Characterization of Ambient Noise

THESIS

Rachel C. Ramirez, Maj, USAF

AFIT-ENS-MS-18-M-155

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-18-M-155

CHARACTERIZATION OF AMBIENT NOISE

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Rachel C. Ramirez, M.S.

Maj, USAF

22 March 2018

AFIT-ENS-MS-18-M-155

CHARACTERIZATION OF AMBIENT NOISE

THESIS

Rachel C. Ramirez, M.S.
Maj, USAF

Committee Membership:

Dr. R. R. Hill,
Chair

Dr. D. K. Ahner,
Reader

AFIT-ENS-MS-18-M-155

Abstract

An Air Force sponsor is interested in improving an acoustic detection model by providing better estimates on how to characterize the background noise of various environments. This would inform decision makers on the probability of acoustic detection of different systems of interest given different levels of noise. Data mining and statistical learning techniques are applied to a National Park Service acoustic summary data set to find overall trends over varying environments. Linear regression, conditional inference trees, and random forest techniques are discussed. Findings indicate only sixteen geospatial variables at different resolutions are necessary to characterize the first ten $\frac{1}{3}$ octave band frequencies of the L_{90} band using just the linear regression. The accuracy of the regression model is within 2 to 6 decibels and depends on the frequency of interest. This research is the first of its kind to apply multiple linear regression and a conditional inference tree to the national park service acoustic dataset for insights on predicting noise levels with dramatically less variables than needed in random forest algorithms. Recommended next steps are to supplement the national park service dataset with more geographic information system variables in common global databases, not unique to the United States.

For my husband, thank you for supporting me

Acknowledgements

I am grateful for the patience, guidance, and mentoring provided by my faculty research advisor, Dr. Raymond Hill. My learning and growth under his tutelage have been invaluable. I would also like to thank AFIT professor Dr. Mbonimpa for assistance with ArcGIS.

Rachel C. Ramirez

Contents

Page

Page

List of Figures

List of Tables

I. Introduction

Problem Statement

An Air Force sponsor is interested in further developing an operational model to more accurately predict the values of ambient background noise in varying outdoor environments. The value of an improved background noise prediction model is to allow better estimates of signal-to-noise levels in environments where noise recordings are unavailable. An ideal model would allow scientists to perform needed signal-to-noise calculations in any terrain by only needing to collect data on a dozen or so geospatial variables that can be found, perhaps through open source databases from ArcGIS, and be able to accurately predict the ambient background noise across all frequencies.

The objective of this study is to investigate whether data mining an existing acoustic dataset of National Parks can help identify a best subset of geospatial variables to predict ambient background sounds at L_{90}–the decibel which 90% of recordings are equal to or louder– across the first 10 one-third octave band frequencies–16, 20, 25, 31.5, 40, 50, 63, 80, 100 and 125 Hertz (Hz).

Research Goals

- Data-mine available National Park Service (NPS) acoustic dataset to visualize data and make inferences on the dataset, to include possible distributions, outliers, and explanatory values. (Chapter 4 - Exploratory Analysis).

- Build predictive models using different statistical techniques and compare results (Chapter 5 - Model Building & Evaluation).

- Review if/how the models compare to previous NPS analyses [1, 2, 3, 4] and Benson [5] (Chapter 6 - Discussion).

- Characterize how well each model derived from the NPS, predicted eight hold out points from the sponsor (in the Philippines). (Chapter 6 - Discussion)

- Discuss general results of how well models performed, cautions on applying the models, and further research. (Chapter 7 - Conclusion)

Research Focus

This project provides a comprehensive exploratory analysis of hundreds of different geographical features aligned with hundreds of recorded noise levels to examine if there is a certain grouping of variables that are more aligned with certain noise levels. Multiple models are built and tested to predict ambient noise levels. One of the sponsor's potential challenges is knowing what features to ask an ArcGIS or geographical information systems specialist given interest in predicting an ambient noise level at that location. Reviewing Benson's [5] database will help narrow down the numerous variables to a select number of useful variables, also known as feature

selection.

Approach

Linear Regression is a well known and widely used procedure for analyzing relationships between variables of interest to a set of related predictor variables. Linear regression is straightforward when there are not as many predictor variables, say $p = 2$. The number of possible main effects models one can build that include or do not include the predictor variables p is 2^p. For example, in a simple project with two variables of *temperature* and *time* there are 2^2, or four, possible linear models that can be built and tested, five when adding an interaction term ($temperature \times time$), and up to seven when including quadratic polynomials $temperature^2$ and $time^2$. One can see how increasing p would increase the complexity of 1) collecting the required data, and 2) evaluating all possible models. Following that logic, modeling 20 variables would require 2^{20} or $1,048,576$ linear models to assess just main effects. If each model took one second to assess, it would take 12.13 days to evaluate all permutations. However this dataset has beyond 20 variables, it has 148 predictor variables. It also has more than one dependent variable, it has thirty three. It becomes computationally infeasible to evaluate each of the 2^{148} models for 33 frequencies. This is where approaches like step-wise regression can help determine the number of sufficient variables to be used per frequency in a more computational feasible time-frame. Forward and backward step-wise regression work through various levels of possible variable combinations at a computationally feasible rate. An R package called 'leaps' allows step-wise regression methods like forward selection and backward elimination [6]. Step-wise linear regression explores linear regression by adding variables sequentially based on the greatest effect on reducing the mean squared error (forward) or

taking out the variables with the least reduction on mean squared error with all the variables already forced in the model (backward) [7]. The disadvantage to the step-wise approach is you can obtain nested results, since "once a regressor has been added, it cannot be removed at a later step" [7] which will most likely lead to a less explanatory model than exhaustive search. Another criticism of stepwise procedures is "... inexperienced analysts may conclude that they have found a model that is in some sense optimal. Part of the problem is that it is likely, not there is one best subset model, but that there are several equally good ones" [7]. This means future teamwork with an acoustic modeling subject matter expert, and geospatial analyst whom knows which geospatial variables are easier to collect for the operational environment of interest, may lend itself to a better model choice for the sponsors' objective.

