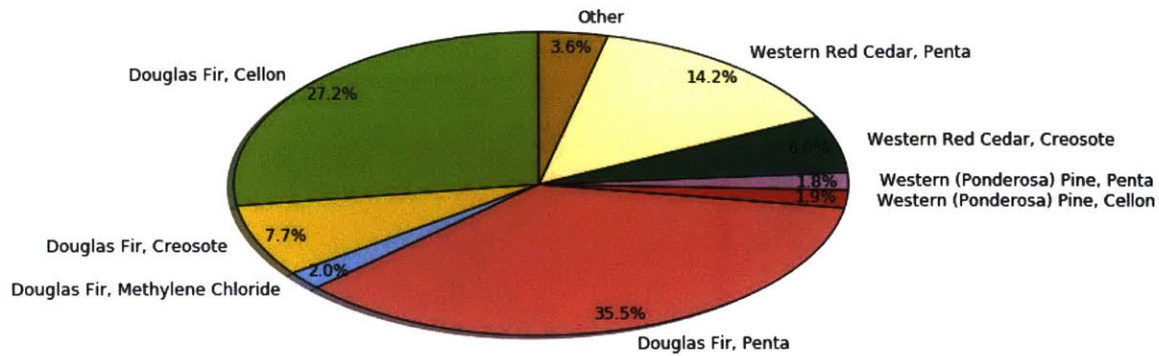


Figure 7: Distribution of poles by species and original treatment

3.3 Rejection rates in different climate zones

Another hypothesis we wanted to explore in this study was that certain weather profiles were more conducive to the formation of decay pockets in poles thereby leading to higher rejection rates. In our preliminary analysis of this hypothesis, we plotted rejection rates by age for poles in different weather conditions by weather element. For example, in the figure below we see the rejection rates for poles by mean temperature as recorded by the station closest to the pole. We have split the data set based on the 20, 40, 60, and 80th percentiles of the values for the corresponding weather element and plotted the rejection rates by age for the groups with temperature in the lower 20 percentile, the ones between the 20th and 40th, and so forth. The graph suggests that there may not be a clear relationship between mean temperatures and rejection rates in the early life of a pole. For poles above 45, however, we see that poles in warmer areas exhibit higher rejection rates.

Such graphs were produced for all weather elements in our data set. In appendix D we have included similar figures for fog and precipitation. The relationship between those weather variables and rejection rates seem less clear, but these three variables turned out to be the most important in our model for estimating rejection rates described in section 5.

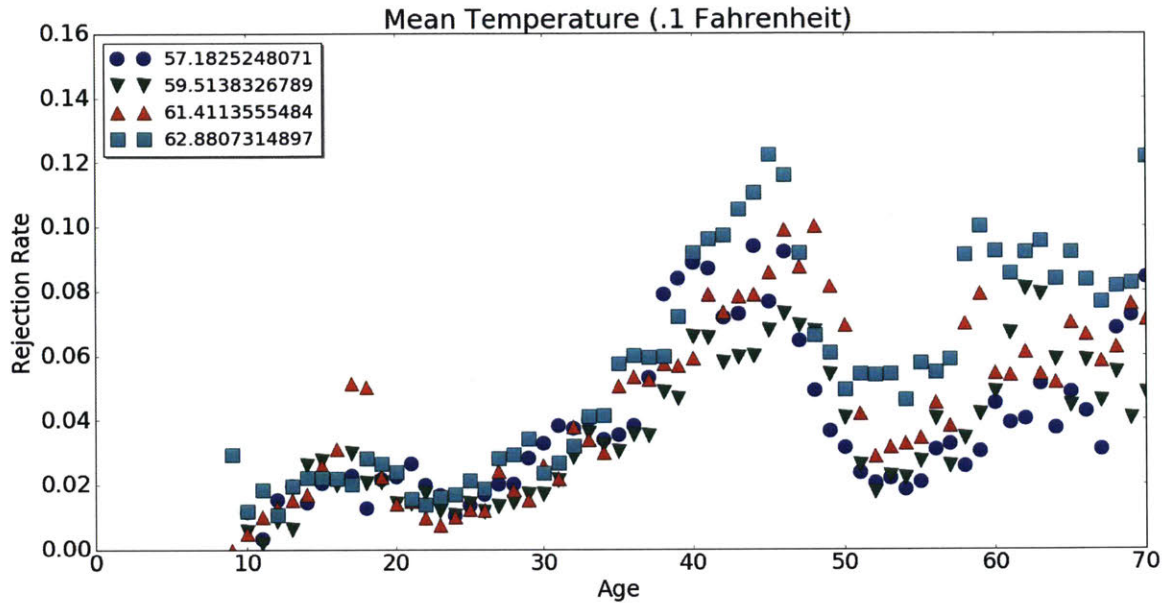


Figure 8: Rejections by Mean Temperature

3.4 Rejection rates for different land cover and soil types

Similar to the analysis based on weather variables, we plotted the rejection rates for different subsets of our data set based on the land cover and soil types associated with each pole. In the first figure below, we see that agricultural and developed urban and suburban areas experience higher rates of rejection whereas more remote areas characterized by forest and shrubland vegetation. Our hypothesis is that the high rejection rates in agricultural (cultivated) areas may be due to regular irrigation and the use of fertilizers, which likely increase the growth of fungi in the wood. In developed areas, the higher rejection rate may be due to the fact that a higher percentage of poles are partially or entirely in solid (concrete, asphalt, etc) surfaces, which limit the ability to excavate around and treat poles with preservatives. The inability to excavate around and check a pole for below ground decay also likely results in more easily rejecting poles that are suspected to have some below ground decay.

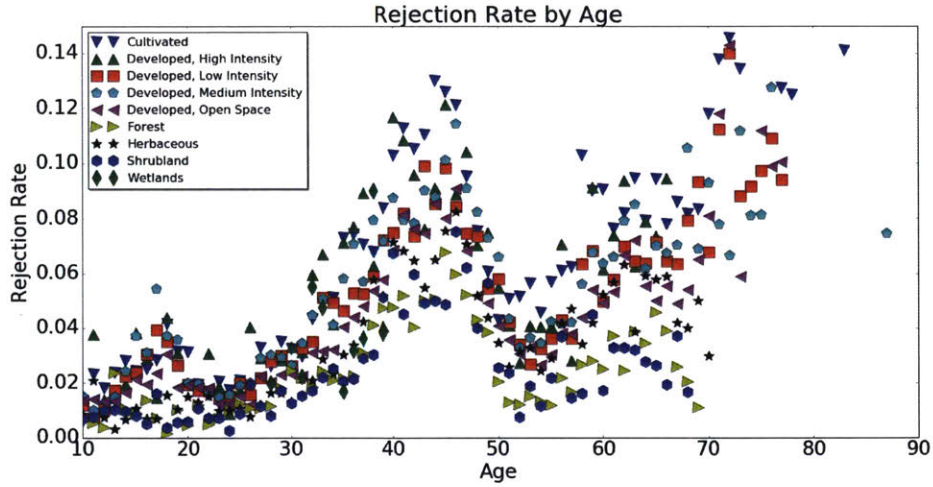


Figure 9: Rejections by age for different land cover types

Plotting rejection rates by age for different soil groups suggests that poles in clay soil demonstrate slightly higher rejection rates than others. Additionally, poles in sandy soil and bedrock often perform better. The relationship between rejection rates and soil group, however, does not seem clear across all age groups.

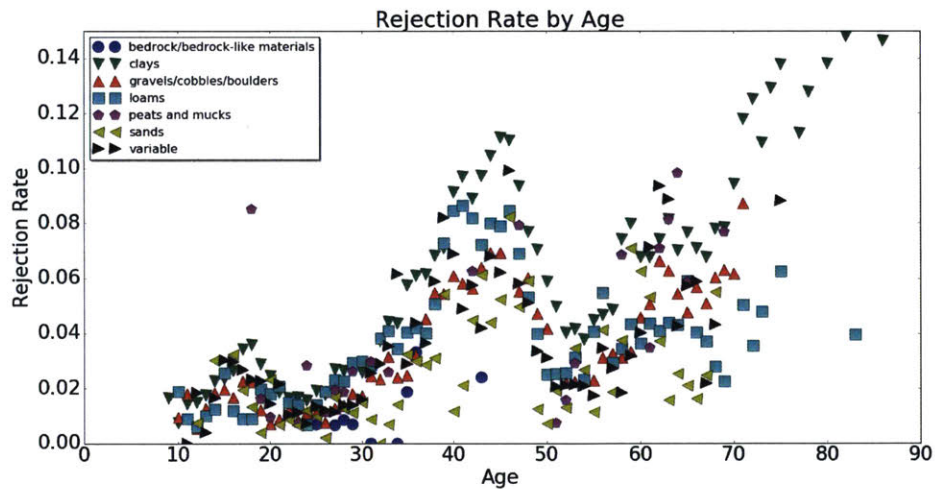


Figure 10: Rejections by age for different soil group type

4 Rejection Classification Model

4.1 Assessing classifier performance using all features

4.1.1 Motivation

The goal of the rejection classification model was to assess the ability of classification models trained on our data set to predict the results of future pole inspections. In order to provide a benchmark for any further analysis, we first built several different classification models using all features in the data set. At this stage, we were not concerned about the complexity of the model or any potential multicollinearity between the variables.