Subset Regression Research Questions

Using forward, backward, and exhaustive (with a maximum number of variables attempted) regression, this research aims to find the best subset regression model to help answer the following questions:

- Are any of the geospatial features useful in predicting the ambient noise spectral data?
- How well can the model predict data?
- How well does the model extrapolate to locations outside of US data?

Methodology

This study considered the following statistical techniques:

- Correlation and Feature Selection, to reduce the number of variables.
- Multiple Linear Regression using Ordinary Least Squares: stepwise regression to select a subset regression model.
- Single Decision Tree using Conditional Inference Trees.
- Random Forests.

Assumption/Limitations

The primary limitation of the study was the provided data were from the National Park Service. Thus a predictive model will be limited to use for sites that are found in the National Parks – assumed to be quiet and remote – not, for example, city data with heavy traffic. The data were also restricted to the contiguous United States. More NPS data were found from Alaska and Hawaii but did not data per frequency, and thus wasn't applicable. This dataset was further limited by not having access to question the original source for the dataset. When necessary, best reasonable approximations of missing or erroneous data were made and explained. For example, multiple distance values were negative, which seemingly can not be possible. Omitting all observations would have reduced the number of observations considerably. Therefore, in the data-exploration stage of this research, each variable that was observed to have some disparity was noted and an explanation on how it was corrected is given.

Preview

Ambient noise is an important but normally not known *a priori* measure for predicting the detectability of acoustic signatures of military vehicles in different environments. In the absence of measurements, scientists seem to use their best guess based on knowledge of rural, suburban, or urban environments [8]. The effect of not knowing this factor is currently unknown. The acoustic background noise may be insignificant to other variables of interest like the visibility or the radar-signature of an object of interest. However, the potential benefits of a generalizable model would be useful for acoustic detection models. Chapter II presents a review of the applicable literature that focuses on previous ambient noise research.

II. Literature Review

Overview

Literature was reviewed to explore the extent of work already done to characterize a relationship between the sound levels and geographic variables of an environment. A goal of the literature search is to learn and apply the best practices and key-terms of the applicable fields and frame results appropriately. Much research was available to leverage on characterizing the noise and/or sound of an area to its physical characteristics using multiple statistical techniques.

Scope

Publications describing the background noise of an area's environment–ambient noise–seemed to begin in the 1960s, and has become increasingly common. This review found measurement and characterization of ambient noise most prevalent in U.S. Navy research and ocean engineering studies, urban traffic-noise studies, and a new niche in wildlife studies called *soundscape ecology* [9]. The U.S. Army and U.S. Navy seem to use ambient noise metrics most often for understanding battle-space situational awareness [10, 11, 12]. The majority of Air Force literature studied ambient noise in context of protecting hearing of personnel working around jets [13]. The Air Force Institute of Technology (AFIT) had three theses that characterized aspects of ambient noise based on wind [9], landscape and geospatial variables [8, 5]. Recent studies in traffic-noise models [14, 15, 16, 17, 18] and land-based wildlife acoustic models [19, 4, 20, 21, 1, 22] were also a good foundation of relevant ambient noise literature for review. Together these sources formed insights for possible methodologies, described

in the next chapter.

Ambient Noise

In this study, ambient noise, is defined as the decibel of noise that was just exceeded at least 90% of the time or higher, written as L_{90}. Most research on ambient noise fell into one of two categories: (1) signal-to-noise studies, which detect signals in the presence of noise, and/or decrease unwanted noise; and (2) as an acoustic metric to measure the effect of different noises on organisms–especially humans, but also birds, fish, and mammals. The first category is similar to noise being something that is obstructing or masking a more important signal. The second category is similar to noise being the metric of interest, like an effect. The first category, signal-to-noise ratio studies, characterize a system or component's signal against a background noise–such as the speech intelligibility in a loud environment of a cockpit [13], or the detectability of aircraft in differing environments [23, 8, 24, 13]. Research in this category is interested in critical ratios for recognition and interpretation, and audio masking. Examples include studying whether the signal of a bird can be communicated to it's intended target with varying types of anthropogenic sources of noise masking the signal [21, 22]. Other examples include hearing-aid research and speech interpretation in the presence of differing types of background noise [13, 25]. Hearing aid research is interested in the ability to detect speech, music, noise, and speech-in-noise; recent techniques use machine learning to identify the typical patterns of noise versus the other two signals [26]. Another use of background noise is the use of differing materials and designs to absorb noise, such as using green barriers and facade designs in cities to reduce the sound of traffic [27]. The second category was studying the physical effects of noise on the health of a system of interest–such as the short- and long-term effects of

loud noise in a neo-natal intensive care unit on growing premature babies [28], or the effect of noisy environments on overall number of official noise complaints written to authorities [29, 13, 12].

Review Method

Most relevant leveraged sources were found in previous AFIT master's thesis students, and their literature reviews. Then, peer-reviewed publications since 2015 were reviewed. Keywords for the research included but were not limited to 'ambient noise', 'sound pressure level', 'white noise', 'acoustic modeling', 'noise mapping', 'acoustic habitat', 'acoustic signature', 'geospatial', 'sound classification', 'feature selection', 'environmental noise', 'spatial', 'temporal', 'spectrum', 'frequency', 'military', and 'defense.' The most comprehensive and relevant studies were from National Park Service's Natural Sounds and Night Skies division [4, 3, 2, 1] and their partnerships [30, 31, 32, 22]. The majority of results returned, but not as relevant to this study, were Navy underwater ambient noise studies. Other common search results, that were outside the scope of this research, were extensive results on machine learning algorithm effectiveness on classifying *environmental sounds*–like the sounds of kids playing in a park, a coffee pot dripping, lawn-mowers–shorter-durations of sound found in a person's environment.

The remainder of this chapter reviews the previous AFIT students' theses results, the areas of foundational research, the latest relevant peer-reviewed articles, and applications and trends in other disciplines.