4.1.2 Classification methods used

We tested 6 different classification models: Bootstrap Aggregating (Bagging), Linear Discriminant Analysis (LDA), Decision Trees, Random Forest, Adaptive Boosting, and Logistic Regression. Following is a brief overview of each method:

- Logistic Regression is a type of generalized linear model (GLM) that models the probability that the dependent variable (Last Inspection Result in our case) belongs to a particular response class (Pass or Reject). The probability is modeled using the logistic function $P(X) = \frac{e^{f(x)}}{1+e^{f(x)}}$ where $f(x)$ is a function (usually linear) of the variables in the data set. [8, p.135]
- Linear Discriminant Analysis (LDA) models the distributions of the variables in our data set separately in each of the response classes (Pass and Reject) and then uses Bayes' theorem to calculate the probability that the dependent variable belongs to a particular response class. [8, p.138]
- Decision Trees recursively split the input space and fit models that predict the outcome of the decision variable in each split. [1, p.663]
- Bootstrap aggregating (also known as Bagging) creates a number of bootstrap samples, fits as many models using each of the bootstrap samples, and combines the results by voting. Thus, bootstrap aggregating is a meta-algorithm designed to improve the stability and accuracy of machine learning algorithms. [9, p.282]
- Random Forest is a special case of Bagging in which a number of other classifiers (decisions trees in this case) form an ensemble classifier (a forest). Each classifier is trained on a slightly modified, bootstrapped, version of the original data, and the results from all classifiers are combined to come up the prediction. In our case, as is customary, we fit the bootstrap samples using Decision Trees [8, p.319]
- Adaptive Boosting (or just Boosting) is an ensemble meta-algorithm that, similarly to Bagging, fits other classifiers (usually small decision trees) and combines their results by putting more weight on previously misclassified training samples. This effectively penalizes misclassifications by the algorithm.[9, p.10]

4.1.3 Methodology

Stubbed (reinforced) vs Unstubbed (not reinforced) poles

Because of the different criteria in rejecting electric poles, we will build and assess two different models - one for poles that have never been reinforced before and one for poles that have been reinforced before.

Poles within 20 miles from a weather station

As we explained in section 2.2.3, the majority of poles in our data set are located more than 20 miles from the closest weather station. We attempted to assess the impact of weather variables more accurately by testing the classification methods with the full data set as well as a reduced data set including only the poles that are no further than 20 miles from a weather station.

Training and testing the classification model

We used 5-fold cross validation to assess the accuracy of the model. In other words, 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 one was used to test the model against the actual results. This was performed 5 times changing the fold used for testing each time until predictions had been made for all records in the data set.

Confusion Matrix, Sensitivity, and Specificity

When we evaluate the accuracy of a classifier, it is useful to compute the confusion matrix. Table 6 shows a sample confusion matrix based on the results of a Linear Discriminant Analysis classifier on a sample of our data. We have coded Pass as 0 (or "negative condition") and Reject as 1 (or "positive condition"). There are 18,961 poles (18,718 + 243) that actually passed an inspection, of which 18,718 were classified correctly (True Negatives) and 243 were classified incorrectly as rejections (False Positives). There are 1,039 poles (881 + 158) that were rejected, of which 881 were misclassified as pass poles (False Negative) and 158 were classified correctly as rejects (True Positives). In other words, on the diagonals we have the correct classifications (True Negatives and True Positives) and on the off-diagonals we have the misclassifications (False Positives and False Negatives).

Two other metrics utilized to assess the accuracy of classifiers are Sensitivity and Specificity. Sensitivity measures the True Positive Rate or how many of the positive outcomes were predicted correctly by the statistical model. Specificity measures the True Negative Rate or how many of the total negative outcomes were classified correctly by the statistical model.

$$\text{Sensitivity} = \text{TruePositiveRate} = \frac{\text{TruePositive}}{\text{AllObservedPositive}} = \frac{TP}{TP+FN} = \frac{158}{881+158} = 0.152$$

$$\text{Specificity} = \text{TrueNegativeRate} = \frac{\text{TrueNegative}}{\text{AllObservedNegative}} = \frac{TN}{TN+FP} = \frac{18718}{18718+243} = 0.987$$

We see that the Specificity of our classifier is very high mostly due to the large number of negative outcomes (passes) in our sample set. The low Sensitivity of the classifier, however, shows that it has a hard time correctly predicting the positive outcomes (rejections).

		Predicted	
		Pass (0)	Reject (1)
Observed	Pass (0)	18718 (TN)	243 (FP)
	Reject (1)	881 (FN)	158 (TP)

Table 6: Sample confusion matrix calculated from classifier results

Cohen's kappa score

Another metric we will use to assess the accuracy of classifiers is Cohen's kappa statistic. While it is normally used to assess the agreement between two raters classifying items, it could also be used to assess the accuracy of a binary classifier. The advantage of Cohen's kappa score is that it takes into account the correct classifications that occur due to the imbalance in our data set (the fact that there are many more passing than rejected poles).

Cohen's kappa score is calculated using the formula $k = \frac{p_o - p_e}{1 - p_e}$, where p_o is the proportion of correct classifications and p_e is the probability of a correct classification by chance. In our example from table 6:

$$p_o = \frac{18718+158}{18718+243+881+158} = \frac{18876}{20000} = 0.9438$$

$$p_e = \frac{18718+243}{18718+243+881+158} \cdot \frac{18718+881}{18718+243+881+158} + \frac{881+158}{18718+243+881+158} \cdot \frac{243+158}{18718+243+881+158} = 0.94805 \cdot 0.97995 + 0.05195 \cdot 0.02005 = 0.930$$

$$k = \frac{0.9438 - 0.930}{1 - 0.930} = 0.197$$

The kappa score ranges from -1 to 1. Scores above 0.8 are indicative of good accuracy of a classifier while scores close to 0 or lower indicate performance comparable to random guessing. In the following sections, we use Specificity, Sensitivity, and Cohen's kappa statistic to assess the accuracy of different classification algorithms used on our data.

4.1.4 Results

The first two tables below describe the results of the classification models tested on poles that have not been reinforced. The first one presents the results for all unstubbed poles and the second one - only for those unstubbed poles located within 20 miles from a weather station.

In both cases the Sensitivity of the model is very low indicating the inability of the model to accurately classify pole rejections. This is also reflected in the low Kappa scores. While there is some variation in the results of the different classifiers, none of them are accurate enough to warrant the use of the model to predict pole rejections in future inspections.

We see in table 8 that all classifiers score higher when trained and testing on the reduced data set consisting only of poles close to weather stations. While the scores are still considerably low, the improvement suggests that accurate weather data increases the overall accuracy of the model.

Classifier	Cohen's kappa score	Sensitivity	Specificity
Logistic Regression	0.067	0.038	0.998
Linear Discriminant Analysis	0.145	0.104	0.991
Decision Trees	0.139	0.191	0.957
Bootstrap aggregating	0.143	0.094	0.994
Random Forest	0.118	0.071	0.997
Adaptive Boosting	0.140	0.085	0.997

Table 7: Results of classifiers for all poles that have not been reinforced

Classifier	Cohen's kappa score	Sensitivity	Specificity
Logistic Regression	0.152	0.088	0.998
Linear Discriminant Analysis	0.219	0.156	0.994
Decision Trees	0.192	0.230	0.976
Bootstrap aggregating	0.202	0.135	0.996
Random Forest	0.180	0.110	0.998
Adaptive Boosting	0.196	0.119	0.998

Table 8: Results of classifiers for poles that have not been reinforced and are located 20 miles or less from a weather station

Tables 9 and 10 display the results for stubbed poles. We see that the scores of the classifiers are low but generally better than in the case of unstubbed poles. Similar to the case of unstubbed poles, when we limit the data set to only those poles within 20 miles from a weather station, the classifiers perform better.

Classifier	Cohen's kappa score	Sensitivity	Specificity
Logistic Regression	0.122	0.094	0.987
Linear Discriminant Analysis	0.144	0.120	0.979
Decision Trees	0.206	0.362	0.847
Bootstrap aggregating	0.244	0.242	0.950
Random Forest	0.244	0.218	0.966
Adaptive Boosting	0.156	0.132	0.977

Table 9: Results of classifiers for all reinforced poles

Classifier	Cohen's kappa score	Sensitivity	Specificity
Logistic Regression	0.226	0.204	0.965
Linear Discriminant Analysis	0.272	0.277	0.944
Decision Trees	0.274	0.432	0.848
Bootstrap aggregating	0.290	0.297	0.942
Random Forest	0.245	0.227	0.960
Adaptive Boosting	0.268	0.310	0.920

Table 10: Results of classifiers for reinforced poles located 20 miles or less from a weather station

4.2 Assessing the importance of predictors

4.2.1 Motivation

In the previous section, we analyzed the performance of different classifiers on the full data set (including all available predictors). In this section, we describe the feature selection process we performed in an attempt to reduce the complexity of the model and see if the scores can be improved by only using a subset of the available variables.

4.2.2 Methodology

First, we selected the classifier that had performed the best in the tests with the full data set described in section 4.1.4: Bootstrap aggregating (Bagging). We used a Bagging classifier to measure the accuracy of models comprised of different numbers of features. The features were selected using the univariate feature selection method (SelectKBest⁵) implemented by the scikit-learn python library. Because we were to perform classification after the feature selection, we used SelectKBest along with the `f_classif`⁶ scoring function, which computes the ANOVA F-value for the provided sample.

For each possible size k of the feature set, we performed feature selection to extract the best k predictors using a (training) subset of our records. We then assessed the accuracy of our classifier on a separate (validation) subset using cross-validation and calculating the Cohen's kappa score. The separation of the test from the validation set ensured that the classifier had no knowledge of the value of the dependent variable for any records that it was scored on.

4.2.3 Results for unstubbed poles

The following graph shows how the accuracy of the classifier changes as we increase the number of variables in the feature set. Notice that, generally, as we increase the size of the feature set, the accuracy of the model increases, but the most important features are the first 20 or so.

When we reduce the set of pole records only to the poles within 20 miles from a weather station, the accuracy of the model increases. Cohen's kappa score doubles from around 0.15 to around 0.30-0.35. This, again, suggests that more accurate weather data may help us improve the accuracy

⁵http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

⁶http://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection

of the model.

Notice the Cohen's kappa scores as compared to the ones listed in Table 8. The difference is likely due to the different sizes of the test sets in the two cases.

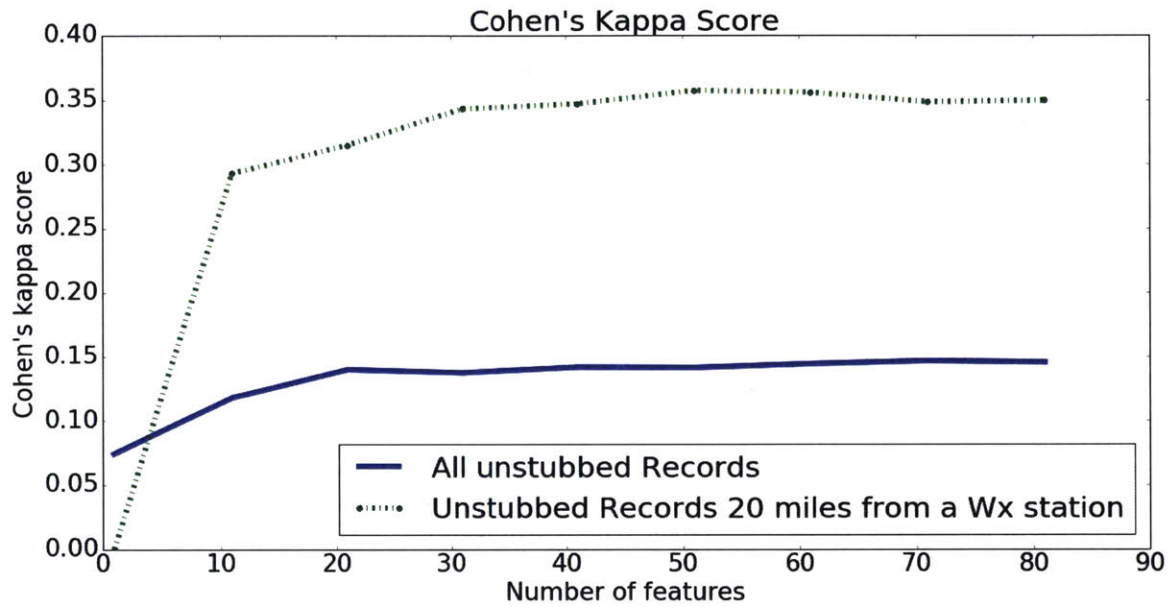


Figure 11: Accuracy of Bagging classifier for different subsets of features

The table below lists the 20 most important variables selected by the univariate feature selection method in descending order of importance. Even though the classifier does not perform well enough to be used for predictions of future inspection results in the field, it is of interest to compare the selected variables when using the full data set and the selected variables when using poles close to weather stations only. There are only two weather variables among the top twenty in the first case. In the second, there are nine. Once again, this highlights the importance of accurate and complete weather data in our model.

All poles	Poles within 20 miles of Wx Stations
Result_Status2	Num_Attachments
Species_Western_Red_Cedar	Species_Western_Red_Cedar
Species_Douglas_Fir	Result_Status2
landcover_Urban	Species_Douglas_Fir
Rmng_Shell	Supplier2_Carney
Pole_Bottom_Condition	THUNDER
YearsInService	HAIL
Rmng_Circ	THUNDER_std
Supplier2_Carney	PRCP
Num_Attachments	PRCP_std
Contractor_OSM	FOG
elevation	FOG_std
SOILGROUP_clays	TORNADO
TEMP	Contractor_OSM
Excavation	elevation
Supplier2_McFarland_Cascade_Co	RAIN_std
Pole_Top_Condition	YearsInService
Orig_Treatment2_Creosote	SOILGROUP_clays
landcover_Cultivated	landcover_Cultivated
MAX	landcover_Urban

Figure 12: Top 20 features selected in analysis of poles that have not been reinforced

4.2.4 Results for stubbed poles

Similar to the results described in the previous section, as we increase the size of the feature set, the accuracy of the model for stubbed poles generally increases. The classifier once again performs better for poles in proximity to weather stations. Cohen's kappa score increases from around 0.25 to around 0.35.

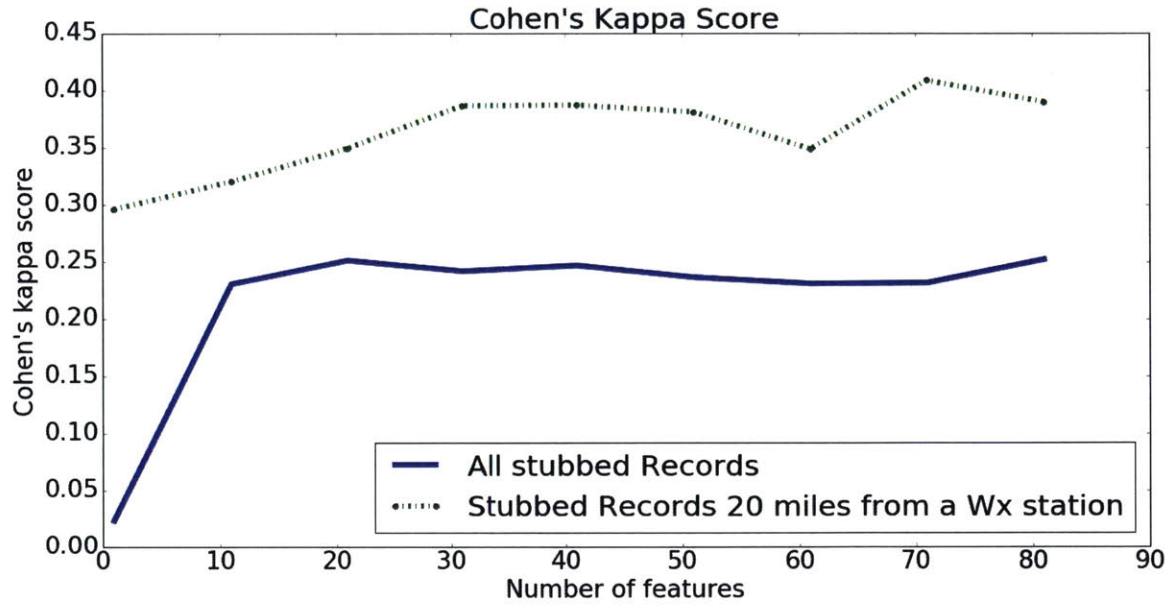


Figure 13: Accuracy of Bagging classifier for different subsets of features

In contrast to the results in the previous section, we see fewer weather variables among the top twenty best predictors when performing feature selection on poles in proximity to weather stations. However, Dew Point and Maximum Wind Speed are among the top five predictors whereas in the case of feature selection using the full data set, the best weather variable is the 9th on the list.

All poles	Poles within 20 miles of Wx Station
Rmng_Lower_Band	Num_Attachments
Result_Status2	landcover_Urban
Rmng_Upper_Band	Orig_Circ
Num_Attachments	DEWP
landcover_Urban	MXSPD
DaysBetweenInspections	Decay_Zone
Excavation	landcover_Cultivated
Decay_Zone	elevation
TEMP_std	Rmng_Lower_Band
MIN_std	Species_Douglas_Fir
TORNADO	TEMP_std
IsJointPole	MIN_std
MAX_std	Population_Density
Supplier2_Koppers_Co	Supplier2_Koppers_Co
Species_Douglas_Fir	SOILGROUP_variable
Contractor_OSM	Species_Western_Ponderosa_Pine
MXSPD_std	Height
MAX	MAX_std
SOILGROUP_variable	Result_Status2
Supplier2_Carney	Orig_Treatment2_Metal

Figure 14: Top 20 features selected in analysis of reinforced poles

4.2.5 Conclusion

The low accuracy of the classification models suggests that there is insufficient predictive power in the data we have available. For this reason, we do not attempt to further analyze the importance of the different features in our model. In fact, the listings in the tables above should NOT be taken as definitive answers to the question of variable importance. Some of the variables in our data set, especially the weather variables, demonstrate multicollinearity, and it is likely that some of them mask the importance of others. In the next section, in which we discuss a model for predicting the overall rejection rates of subpopulations of poles, we analyze the multicollinearity among the variables and discuss variable importance in the context of the rejection rate prediction/estimation model.

5 Predicting Rejection Rates

5.1 Motivation and hypothesis

The results of our classification analysis suggest that we are unable to achieve high accuracy in predicting inspection results at the pole level with the data we have available. However, we may be able to analyze the poles at an aggregate level in order to estimate the overall rejection rates for subpopulations of poles and achieve a higher accuracy model that could be used in practice.

The hypothesis behind this part of the analysis is that we can use a probabilistic model that outputs a probability of rejection for each pole and average these probabilities to come up with a reasonably accurate estimate of the expected rejection rate for the given group of poles. One of the classification methods we described in the previous section, logistic regression, is a good candidate for this. It models exactly what we are interested in - the probability that the dependent variable (Last Inspection Result) belongs to a particular category (Pass or Reject).

5.2 Methodology

In order to calculate the probability of rejection of each pole we used scikit-learn's Logistic Regression implementation. First, we standardized the features in our data set, then we fit the logistic regression model, and, lastly, calculated the probability of rejection of each pole. Once again we used cross-validation fitting the model on the training portion of our data set, and calculating the probabilities for the poles in the validation portion of the data set.