Previous Thesis Work

This work builds on the work of three previous AFIT theses: Benson [5], Gaski [8], Popovich [33]. The database of study is from Benson, who paired 513 National Park Service audio data measurements with approximately 150 geographical information systems variables (GIS) data. Gaski [8] characterized nine different ambient profiles of human-performance hearing data and the effect of ambient noise on audibility of aircraft. Popovich [33] looked at characterizing ambient noise due to head-angle orientation to wind and found a best-fit polynomial linear regression.

Predicting Audibility with Logistic Regression Models

Gaski's [8] work largely focused on two things: 1) examining nominal logistic modeling of human performance data and audibility of aircraft, and 2) developing a best-match algorithm to help compare a sample of ambient noise against one of nine other recorded ambient noise profiles from rural, suburban and urban areas [8].

This work expands on the nine standard ambient noise profiles by understanding what factors seem to be most correlated with different frequencies and ambient noise levels. It will also research the feasibility of matching a future point to the database Benson [5] developed, with the potential to represent roughly 500 more ambient noise profiles.

The Effect of Wind and Head Angle with Polynomial Linear Regression Models

Popovich [33] examined human-performance data of the effect of wind direction and head-angle on an observers ambient noise environment. Sound level measurements were obtained from extensive wind tunnel tests simulating what a human would experience of different wind speeds and different directions. He produced a polynomial-fit model for predicting wind noise levels at different frequencies using wind speed and head-angle [33]. Furthermore, he used directional and omni-directional graphics to visualize the results of prediction versus performance. Popovich stated the military application of wind models:

> "With an increasing amount of surveillance provided by remotely piloted aircrafts and other technologies, there is hope that the United States military will eventually integrate the visual feeds produced by this data with other forms of data. For instance, wind speed and direction, coupled with aircraft sound profiles in the operational environment could be used to adapt vehicle flight paths to decrease audible detection" [33].

Popovich developed 31 frequency-based models, and a combination of forward step and backward stepwise regression, using up to the fourth-order polynomial terms. Developing frequency-based models is also an important step for model-building in this research.

Database and Initial Random Forests

Benson [5] added a variable called Land Cover Land Use (LCLU), to a National Park Service (NPS) dataset of 513 audio observations to see if a machine learning technique

called Random Forest could predict the acoustic metrics given certain landscape data not necessarily in the database. He also created a matching schema in R– an open-access statistical programming language–to find the five 'best matching' observations to five of the 513 semi-randomly selected observations. The random observations were semi-random because he ensured he sampled five different landscape types. The results showed fair results for four of the five landscape types but poor results to the 'Shrubland' landscape [5].

The poor predictive performance of Benson's model for the Shrubland landscape may be because 38%–a majority of the 513 observations–were identified as 'Shrubland'. The other 62% of observations were categorized as 12 different landscape types, thus several landscapes had 5% or less representation of the total observations from which the model was created. Therefore, it is possible the abundance of data on the Shrubland landscape led to a more varied dataset for shrubland. In other words, the Shrubland's poor performance in Benson's model was probably a more realistic measure for real-world random observations outside of the existing dataset, not a sign that the Shrubland prediction method was not accurate enough.

A proposed follow-on to Benson's work is to recreate the analysis done by NPS's Natural Sounds and Night Skies (NSNS) division using Random Forests–and see whether classical statistical learning methods like linear regression can make better predictive models for observations not existing in the NPS dataset. This is further explained in chapter III, Methodology.

Disciplines that use Ambient Noise

Disciplines that traditionally study the effects and characteristics of ambient noise include, but are not limited to the following: human-health studies, naval/marine/underwater studies, seismic monitoring of earthquakes, volcanoes and nuclear test compliance, aeroacoustic studies interested in absorptive properties to reduce engine noise inside the aircraft and outside the aircraft, urban planning and mitigation of loud noise in residential zones.

Human Health studies

In human-health studies, researchers are interested in how noise can be interpreted as annoyance [34] or pain, or how noise can physically damage one's hearing, as well as the holistic life-time effects on behavior. Extensive health studies have led to conclusions that long-term exposure to unwanted noise abundant in city environments is correlated with increased risk in heart-problems, tinnitus, hypertension, decreased hearing, sleep problems and cognitive impairments of children [35, 36, 37, 38, 39, 17, 15].

Traffic-Noise studies

Traffic-noise prediction studies focus on using existing land-use regression (LUR) methods for air-based pollution models for noise prediction [38, 15, 40, 35], and create traffic-noise predictions. Most of the LUR publications stated favorable results, for example, Torija et al. [39] reported a nonlinear model for LA_{eq} with an $R^2 = 0.94$ and MAPE=1.15 dB using feature-selection technique *wrapper for feature-subset se-*

lection (WFS) and a machine-learning regression method called *sequential minimal optimization (SMO)*. However, one study reported low-accuracy findings for acoustic measure L_{DEN}–an A-weighted day-evening-night equivalent sound level– from geographic information system (GIS) variables, with an $R^2 = 0.130$ [38].

There an increasing trend to improve models using machine-learning algorithms like Artificial Neural Nets (ANN) [18, 39] to get more precise noise level predictions. There is also an increasing trend to incorporate frequencies into traffic-noise models, instead of just overall noise levels. Some studies found using $\frac{1}{3}$ octave band frequencies as dependent variables in their models caused other independent variables, like vegetation that previously were not important to the overall loudness of the area, to become important to select frequencies [17]. Adding frequencies to the model can help for understanding audibility over large areas [3]. They can also help influence policies to help monitor and mitigate the effects of harmful levels of noise.

Navy Research

Ambient noise research in a completely different medium–water–is well-established and published metric since at least the 1950s. Navy military research and marine mammal research studies almost always characterize the underwater ambient noise as a recognized 'acoustic signature' to characterize the water environment. A study of battlespace awareness from 2004 describes the importances of Navy acoustics: "Acoustic sensing technologies are used to detect, identify, and locate sound wave and seismic activity to characterize underground or underwater activities and facilities. These measurements allow characterizations for targeting and battle damage assessment" [11].

The importance of the U.S. Navy studying the ocean's ambient sound is to increase their understanding on detecting foreign submarines and mine signatures, while keeping their own U.S. assets' location undetected [41]. Prior to 1970s, it was relatively easy to detect a submarine due to how loud they were, but since then, submarines have better technology to be essentially silent objects traveling in the ocean. The ability for a submarine to travel close to our country's border, undetected, and deliver a missile in a short time span is a possible driver for the extensive research available from the Navy on this metric. Most of the tactics developed to help understand these behaviors, like SONAR, were from studying how animals communicate underwater [41].

Marine Animal studies

Fish and marine animals have evolved over millions of years and have refined senses to help navigate, communicate, and sense threats in an underwater vision-limited environment [42]. Whales, for example, communicate in a frequency range that allows them to communicate up to 100 kilometers away underwater [43]. Some fish rely on their sense of hearing to find historical breeding grounds, or evade the sounds of predators like snapping/clicking shrimp [19]. Similar to fish and marine animals, naval ships and submarines rely on technological sensing devices cued in to ocean ambient noise to better inform underwater battle-space situational awareness, to navigate, communicate and sense threats as well.

Ambient noise in water helps identify changes in anthropogenic noise and understand the impact on animals. Navy researchers, energy/oil companies, fishing industries, recreational boaters, and marine biologists are concerned about how increasing human-caused noise activities can mask signals that are important to fish and mam-

mals' ability to adapt and survive in their changing environment. Ambient noise is not only used to characterize their 'normal' environment, it is also used to track, measure speeds [44], and even visualize objects in water—see 'Acoustic Daylight' [45]. A review of the marine biology studies is not covered here, but a thorough literature review from over 100 marine noise studies is found in Erbe et al.[46].

In general, the climate variables, topography variables and compactness of terrain seems relatively well understood in ocean science to shaping ambient noise. Unfortunately the extensive research on underwater models does not easily help inform Air Force models–one key difference is sound travels faster underwater, because liquid is a denser material than gas. However, much like the Navy can learn from marine-life acoustic studies, there is potential to learn from land-based acoustic studies. An increasing number of land-based animal studies are attempting to characterize ambient noise as a product of climate, topography and/or terrain.

Increasing Literature Trends in Ambient Noise: Acoustic Habitats and Strategic Noise Mapping

Ambient noise is an increasingly important metric in acoustic habitat studies like landscape ecology, and strategic noise studies. Landscape ecology studies focus on the natural background noise of an environment, which traditionally is defined as the background noise without human-sources of noise, although some choose to keep human-sources of noise. These studies refer to natural background noise as the 'ambient sound environment' [47], 'acoustic habitat' [19], 'acoustic environment' [30] and 'soundscape ecology' [9]. In these studies ambient noise is used as one of several measure to help characterize the health and diversity of the natural environment, and to help identify changes over long-term studies, usually caused by anthropogenic sources

of noise.

Another source for increasing number of publications with interest in measuring ambient background noise is policy-based. In 2002, the European Union (EU) mandated all member countries implement strategic noise maps–prediction models of the sound propagation in an area–to monitor noise levels of their countries and help inform policies to mitigate rising noise levels due to their known effect on human health [14, 35]. Of further importance was the connection that the well-known environmental land-use regression–which is just multiple linear regression using geospatial variables–is a helpful technique not only for modelling traffic air pollutants, but also noise. Strategic noise maps could also help protect existing quiet areas [14]. These strategic noise mapping studies center in Europe [14]. Within the last two years many other countries have directed studies based on their own interest to study noise: Canada [40], India [48], South Korea [49].

Most published land-use-regression (LUR) studies seem to report an R^2 value of at least 0.5 when reporting model accuracy of the location studied, and using expertly-advised geospatial features, but one study stood out as a potential warning of the misapplication of using tailored models to a location that was not in the study. In a LUR study of South Africa, the authors reported an R^2 value of 0.1 despite using many of the features identified as important from previous traffic-noise studies [38]. The authors speculate that because the location was in a developing country the GIS data may not have been accurate, and other variables were probably more important that were not easily derived from GIS, such as neighborhood noise. "Beside traffic, the household density was also a significant noise predictor variable. This result was expected because these areas are crowded, and thus the noise coming from the neighborhood is expected to be substantial. However, derivation of GIS predictors

as a surrogate for neighborhood noise is tricky, and thus another reason for the low noise variability explained by the LUR model may be the underestimation of the neighborhood noise by the GIS variables available." [38] This is further motivation to help understand that data mining the National Park Service dataset will most likely carry the same limitations as those seen when applying the insights gained from European and American traffic-noise studies to locations that are vastly different than those sampled.

In addition to traffic-noise studies, there is an increasing trend of literature on the effects of anthropogenic noise on non-humans. For an extensive literature review which consolidates 242 peer-reviewed articles on the effects of anthropogenic noise on wildlife and gives recommendations for future acoustic measurements of importance, see the Shannon et al. [22] article. In November 2016, an extensive literature review was published on the research from 1990 to 2013 on the effect of noise on wildlife [22]. The greatest insights offered by both the acoustic habitat and strategic noise mapping studies is the methodologies they used to pick geospatial variables of importance and their findings with what natural and anthropogenic noises were associated with what frequencies and the methodologies they use to predict or model ambient noise. Some of the methodologies used employ decision-tree random forests (RF), and self-organizing maps (SOM) [1, 2, 3, 4, 30, 5, 50].

Methodologies of Interest

The methodologies of interest for this research are linear regression method used in the traffic studies, the decision-tree random forests (RF) used in mostly the National Park Service studies, [1, 2, 3, 4, 30], and the elastic-net penalized linear regression model used in the Boston-traffic noise study [16] which are explained below.