In order to assess the accuracy of our rejection rate predictive model, we compared the predicted rejection rate for each district in PG&E's area of service to the actual observed rejection rate. We chose to compare the rates at the district level because districts provide a logical grouping of poles, one that the PTT team is familiar with. Moreover, the ability to predict rejection rates at the district level is of interest to the PTT team because of the way the program operates currently - each inspection cycle poles are inspected district by district. The metric we used to assess the accuracy of our model was Mean Absolute Percentage Error (MAPE). We calculated MAPE of our predictions for each district.

In the end, we performed variable selection to simplify the model. We used the vif function in R's usdm package to calculate the variance inflation factor for the variables in our data set and estimate the multicollinearity among them. Then, we manually tested different subsets of our feature set until we found a model that performed as well as the model with all features.

5.3 Results

To provide a benchmark for all subsequent results, we started by building a model using all features in the data set. The predicted rejection rates for all districts had a MAPE of about 27%.

Then we calculated the variance inflation factor for the different variables in our model. We chose the commonly used threshold of 10 to determine whether a variable is correlated with other vari-

ables or not. In other words, any variables that had a variance inflation factor above ten were deemed highly correlated. Appendix F provides a listing of the calculated values. We can see that there is strong multicollinearity between the weather variables. This was expected especially for variables that are closely related to one another such as the daily mean, maximum, and minimum temperatures or the mean and maximum wind speeds. It is important, however, to note that there are no non-weather variables with a factor above 10. This allowed us to use the results from tables 9 and 7 to select the most important non-weather variables. In contrast, we used the results of the exploratory analysis phase and the hypotheses of experts from the PTT team to inform the decision what combinations of weather variables to include in testing the models with reduced feature sets.

The best model we obtained consisted of the following variables: External Treatment, Pole Bottom Condition, Result Status (of prior inspection), Remaining Shell, Years In Service, Stubbed (whether pole was stubbed before), Excavation, TEMP, PRCP, PRCP_std, FOG, land cover, SOIL GROUP.

Figure 15 demonstrates that for different sizes of the training set, the simplified model performs comparably to the model that includes all variables. Furthermore, a near-optimal accuracy of the model (MAPE of around 0.30) is achieved when the training set is comprised of 75,000 or more records.

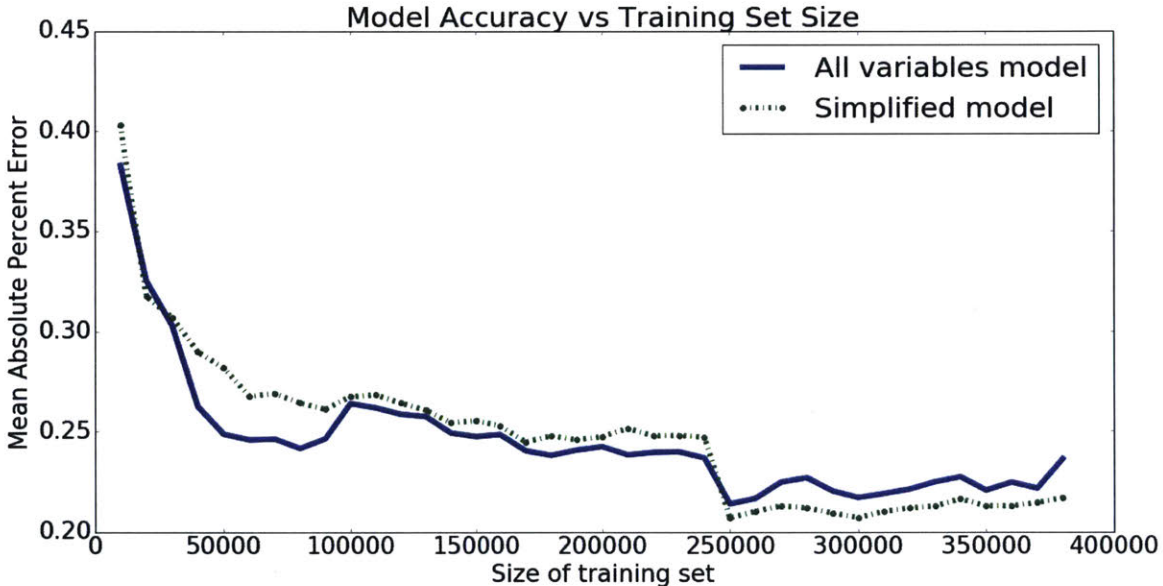


Figure 15: Model accuracy comparison between simplified model and full model

Of interest to us was what category of variables (pole inspection, weather, or geography) had the most predictive power in our model. To assess this, we created models comprised of different combinations of the three categories of variables and calculated the MAPE of their predictions. The table below shows that the weather variables explain the majority of the variance in the model, followed by the pole asset variables, and lastly the land cover and soil type variables. The cells

highlighted in yellow show the closest prediction for each district. We can see that neither combination of variables is best in all cases, but the ones that include the weather components and pole data perform the best in most cases.

District	Num Records	Actual	Pole Data	Pole + Wx	Pole + GIS	Pole + Wx + GIS
TEHAMA	25518	3.73	4.13	4.37	4.28	4.48
UKIAH	45023	3.58	4.66	4.2	4.16	3.86
SAN JOSE	49332	6.29	5.71	4.92	5.24	5.13
STOCKTON	53829	5.78	5.46	5.98	6.02	6.09
FRESNO	113832	8.34	5.43	7.19	6.2	7.28
PLACER	15534	4.13	5.22	3.97	4.29	3.58
SANTA ROSA	25318	7.82	5.46	5.22	5.24	5.37
MÖTHERLÖDE	34670	1.47	4.58	2.92	3.54	2.59
BAY	13883	4.23	5.72	3.68	5.06	3.85
SHASTA	37590	3.07	4.19	3.81	3.64	3.61
YOSEMITE	70254	3.97	5.17	5.59	5.24	5.29
DIABLO	32749	4.07	5.07	4.99	5.06	4.93
CENTRAL	9142	4.67	5.47	4.94	4.99	5.08
CUPERTINO	18326	5.44	5.31	5.12	4.81	5.15
EUREKA	8600	1.57	3.75	1.44	3.96	1.7
NORTH BAY	17037	3.15	4.05	2.98	3.68	3.09
MISSION	24840	3.59	4.85	4.1	5.24	4.39
SILVERADO	21509	4.17	4.48	4.79	4.91	4.96
GARBERVILLE	4598	2.37	3.88	3.48	2.95	2.73
FORTUNA	2506	1.68	3.63	2.09	4.09	2.21
STANISLAUS	6184	3.07	5.77	4.93	5.95	5.04
WILLOW CREEK	3056	1.51	3.54	1.48	2.28	1.2
MAPE			0.51	0.23	0.42	0.21

Figure 16: Model predictions vs actuals for a single run of the prediction model

In this part of our analysis, we also attempted to build separate models for poles that had been reinforced previously and those that had not similar to our analysis of the classification models. In the case of the predictive model for rejection rates, however, distinguishing between the two subsets of poles did not increase the accuracy of our results. In fact, the general model performed slightly better (MAPE of around 0.45) for the subpopulation of reinforced poles than the model that uses variables specific to reinforced poles such as the Remaining Shell at the level of the reinforcement bands (MAPE of around 0.50). The table below lists the results.

This was unexpected and may be because there are only about 35,000 poles in our data set that have previously been reinforced. Because of this, it is advisable to perform the analysis again in the future when the number of reinforced poles increases. (Remember that, as described in Section 2.2, our data set consists only of those pole records with valid data that we have at least two consecutive inspection records for.)

Model	All poles	Unstubbed	Stubbed
All variables	0.253	0.259	0.353
Pole asset features only	0.489	0.545	1.013
Pole asset features + Weather	0.252	0.287	0.683
Pole asset features + land cover and soil type	0.421	0.440	0.932
Pole asset features + Weather + land cover and soil type	0.249	0.255	0.584
Stubbed pole features only ⁷	-	-	1.089
Stubbed pole features + Weather ⁷	-	-	0.794
Stubbed pole features + land cover and soil type ⁷	-	-	0.976
Stubbed pole features + Weather + land cover and soil type ⁷	-	-	0.660

Table 11: Mean Absolute Percent Error (MAPE)

6 Pole Population Aging Simulation Model

6.1 Motivation

In the previous sections, we saw that the age (or years in service) of a pole is an important factor in determining the probability of rejection and the need of replacement. Few utility poles survive past 100 years of age. In section 3 we saw that the average age of the poles in PG&E's pole population is about 40 years. The company currently inspects poles every 10 years. If we make the simplifying assumptions that the maximum age of a pole is 100 years and the oldest poles will be the ones that are replaced, we would need to replace roughly about 10% of our poles every inspection cycle in order to maintain the average age of our pole population. During the last two inspection cycles, the replacements resulting from PTT inspections has been around 1%. There are other sources of pole replacements, however, and according to data from PG&E's Tangible Property List, there are about 206,82 poles replaced per year between 2005 and 2014. This is equivalent to 206,820 poles per inspection cycle or a replacement rate of about 8.6% per inspection cycle. Looking back at the age profiles for poles inspected in cycle 2 and cycle 3 in figure 6, we see

⁷Model trained with 10,000 records and validated against 25,000 records

that there is an observable increase in the average age of poles from one inspection cycle to the next.

The ability to predict the rate at which pole assets will have to be reinforced and replaced in the future is of great importance to PG&E. It allows to make an accurate estimate of the budget and staffing required. Furthermore, estimating rejection rates in the more distant future can help the company foresee some of the challenges it may be presented with and plan for them accordingly. In this section, we describe a simulation model we built for this purpose and an analysis of its results.

6.2 Methodology

In order to estimate the reinforcement and replacement rates in the future, we simulate successive inspection cycles by adding 10 years to the age variable, and each time we rerun the logistic regression model described in the previous section to estimate the rejection rates we expect to see in that cycle. Because separating stubbed from unstubbed poles did not result in a more accurate prediction model as we explained in the previous section, we use the same logistic regression model for all poles. In the future, as more data is gathered on the performance of stubbed poles, the statistical analysis can be repeated and a separate regression model for stubbed poles could be developed.

We use the results of the logistic regression (logit) model on each iteration of the simulation to determine what percentage of poles to reinforce and replace as well as which poles are reinforced and replaced during that cycle. To do so, we make the simplifying assumption that the poles with the highest probabilities of reinforcement and replacement returned by the logit model are the ones that are reinforced and replaced. In the previous sections we saw that in reality this was often not the case, but this is our most accurate possible prediction.

Additionally, we choose to "replace" all poles that reach 100 years. The relatively low average age of our pole population means that we have not seen the majority of poles reach their maximum lifespan. However, few poles survive to become 100 years of age.

As described earlier, a number of poles are replaced for reasons other than wood decay. For example, poles may have to be replaced if additional equipment is to be installed and the existing cannot handle the load or if distribution lines are to be put underground. In our simulation, we assume that the number of pole replacements due to other reasons is constant each inspection cycle, and that it is equal to the number of such replacements performed during the most recent full inspection cycle - 7.6% of the pole population. In other words, in the simulation, we randomly replace 7.6% of our poles each cycle.

To replace and reinforce the poles in our simulation, we update their respective attributes. In the case of replacement, we set the age to 0, remaining shell to 100%, the pole bottom condition to "good", and reset the indicators for Excavation, External Treatment and result of the prior inspection. In the case of reinforcement, we set the indicator of whether a pole has been stubbed to 1. Additionally, for all stubbed and unstubbed poles we estimate the values of their attributes (Remaining Shell, Pole Bottom Condition, Excavation, External Treatment) based on their updated age, whether they have been stubbed, and the average value of the respective attribute for poles of that age from the distribution of the variable in our current population. For example, if a pole's

updated age in an iteration of the simulation is 35 and it has not yet been stubbed, we update its remaining shell to the average remaining shell among the poles of age 35 that have not yet been stubbed in our training set.

It is clear that this model has significant limitations. Most importantly, we are unable to predict, so we assume that the poles will experience the same weather conditions in the future. Since the weather variables in our data set are averages over many years, thereby representing general weather patterns rather than specific weather conditions, we believe that this assumption is reasonable. Another limitation is that when we estimate that values of variables such as Remaining Shell, we do not take into account variables other than age and whether a pole has been stubbed or not. This choice was made in the interest of simplicity, but we can imagine a model in which these estimates are made by separate regression models for each attribute.

6.3 Results

The figure below shows the results of the simulation. We can see that the reinforcement and replacement rates rise in the next 30-40 years but not at a level that can prevent the average age of our pole population from reaching upwards of 80 years. According to the model, in about 40-50 years we can expect a large percentage of our poles to reach the assumed maximum age of 100 years, which will lead to an increase in replacement rates.

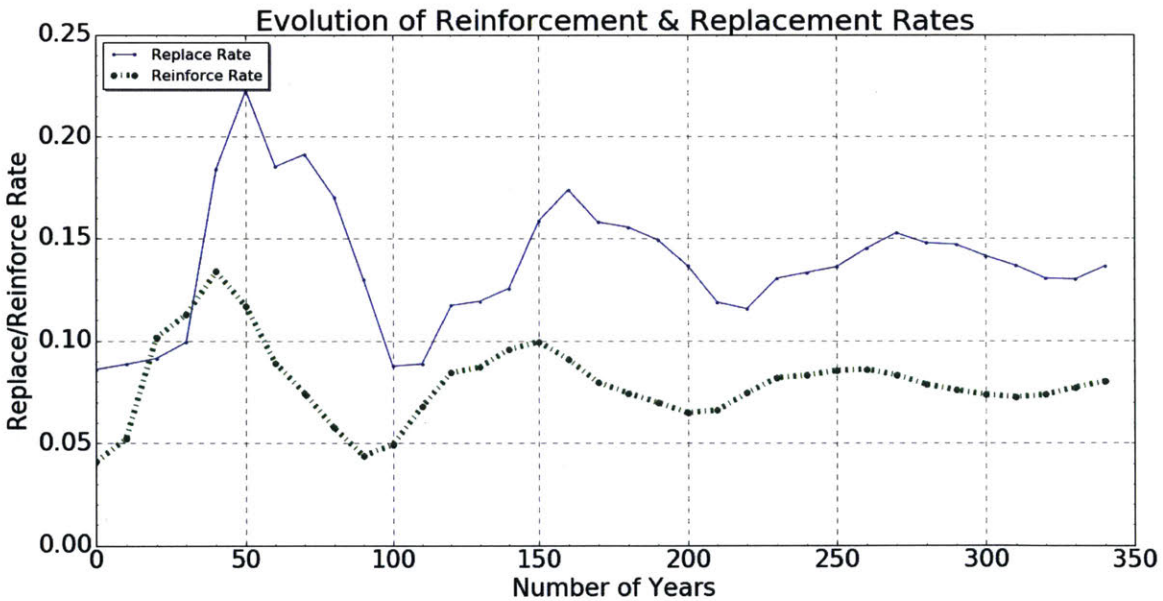


Figure 17: Simulation of rejection rates as pole population is aged

As we mentioned in the previous section, our model for estimating rejection rates has a MAPE of about 30% when using it to predict the rate of reinforcement and about 60% in the case of replace-

ments. Figure 18 below helps visualize the impact of this error. Three separate curves are drawn for both the reinforcement and replacement rates representing the original estimate along with the upper and lower-bound estimates based on the MAPE of the logistic regression models. Notice that the prediction errors have a bigger impact on the estimates of the rate of reinforcements even though the percentage change in them is smaller than that for replacements. This is because a large portion of the pole replacements in our simulation is driven by the assumption that any pole that reaches a maximum age of 100 years is replaced. This highlights the importance of the age factor. In other words, the growth in the replacement rates is largely due to poles reaching their maximum age.



Figure 18: Simulation of rejection rates as pole population is aged

The results of the simulation model suggest that unless we replace poles proactively, we will experience cyclical increases in replacement rates. As the average age of poles reaches their maximum lifespan, a large number of poles will be replaced, thereby lowering the average age significantly. This, in turn, will lead to lower rejection rates until the mean age of the population rises again and the cycle repeats, albeit at a smaller scale (as we can see in figure 18, the peaks in the replacement and reinforcement rates flatten).

Figures 19, 20, 21 below show what occurs if we replace poles proactively at rates of 10%, 11%, and 12% respectively. Note that this is only 2-4 percentage points higher than the current replacement rates (including PTT replacements and replacements from other jobs). We see that increasing the replacement rates proactively can help us reduce the first peak in Figure 18 from approximately 22% to 14-15%. Furthermore, we notice that increasing the replacement rate beyond 10% has a relatively small impact on the first replacement rate peak (bringing it down from 15 to about 14%) and virtually no impact on the second and third peaks (the peaks are spread across fewer inspection

cycles i.e. years, but the peak value remains approximately 12%). Comparing the three figures, we can also better visualize the trade-off between replacements and reinforcements. Keeping the replacement rates above 10% results in higher numbers of replacements, which lowers the rate of reinforcements. Because replacements are much more costly than reinforcements, however, it may not be cost-effective to maintain replacement rates higher than 10%.

In summary, proactively replacing poles at a rate of 10% per cycle could help us reduce the first replacement rates peak from approximately 22% to 14-15%. Increasing the replacement rates further seems imprudent.