Methodology 1: Random Forests

The US NPS NSNSD is one of the more prolific sources of research on the influence of geography and environmental factors on ambient noise. The NPS NSNSD has extensive research on modeling the influence of geospatial, temporal, and terrain features on noise. Across 10 years of acoustic monitoring over 1 million hours of audio from over 400 unique locations in the contiguous United States, Hawaii, and Alaska has been recorded. A US Public Law [51] requires the NPS to monitor the noise of its natural parks and report each year on aircraft noise and whether it is negatively impacting the natural quiet enjoyed by visitors. The reports are also supposed to better inform aircraft on the elevations to fly to lessen noise impact in the parks. The 10 years of acoustic metric summaries from NPS NSNSD were evaluated using random forests.

The key research from NPS that serve as the background information for most of this thesis project are the following: a geospatial sound model to map sound pressure levels on a continental scale [1], the influencing factors and spatiotemporal patterns of environmental sound levels [3], the explanatory variables generation for geospatial sound modeling standard operating procedure [30], how to measure acoustic habitats [19], and GIS Metrics for Soundscape Modeling standard operating procedure [31]. The article about influencing factors [3] is most relevant since its employs the use of random forests for describing the relation of ambient noise to landscape values. It is also our speculation that in the future the NPS, together with its partnerships, may offer an approximation of ambient noise in the absence of real-data as part of another open-source software called Sound Mapping Tools (SMT) [52], since it mentions ambient sound as an optional input into its model, and references Mennitt 2014 et al [2] as an example of a data source for ambient noise although not used.

Methodology 2: Elastic Nets Regression

Harvard Researchers looked at 400 sites in Boston, Massachusetts, from February 2015 to February 2016, to develop a model for predicting A-weighted sound pressure levels of low, medium and high frequency sounds. The authors used an elastic net variable selection technique and the final model explained approximately 60% of the variability in each measure.

An elastic net variable selection technique is a linear regression technique that is mathematically between a lasso regression technique and ridge regression technique [16]. It is a technique more known in genome research than noise studies [16]. The authors state the results were similar to other A-weighted models of urban environments, because they included "transportation related variables such as length of roads and bus lines in the surrounding area; distance to road and rail lines; traffic volume, vehicle mix, residential and commercial land use." However, making frequency specific models allowed other variables to appear, such as "temperature, vegetation, impervious surfaces, vehicle mix, and density of entertainment establishments and restaurants." As a result of using the elastic net technique 239 potential predictors were considered, and a total of 58 were included in at least one of the final prediction models. The use of elastic net supposedly allows 'better grouping of variables with correlations' and was attempted on the NPS NSNS data but analyzed results were not available in time to be in this report.

General Military Applications of Acoustics

Sound is most useful in the military for beyond line of sight measurements. Beyond line of sight is where one cannot see an approaching vehicle but sensors be able to

hear/feel and locate its approach, according to Becker [53]. For applications where vision will not suffice, better hearing can help get the tactical edge [53]. Becker predicted rising importance of acoustics in the following military applications due to its relatively low cost: low-power wakeup, networked unattended ground sensors, stand alone services to help detect gun locations, sniper locations, or 'wide area mine fuzing, ' non-lethal weapons, IR and acoustics combined solutions for vehicle based helicopter detection systems, and vehicle based self protection systems [53]. Similarly to detecting ground movement of Army or other aircraft, seismic and acoustic mine devices can be easily camouflaged and networked to allow better battlefield awareness and sensing [53]. Becker [53] says vehicle interiors are hard for soldiers to understand what is going on outside, so they could be equipped with microphones to allow a vehicle to know what sounds are occurring outside of the vehicle. For example if they can hear a sniper shooting or a missile is inbound, they can react faster. Similarly a network of acoustic sensors may be able to help pinpoint acoustic events like a sniper shot. Becker says most artillery situations comprise of at least three intense acoustic events, which can make tracking complex. In general, surveys of publicly available Army budgets show continuing interest in military applications of acoustics.

Air Force

In the Air Force, traditional noise studies seem to center on noise-abatement and noise-annoyance studies for hearing safety of jet engine maintainers and sleep-interruption of the general public surrounding military bases with active runways. Noise studies are important for determining the type of zoning around military bases. Most Air Force projects also have an environmental impact statement and safety procedures in place to prove that no wildlife or people will be harmed with

the impact of their testing or construction. This includes monitoring the effect of sound on wildlife, especially endangered wildlife. However no public records were found that included the ambient level sounds.

Noise studies are also as an emerging field of interest to UAV survivability [24]. Unmanned aircraft were designed without survivability as a critical system characteristic "due to their lower cost and the obvious reason that a human was not on board" [24]. Therefore since survivability was not designed-in from the beginning, McDaniel argues "situational awareness and route planning provides the best means of aircraft survivability." McDaniel states the acoustic profile may be the most important part of a vehicle's signature, with which the "proper knowledge of the UAS's acoustic profile, mission planning can be used to increase aircraft survivability and enhance mission effectiveness" [24]. Finally, McDaniel states one of the more useful techniques for UAS survivability is route planning: "Using knowledge of the vehicle's acoustic profile, the terrain, and the atmospheric conditions, a specific mission can be tailored to avoid or minimize audible detection by a listener on the ground." McDaniel explains how mission planning can be conducted in real time as "updated threat information becomes available" or as a pre-planning exercise. McDaniel also states the biggest drawback to real-time mission-planning is the computational requirements for ground control station computers [24].

Army

One Army experiment studied the acoustic detection of different acoustic signals in the presence of ambient noise. In addition to being an interesting and valuable study, it articulated the importance of understanding noise accurately in combat:

> "Accurate sound perception is directly related to both Soldier mission and safety (Katzell et al. 1952; Abouchacra et al. 2007). Infantry Soldiers generally agree that in limited-visibility environments, the sense of hearing is their main survival resource. In combat, detected sounds alert the Soldier that something is there, and early recognition of the sound source allows the Soldier to take swift action (Price and Hodge 1976a, 1976b). The farther away the sound source is detected and recognized, the longer time the Soldier has to respond to the threat (Abouchacra et al. 2007). Therefore, early detection, localization, and recognition of surrounding sound sources are critical in any military operation," [10].