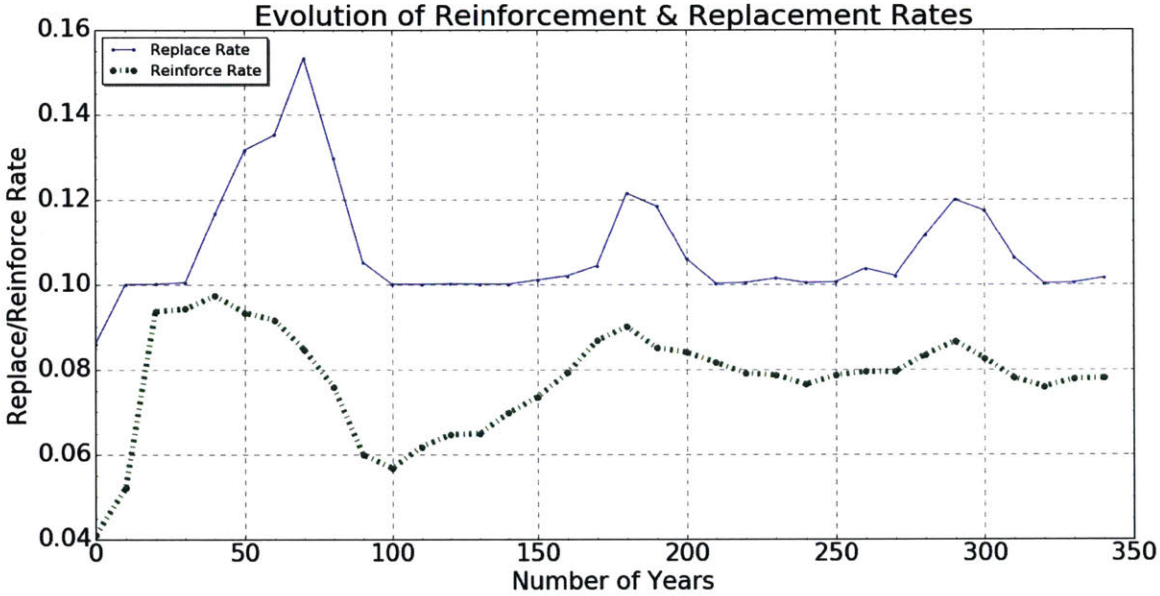


Figure 19: Simulation of rejection rates forcing a 10% replacement rate

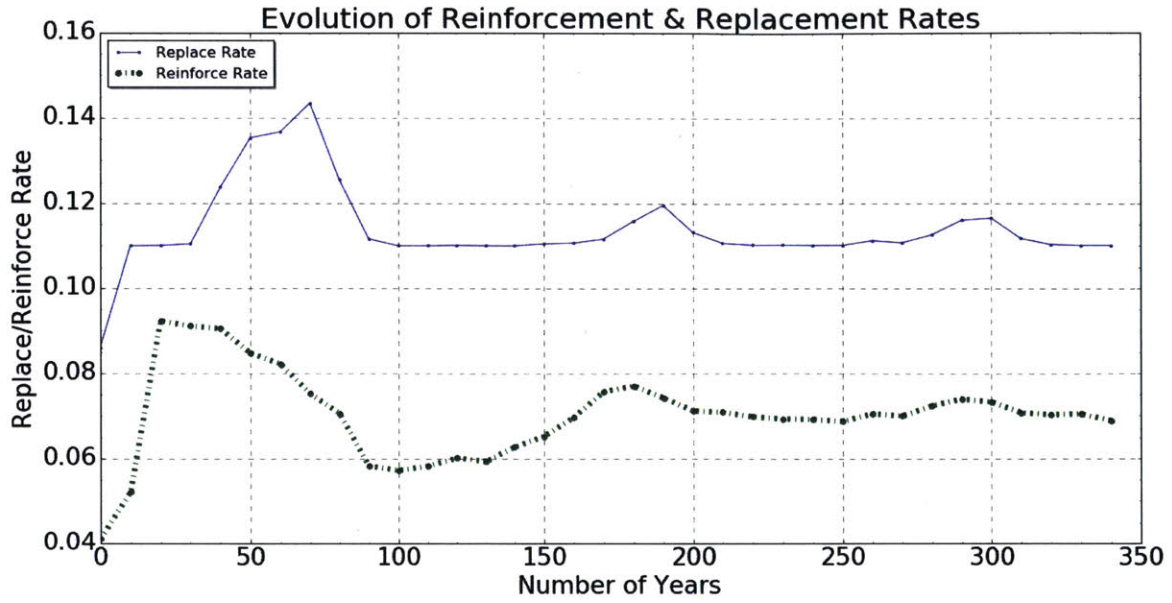


Figure 20: Simulation of rejection rates forcing a 11% replacement rate

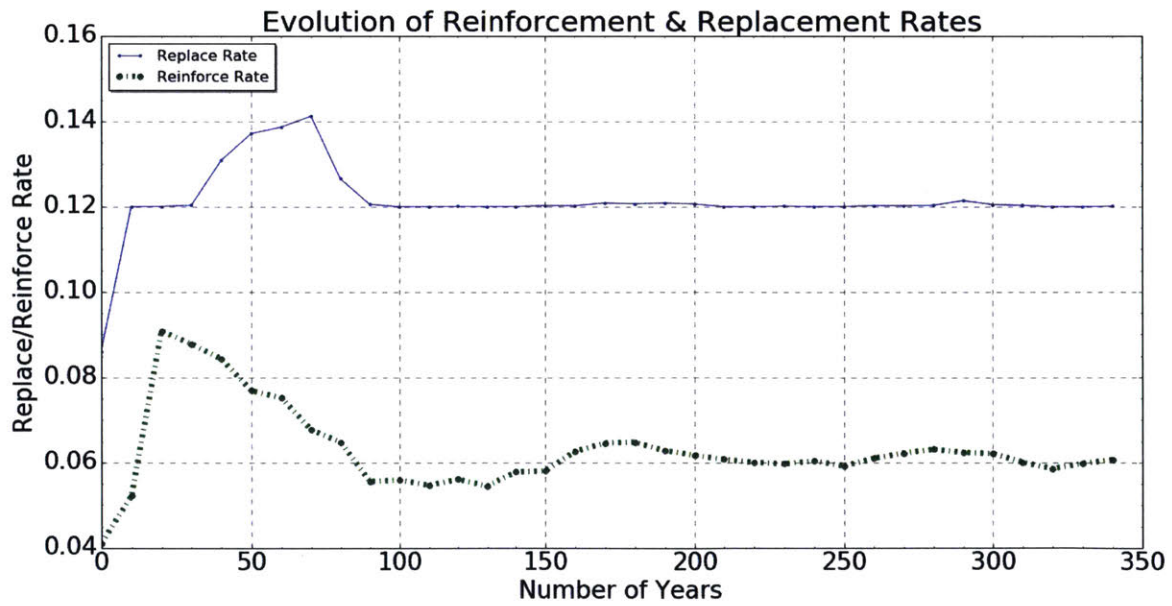


Figure 21: Simulation of rejection rates forcing a 12% replacement rate

7 Conclusions and Future Work

7.1 Improving data quality to facilitate future analyses

In the beginning of this research project, we set out to build a statistical model that takes pole inspection records and environmental information and calculates a probability of rejection for each pole. The analysis described in section 4 showed us that it seems impossible to obtain a model with such accuracy with the data we have available. We can only conjecture whether more precise inspection measurements and notes or more detailed weather data can help us obtain such results. There are numerous external factors that make such a prediction very difficult. A few classic examples, frequently mentioned by the PTT team, are the presence of woodpeckers, the proximity to lawns that are irrigated often, which speeds up the wood decay process, and mechanical damage caused by vehicles or other sources.

Even though it is difficult to predict exactly what poles will not meet safety requirements in future inspections, in section 5, we saw that it is possible to estimate the overall rejection rate in the pole population or in a large enough subpopulation of poles. We used a logistic regression model to estimate the probability of rejection of each pole based on past data. While these probabilities were not accurate at the individual pole level, averaging them across a subpopulation of poles gave us a reasonable estimate of the overall rejection rate for that subpopulation. In particular, our estimates of the rejection rates in all PG&E districts had a MAPE of 30% compared to the observed rejection rates.

Obtaining a fairly accurate estimate of the overall rejection rates suggests that there is predictive power in the data. As PG&E further develops its ability to use modern statistical techniques to inform operational and other business decisions, it is essential to keep improving the accuracy of available data. First, processes could be improved so that the measurements taken during inspections are as precise as possible. This could be done by using better equipment (when justified financially) to measure shell thickness and internal decay for example. It can also be accomplished by educating and encouraging personnel to measure pole circumference and existing mechanical damage more precisely. Second, data collection and storage mechanisms can be designed with more data integrity checks built in. This may create some additional work, but will ensure higher accuracy of the data in the long run. Furthermore, it will force process changes to be implemented and documented more carefully because each significant process change will require revisiting the design and checks implemented in the data storage mechanisms. Third, performing different statistical analyses more frequently will ensure that errors and inconsistencies in the data are found more quickly. As long as there is a will to rectify such errors, the integrity of PG&E's data will inevitably increase. In particular, it is of interest to repeat the analysis described in this research project in the future as more data is collected and as weather patterns in Central and Northern California change. Other analyses focusing on particular variables in the data such as shell thickness changes over time or the effects of mechanical damage, woodpeckers, insects, or the positioning of poles next to irrigated lawns or cultivated land can inform decisions on what additional information should be gathered during inspections.

In order to make statistical analyses in the future easier, it may be of interest for PG&E to create an informal, easily accessible, and editable set of documentation for asset classes of interest. This way, anyone who would like to perform statistical analysis for an asset type can quickly understand

what data is available for the given asset and what some of the characteristics and idiosyncrasies of the data are. For example, such documentation can describe the change in processes and rejection criteria for poles throughout the years. It could also mention that a data migration effort in the past has resulted in tens of thousands of records with an installation year of 1900, which should not be used in the calculation of the actual age of these poles.

This documentation should be informal and editable (such as a wiki) so that anyone working with the data set for the given asset could improve it by adding more information or explanations. It should reference official, static PG&E documents that provide detailed explanation of inspection processes. The documentation should also be easily accessible, so that it can be used by both the PTT team and the personnel performing statistical analysis using the data. Such documentation will only provide added value to the company if it is used frequently. PG&E's strong recent efforts in leveraging asset information to measure asset health and risk suggests that this will be the case.

7.2 Using results of simulation model to inform operational decisions

The simulation model described in section 6 presented a rough order of magnitude estimate of the growth in the reinforcement and replacement rates of poles in the future. The analysis highlights the issue of aging infrastructure, which is exacerbated by the fact that the distribution of the age of poles is not uniform. The majority of poles in PG&E's area of service are between 35 and 55 years of age. This will likely lead to significant increases in the rejection rates of poles in the next 30 to 50 years. The pole reinforcement and replacement programs will therefore likely experience big increases in the amount of work in the following decades. The impact of such increases is two-fold. First, it will increase the budget required for the pole maintenance program. Second, it will require a larger number of reinforcements and replacements to be performed per day, which, in turn, will require a larger number of qualified personnel to perform the work.

PG&E already relies on subcontractors for the majority of the work performed for its pole maintenance program. This gives it more flexibility to deploy more human resources when the amount of maintenance work increases. It is still essential, however, to estimate whether enough manpower will be available to perform the larger number of reinforcements and replacements in the future. It is important to monitor the rise of rejection rates each year, rerun the analysis described in this paper in order to adjust the estimates based on new information, and carefully plan for the increased work.

The estimates resulting from the simulation model are higher than what other utilities in the industry have experienced so far. That being said, even a more modest increase in replacement rates than the one our simulation model projects can lead to significant increases in the work and man hours required to perform pole replacements. If PG&E determines it cannot easily handle such an increase, the issue could be mitigated by smoothing the age distribution of its pole assets by proactively replacing poles that may not have reached their maximum age. This is a difficult decision to make, but one that could prevent operational difficulties in the future.

A Appendix: Acronyms Used

Abbreviation	Meaning
CPUC	California Public Utilities Commission
GLM	Generalized Linear Model
MAPE	Mean Absolute Percentage Error
LDA	Linear Discriminant Analysis (a classification algorithm)
NCDC	National Climatic Data Center's
NLCD	National Landcover Dataset
NRCS	Natural Resources Conservation Service
PG&E	Pacific Gas and Electric Company
PTT	Pole Test and Treat (Pole asset management program)

Table 12: Abbreviations used

B Appendix: Variables in Original Data Set

B.1 Pole Asset Information

Data column	Description
Class	Pole Class (classification)
Height	Height of the pole
Orig Circ	Original circumference
Orig Treatment	Original chemical treatment performed during manufacture to ensure the durability of the pole
Mfr Year	Manufacture year
Supplier	Supplier company name
Ins Year	Installation year
Pole Type	Type of pole (e.g. wood, steel, guy pole, etc)
Species	Wood species type (e.g. Douglas Fir, Western Red Cedar, Western Pine, Lodgepole Pine)
Surface	A list of the types of surface surrounding a pole (e.g. concrete, asphalt, grass, brick, gravel)

Table 13: Original pole asset information

B.2 Pole Inspection Records

Data column	Description
Visit Date	Date of the inspection
Existing Reinforcement	Type of existing reinforcement
Attachment	A list of the pieces of equipment installed on a pole (Distribution equipment, Cable TV, Communications, etc)
Excavation	Whether an excavation was performed and if so what type - partial or full
External Treatment	Whether below ground preservative was applied to the pole
Internal Treatment	Whether a pole was fumigated
Internal Test	What type of test for internal decay was performed (e.g. visual only, visual and boring)
Pole Top Condition	Qualitative assessment (bad, fair, good) of the condition of the top part of the pole
Pole Bottom Condition	Qualitative assessment (bad, fair, good) of the condition of the bottom part of the pole
Steel Installed	Field used in stubbing projects specifying whether a steel truss/stub was installed
Groundline Shell Avg	Average shell thickness at ground line
Below GL Shell Avg	Average shell thickness below ground
Shell at 15av	Average shell thickness at 15" above ground
Shell at 26av	Average shell thickness at 26" above ground
Shell at 42av	Average shell thickness at 42" above ground
Shell at 54av	Average shell thickness at 54" above ground
Shell at 66av	Average shell thickness at 66" above ground
Current Circ	Current circumference
Effective Circ	Current circumference accounting for any external damage
Pole Load	Percent at which the pole is loaded (e.g. 100% means a pole is loaded up to capacity)
Wood Strength	A calculated field specifying the strength of the pole. In the past, this was estimated using the values for ground line shell thickness and current circumference. Since 2014 a proprietary formula unknown to PG&E has been used.
Rmng Strength	Remaining strength equal to wood strength divided by pole loading (recorded as a percentage)
Result Status	Result of the inspection (Pass, Reinforce, Replace)
Immediate Response Conditions	

Table 14: Original pole inspection information

B.3 Appendix: Land cover classes

The descriptions below are from the National Land Cover Database 2011 classification available at the Multi-Resolution Land Characteristic Consortium (MRLC)⁸. A full listing of land cover types and the accompanying descriptions can be found on the website.

Class	Land cover type and Description
Water	Open Water - areas of open water, generally with less than 25% cover of vegetation or soil.
Developed	Developed, Open Space - areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes.
	Developed, Low Intensity - areas with a mixture of constructed materials and vegetation. These areas most commonly include single-family housing units.
	Developed, Medium Intensity - areas with a mixture of constructed materials and vegetation. These areas most commonly include single-family housing units.
	Developed High Intensity - highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses and commercial/industrial.
Barren	Barren Land (Rock/Sand/Clay) - areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits and other accumulations of earthen material. Generally, vegetation accounts for less than 15% of total cover.
Forest	Deciduous Forest - areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species shed foliage simultaneously in response to seasonal change.
	Evergreen Forest - areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species maintain their leaves all year. Canopy is never without green foliage.
	Mixed Forest - areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. Neither deciduous nor evergreen species are greater than 75% of total tree cover
Shrubland	Shrub/Scrub - areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation.
Herbaceous	Grassland/Herbaceous - areas dominated by graminoid or herbaceous vegetation. These areas are not subject to intensive management, but can be utilized for grazing.
Planted / Cultivated	Pasture/Hay - areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle.

Table 15: Land cover classification

⁸http://www.mrlc.gov/nlcd11_leg.php

Class	Land cover type and Description
	Cultivated Crops - areas used for the production of annual crops. Crop vegetation accounts for greater than 20% of total vegetation.
Wetlands	Woody Wetlands - areas where forest or shrubland vegetation accounts for greater than 20% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.
	Emergent Herbaceous Wetlands - Areas where perennial herbaceous vegetation accounts for greater than 80% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

Table 16: Land cover classification

B.4 Appendix: Soil Group Types

The descriptions below are adapted from various online sources. More detailed descriptions of the soil types that comprise each soil group can be found on the Natural Resources Conservation Service's website⁹ or the State Soil Geographic (STATSG0) Data Base User Guide¹⁰.

Class	Description
Clays	Fine-grained natural rock or soil material that combines one or more clay minerals with traces of metal oxides and organic matter. Clays are plastic due to their water content.
Loams	Soil composed mostly of sand, silt, and a smaller amount of clay. Most commonly, its composition is about 40%-40%-20% concentration of sand-silt-clay.
gravels/ cobbles/ boulders	Unconsolidated rock fragments that have a general particle size range and include size classes from granule- to boulder-sized fragments
Variable	Variable soils
Sands	Sandy soils
bedrock/ bedrock-like materials	Consolidated (solid and tightly bound) undisturbed rock usually located beneath a surface layer of soil or other material
Peats and mucks	Soils made up primarily of organic materials. Peats most commonly refers to an accumulation of partially decayed vegetation or organic matter. Mucks is soil made up primarily of humus from drained swampland.