Studying sound to help assess the location of a enemy weapon firing as indicated earlier general military applications [53] is important and just one facet of the use of studying sound in the Army. A recent U.S. Army news article states the U.S. Army Corps of Engineers' Engineer Research and Development Center (ERDC) developed the ability to pair acoustic sensors with light detection and ranging (LIDAR) to help soldiers and marines anticipate enemy positions and plan terrain movement while moving from ship to seashore in possible Anti-Access Area-Denial environments. By deploying acoustic sensors into the field with reconnaissance aircraft or UAV, a 3-D map can be created with all the sensors' data. The article stated one could "track the movement of Soldiers, who unknowingly set off the acoustic and seismic sensors on the desert floor," [54].

The Army Test and Evaluation Command (ATEC) has an Acoustic Research Complex (ARC) which measures the acoustic signature of aerial targets. ARC is described as following:

> "The ARC facility is the first of its kind within DoD and the research community as a whole. It is used to help with the design, modification, and increasing combat survivability of current and future aircraft. The ARC provides collection capabilities of three-dimensional (3D) acoustic data from operating aircraft that are not available anywhere else within DoD or in the private sector. This capability responds to a critical need

> for validation of existing predictive acoustic models. Such models are used for aircraft design, survivability, non-linear acoustic propagation research and assessing noise exposure to residents living adjacent to airfields. A large area microphone array, with sensors along the flight path as well as in the vertical, enable 3-D capture of the radiated acoustics for any air vehicle. The ARC measures the noise radiated in all three dimensions simultaneously under various operational dynamic flight conditions. The measurement station consists of microphone locations near the ground and acoustically instrumented tall towers for rotary wing, UAV, fixed wing heavy, and high performance aircraft." [55]

Current Machine Learning Uses in Acoustics

There is increasing research available on classifying acoustic data with machine learning methods. The acoustic data is usually for characterizing natural speech to text and natural language context, as well as automatic recognition of music songs, like those algorithms available in popular mobile-phone applications such as 'Shazam' or ''SoundHound'. There are also increasing trends to look at non-speech acoustic data, sometimes called 'acoustic events'.

In robotics, automatic speech detection and recognition are important and there is growing interest in audio scene analysis. Many articles propose better algorithms or methods to detect sound events, [56]. One of the reasons it will be important is for automated vehicles to detect when they are moving between indoors to outdoors. It is also important in forensics to be able to detect the authenticity of a received video, and identifying inserted splices of forged video by their audio and image characteristics [57].

A growing number of online data-science communities, like Kaggle.com, reward people for best classification algorithms on datasets, to include a growing number of audio datasets. This has led to a drastic increase in publishing and sharing acous-

tic machine-learning driven scholarly articles. Machine Learning and databases are largely growing to support the research demands for better acoustic detection [58].

Methods employed already include: unsupervised machine learning, [59], semi-supervised machine learning [60], deep neural nets [61], feature extraction [62]. Other research proposes classification benchmarks for robots, [63], measures and methods for passive audio surveillance [64], vehicle detection using neural nets [65]. A thorough survey of the machine learning techniques employed can be found in [66] and [67].

The most recent studies pertaining to audio events and audio scene recognition use other machine learning techniques beyond the scope of this research but may be pertinent to future studies. Those include: "document-event co-occurence matrix for topic analysis," [34], deep neural nets, [68], gabor-matching pursuit [69], and information bottleneck principle [70].

Machine-learning classification of audio events is outside of the scope of this thesis since the audio data available is mostly numeric and not categorical. With the original raw audio files (millions of hours of recordings) some of these machine learning techniques could apply for more limited purposes, such as identifying aircraft flying in the audio recordings to be deleted, or measuring the health of an eco-habitat by counting the number of indigenous bird-calls. Extending the applications of machine-learning classification algorithms outside of the NPS dataset are also beyond the scope of this thesis.

Summary

The literature reviewed provided background on three focus areas: studies that leveraged a large dataset of acoustics and different GIS-paired geospatial variables, statistical techniques to gain insights on which geospatial variables were influential, and potential military applications of ambient noise. It also showcased some current uses of machine learning in classifying different sounds. Many traffic noise studies use a technique called 'land use regression' (LUR)—which is really just multiple linear regression using geospatial variables. One study highlighted the weaknesses of applying insights made from a LUR noise model on a location that wasn't used to develop the model [38].

Multiple linear regression models will be employed on the NPS dataset modified by Benson [5], and time durations to develop and test models will be compared to other methods.

III. Methodology

Research Purpose

There are no known methods to predict ambient sound pressure levels in non-US geographical environments based on geospatial variables (such as population density, distance to roads, or amount of trees in the area). The intent of this research is to improve potential operational signal-to-noise acoustic prediction models using terrain-type descriptors as the predictor variables. There are many studies done in various different geographical areas–India [48], Europe , South Africa, Middle East, United States–but most focused on busy-urban traffic noise levels, and a subset of those report the sound based on frequency. There are a growing number of environmental studies, that focus on studying the acoustic habitat of quiet and remote places and used frequency [4, 3, 2, 1]. The purpose of this current study is to explore the utility of various statistical techniques on existing National Park Service acoustic data from contiguous United States locations, and see if that model can predict points outside the modeled set. The techniques investigated are linear regression, penalized regression, and random forests.

Step 1: Data Prep and Exploratory Analysis [Chapter 3]

The first step was to explore the dataset. Understanding data helps inform appropriate statistical techniques. Outliers with leverage were identified and further analyzed, transformations on non-linear variables were tested. Exploratory data analysis led to creative ways to visualize the data. The data were highly-dimensional, so scatter-plots were useful but limited to examining two variables at a time. Since examining $\binom{200}{2}$

pairings was not possible, automated techniques were used to help identify outliers and correlations between the main effects. Examining interactions and polynomial effects occurred after reducing the main x-variables to the smallest subset possible. Correlation-based feature extraction and elimination of variables with near-zero variance resulted in a dataset with a smaller subset of columns/variables which would help expedite model-building and evaluation. Chapter three details these efforts.

Step 2: Model Building [Chapter 4]

The second step uses the reduced data set to perform model-building and model-evaluation. Modeling techniques include:

- **Multiple Linear Regression** using forward, backwards, and exhaustive search.
- **Penalized regression** using lasso, ridge regression, and elastic net methods.
- **Random Forests** using different number of trees, nodes, and splitting techniques.

Chapter four details these efforts.

Step 3: Model Evaluation and Prediction [Chapter 5]

Common model evaluation measures for prediction and estimation models are as follows :

- R^2 is the proportion of variability in the response due to model fit, in general,

the models with greater values fit the model better.

- **Adjusted** R^2_{Adj}, is like R^2 but with a penalty on the number of variables to help more simple models score higher over overly complex models–the closer to 1 the better.

- **Mean Square Error** (MSE) is the square-distance of residuals divided by $n - p - 1$ where n is total observations used, and p is variables in model–the smaller the better. MSE $= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

- **Variance Inflation Factor** (VIF) is to detect multicollinearity; if this is high, the model is not easily interpretable, as some of the variables are correlated with each other; meaning the model suffers from multi-collinearity. Generally VIF values under 10 are fine. $\text{VIF}_\text{i} = \frac{1}{1-R_i^2}$

- **Mean Absolute Error** is similar to MSE but uses the absolute difference versus squared difference to help minimize effects of outliers– the smaller the better.

- **PRESS** predicted residual error sum of squares statistic is a form of cross-validation where each point is left out of the model, the model is fit, the point is estimated, and the holdout residual is calculated. The smaller the better the model is at predicting.

- **Mallows** C_p, when a model fits well, the expected Mallows Cp is equal to $p+1$. A plot of Mallows C_p against the number of predictors can help pick the best model size. In general the smaller the better.

- **Bayesian Information Criterion** a way of penalizing complex models. The BIC helps generalize when the model is becoming over fit.

$BIC = n \cdot \ln(RSS/n) + k \cdot \ln(n)$ where n is the number of observations in the model, RSS is the Residual Sum of Squares, and k is the number of parameters in the model plus two.

- **AIC** the Akaike Information Criterion is used to rank the quality of the models relative to each other–the smaller the value the better.

Chapter five details these efforts.

Multiple Linear Regression Model Building

The data was further reduced in this stage to eliminate approximately thirty nearly-zero variance columns, and also split the data into training and testing/validation sets. Backwards and forwards regression were employed. Backwards and forwards regression helped estimate the number of parameters that were needed in the training set before over-fitting the model. It looked like no more than 10 variables were needed before the Adjusted R^2 value plateaued. Having an idea of the number of variables in the model, and taking out near-zero variance columns, and the columns with the least amount of unique variables, helped scope down efforts for exhaustive linear regression.

The results of the linear regression appear unique, as no known publicized efforts exist examining linear regression on the NPS data.

Penalized Linear Regression Model Building

Penalized regression was also attempted, motivated by the Boston traffic-study on Elastic Net feature selection [69]. Lasso, Ridge and Elastic Nets are similar to linear regression but with added penalties on the number of variables and the size of

coefficients. The results of the elastic net procedure were not available in time for this report, but are implemented in the deliverable R-Markdown report for future researchers or sponsor if time and interest allow.

Random Forests Model Building

Random Forests is a classification method that improves on decision trees. Decision trees are characteristically simple models, but to get great accuracy usually requires over-fitting the model. Ho [71] increased the predictive accuracy of generalized decision trees by using thousands of decision trees generated with a subset of random variables at each node, and using the average predicted value given across some number of decision trees (hundreds or thousands) in regression. Multiple decision trees using random variables are called "Random Forests." On the NPS dataset, one decision tree may start with one node that defines the distance to nearest airport as above or below 5000 feet, and then those data points that are between zero and 5000 feet go on to be split to amount of water in a 5km radius, and then are split by more nodes like the above process. Continuing this example, another decision tree might use the same data but start with the historical wind levels in the area, and then split the data among the amount of shrubland in a 5km area. The same data would go through both of these trees—and thousands of other decision trees which were randomly selecting variables for each node to minimize some object function—and the result of the predicted dependent variable would be the average value over all of the decision trees. Another motivation for using random forests was to compare results from this study to those already published by the National Park Service.

Step 4: Comparative Model Analysis and Hold-Out Data Testing

The final step was a requested proof-of-concept on the model from an out-of-data set point where sound information was gathered by the sponsor. Originally it was assumed if the sponsor was interested in using the National Park Service dataset for inferences on how to quantify and predict sound in remote quiet places, it was hypothesized the purposes would be used to predict the sound of areas like the mountains of Afghanistan or a cold desolate hazardous place like Chernobyl. However it turned out the proof-of-concept data was from a heavily populated metropolis in the Philippines, Cebu City, and no geospatial variables were provided. Those that were collected would most likely be far outside of the parameters used in the NPS model. The resulting regression models were applied to the Philippines dataset to complete the task, and serve as a warning to why one should not extrapolate a model outside of the parameters it was built from.

The out-of-data set had known acoustic metrics given by frequency and latitude/longitude but no further information was provided. Since no geospatials were provided some information was obtained from ArcGIS. The Philippine locations were not identical to any of the thirteen LCLUI types from the National Park Service–it was a land-use of cropland/coconut-plantation mix. Cropland was an underrepresented land-use type in the NPS data. Coconut-plantation was also not in the NPS data.

The Data

National Park's Natural Sounds and Night Skies Division (NPS NSNSD) merged historical audio recordings in multiple national park sites with ArcGIS data to provide

hundreds of different measures of the land around the given latitude and longitude of each audio recordings precise GPS location. The data source received for this research was an excel file from Benson [5]. The primary objective for NPS NSNSD research was to understand what features were responsible for rising noise levels heard in all of the parks, and they determined it was anthropogenic. Benson added a variable called Land Cover Land Use Color Index to investigate whether certain land-types could be used to predict and characterize ambient noise.