Table 17: Soil classification

⁹<http://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/survey/class/>

¹⁰http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053629

C Appendix: Weather Stations



Figure 22: Location of weather stations used in analysis

D Appendix: Rejections by Age for Different Weather Profiles

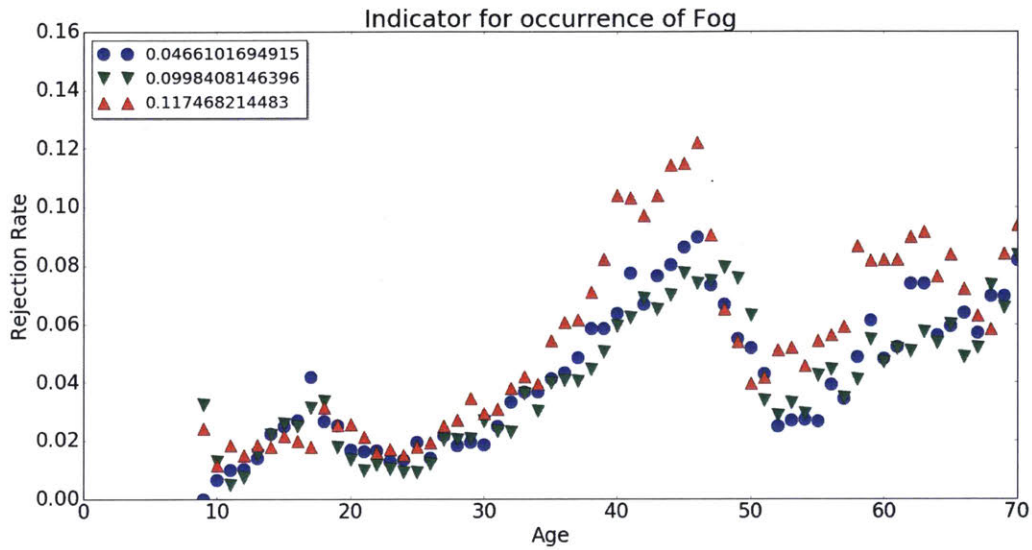


Figure 23: Rejections by Age and Average Occurrence of Fog

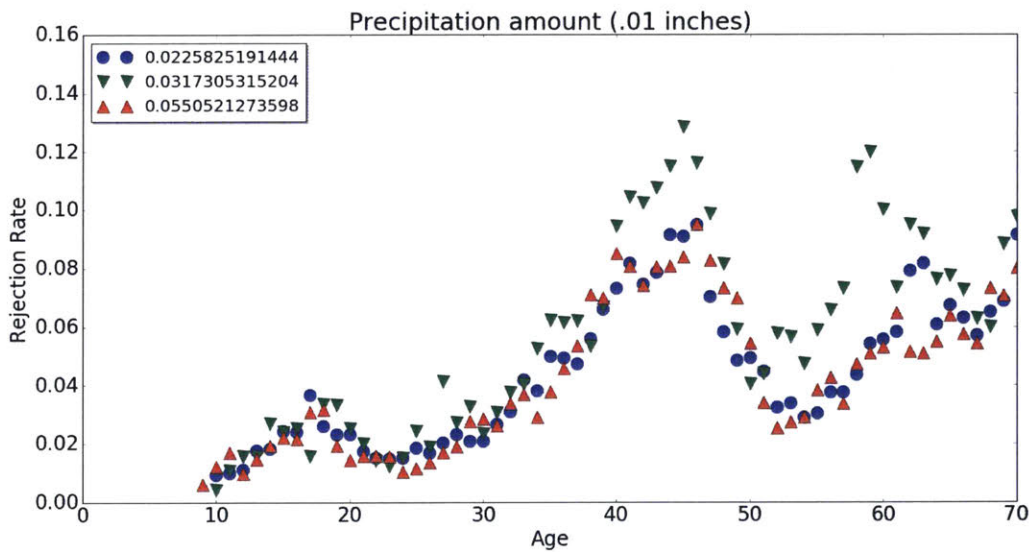


Figure 24: Rejections by Age and Average Precipitation Amount

E Appendix: Rejections by Age for Different Species and Original Treatments

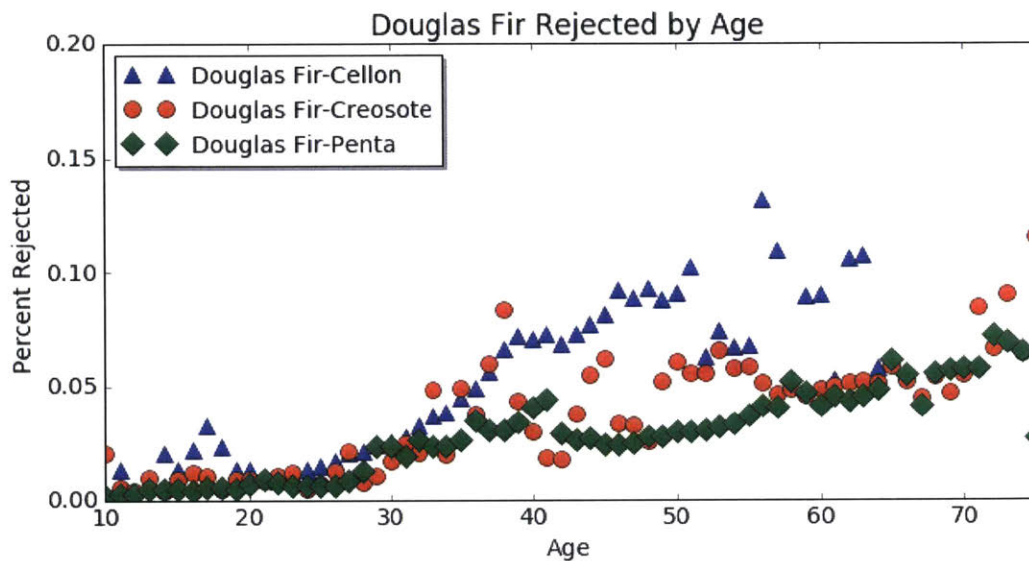


Figure 25: Douglas Fir rejection rates by age

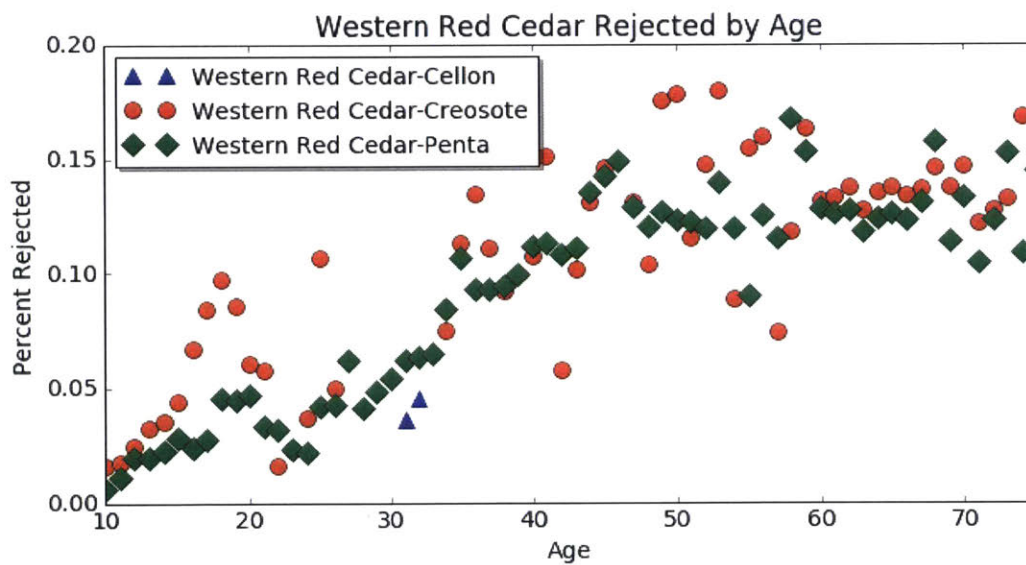


Figure 26: Western Cedar rejection rates by age

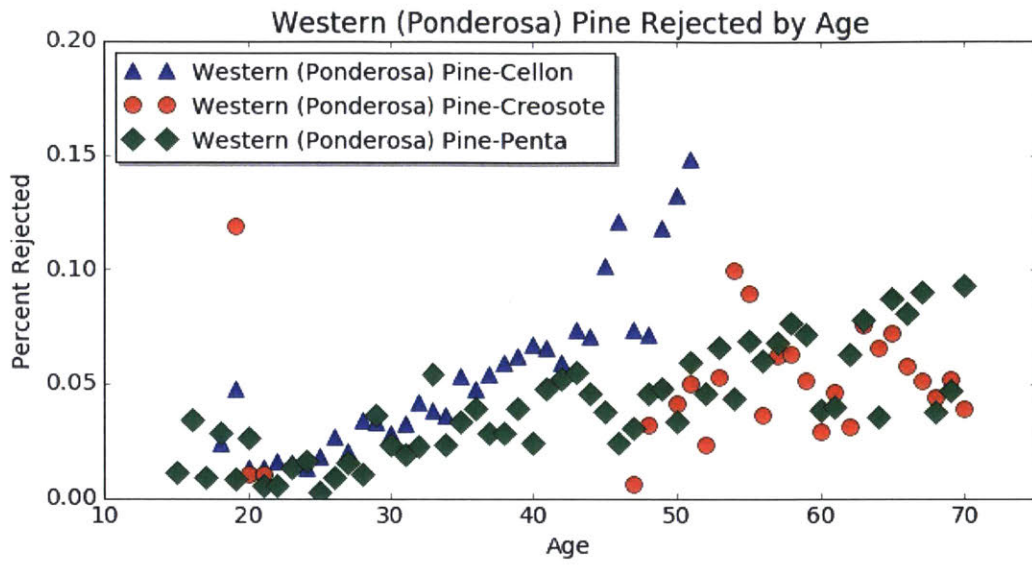


Figure 27: Western Pine rejection rates by age

F Appendix: Assessing Multicollinearity using Variance Inflation Factor

Variables	VIF
TEMP_std	354.176
MAX_std	268.826
TEMP	98.885
MAX	96.484
FOG_std	85.427
RAIN_std	73.166
PRCP	58.240
PRCP_std	53.764
THUNDER_std	51.946
MIN_std	43.457
RAIN	37.552
THUNDER	35.572
FOG	23.530
MXSPD	22.419
WDSP	21.250
MIN	16.658
WDSP_std	16.264
MXSPD_std	11.784
GUST	8.018
Decay_Zone	7.513
External_Treatment	6.658
Excavation	5.606
TORNADO	5.329
Result_Status2	4.742
DEWP	4.676
Pole_Top_Condition	3.858
Stubbed	3.709
Orig_Circ	3.129
SNOW	2.971
Height	2.884
Num_Attachments	2.743
YearsInService	2.671348343
elevation	2.562445457
HAIL	2.503530809
Population_Density	2.432180106
DaysBetweenInspections	2.26422012
Rmng_Shell	1.604404365
IsJointPole	1.482076853
Rmng_Circ	1.398388317
Immediate_Response_Conditions	1.13397553

Figure 28: Variance inflation factor of variables in data set

References

- [1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer.
- [2] Adam Christopher Chao. “Anomaly Detection For Natural Gas Regulator Stations”. MA thesis. Massachusetts Institute of Technology, 2016.
- [3] Gregory D. Eschelbach. “Wires-Down Predictive Modeling and Preventative Measures Optimization”. MA thesis. Massachusetts Institute of Technology, 2016.
- [4] T. Blanc K. Pierson. “The Operational Determinants of Utility Pole Decay and Optimal Replacement in the Pacific Northwest”. In: *IEEE Transactions on Power Delivery IEEE Trans. Power Delivery Power Delivery, IEEE Transactions* (Oct. 2016).
- [5] Lillian Ruth Meyer. “Predicting Corrosion on Protected Buried Steel Natural Gas Distribution Pipelines”. MA thesis. Massachusetts Institute of Technology, 2016.
- [6] Jeffrey J. Morrell. *Wood Pole Maintenance Manual: 2012 edition*. Tech. rep. Oregon State University, 2012.
- [7] Seth Guikema Seung-Ryong Han David Rosowsky. “Integrating Models and Data to Estimate the Structural Reliability of Utility Poles During Hurricanes”. In: *Risk Analysis: An International Journal* 34 (June 2014), pp. 1079–1094.
- [8] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer.
- [9] Jerome Friedman Trevor Hastie Robert Tibshirani. *Elements of Statistical Learning*.