Code

All code to do the methods described in this chapter, are written in R [72], and delivered to sponsor. All others may need to request. Packages used in R are noted under 'libraries' and include but are not limited to 'stargazer', 'MASS', 'leaps', 'caret', 'ranger', 'moments', 'MPV', for analysis, and 'ggplot2', 'HMisc', 'corrplot' for data visualization [73, 6, 74, 75, 76, 77, 78, 79]. The files included are listed below. Before running these reports in R one must ensure the active working directory is set to be in the folder where these files are located.

- **dataSource_r2.xlsx** This is the original file created and delivered by Benson.
- **Start.R** The start.R file will load the source data as-is and performs a number of steps to 'clean' the data—like imputing values for zero-wind values, zero-sound values in the first and second octave frequency, correcting negative distances, and correcting proportion values that are less than zero or exceed one. Future researchers should review this thoroughly documented Rfile for all assumptions and corrections. They can try changing the assumptions and corrections here too. A little bit of tinkering with the 'read_excel' function at the beginning of

the Start.R file will be necessary when the source of the data changes. When all lines of this file are run, the result is an R-object saved to the users working-directory called, 'cleandata'. The 'cleandata' object is used by the next two files. In addition to this file, it also creates files that list the variables that were eliminated due to near-zero variance, values that were over 100% or negative (when they shouldn't be negative), and values that were removed that had 75% or greater correlation. This value can also be changed, to 90% for example, to keep more variables in the running for model-building. However the greater number of variables, the slower exhaustive regression and random-forest methods will take.

- **train.R** the Train.R code file contains information on how the test and training sets were initially split to create an initial forward, backward, and exhaustive regression model. The data was initially split 50% training and 50% testing. It reads in the 'cleandata' created from the 'Start.R' file. This file is not really necessary but is included for reference. The R package 'caret' was used to create 10 semi-random folds for 10-fold cross-validation so the split data is no longer necessary.

- **Analysis_firstpart.rmd** is a reproducible RMarkdown report that does the forward, backward, and exhaustive regression. It can compute a lot of information for each model. Change $run = FALSE$ to $run = TRUE$ before each section to run the model. When set back to $FALSE$ it will import the models instead of creating them. This code creates the multiple linear regression models for forward, backward, and exhaustive regression and saves them as R-objects in the specified working directory. Upon knitting it will produce a PDF/HTML/ or Word report as desired. If you change the 'runfast' to 'FALSE' it will print

out outliers, residuals, and suggested polynomials for each model, but the report will easily be over 200 pages long so care is advised. When printed the code is approximately 40 pages long so it is not included so it is not included in the appendix.

- **Analysis_secondpart.Rmd** is a reproducible RMarkdown report that imports the forward, backward, and exhaustive regression models created in the first part, and creates models using conditional inference trees, and random forests, and stores these objects in the active working directory for easy reference later (without rerunning analysis). It would be straightforward to add another methodology from caret into this report, like 'elasticnet' which was originally planned but analysis were not completed in time for this report. Upon knitting this document in R-Studio, the results of all the models are compared and printed as a PDF/HTML/ or Word report as desired. When printed the code is approximately 20 pages long so it is not included in the appendix.

- **parallel_coordinates.Rmd** is a small amount of code modified from the github account 'timelyportfolio' to enable a visualization of the data as an interactive parallel coordinates plot. One can change the axes to represent different factors of interest. Our code defaults to the factors important to determining L90f1 and the factors found for the philippines. This code is small enough that it is included in the appendix.

- **philippines.xlsx** a small excel file that contains the geospatial data collected for the eight philippine locations. The data was not collected in accordance with previous NPS procedures and is not validated.

IV. Data Exploration

Data Preparation

The initial data set had 513 rows and 248 variables. Appendix A describes the exploratory variables in the NPS data set as described in Benson's appendix [5]. Variables were classified as location-variables, land cover variables, land use variables, and other environmental factors. The location variables were generally site specific details like the lat/long, year or years recordings were taken, season, elevation, slope, etc. Land cover variables were generally what physical type of environment is in the 200 meter or 5 kilometer area: forest, barren, shrubland, wetlands, water, etc. Land use variables were specific to how the land is used: conserved park lands, timber harvesting, livestock grazing, cropland, residential suburban or urban, industrial, etc. The environmental factors category was a collection of many different metrics. Sixteen metrics were distances to the nearest facilities of different types–airports, roads, railroads, streams, coast, etc. Two metrics were aircraft specific to the sum of weekly flight observations within 25 miles, and the sum of military flight paths. Six metrics were descriptive of the amount of roads in an area. An additional six metrics were precipitation and temperature specific. Two measures provided a raw and ordinal value for the topographic positions which described the six different landscape the acoustic recorder was on–ridge, slope, flat, etc. One measure called Wilderness was the sum of designated wilderness in meters squared. Finally, the last measure Wind_CRU was the historical value of wind in that area. Benson's appendix states it is "Wind power class potential density (50m AOA)" [5], and Nelson source states "Wind speed annual mean (1960-1990) in meters/second, 10 meters above ground" [30]. So from these two descriptions we assume that "Wind_CRU" is both a 30-year

mean of wind 10 meters above the ground and within 50 meters of the siteID.

Table 1 shows a partial illustration of the dataset. Only 9 of the 216 variables are shown: siteID, Season, park, Latitude, Latitude, Longitude, Elevation, Slope, Barren200m, and Developed200m. The first three variables are categorical information as they contain text, whereas the next six variables are numeric continuous values. Only five of the 216 variables were text: siteID, season, park, LCLUCI, and TPI.

Table 1. Partial Illustration of Source Data: 5 Rows and 9 Variables

	siteID	Season	park	Latitude	Longitude	Elevation	Slope	Barren200m
1	ACAD001	Summer	ACAD	44.419	-68.320	4	0.700	0
2	ACAD002	Summer	ACAD	44.300	-68.366	89	10.953	0
3	ACAD004	Summer	ACAD	44.362	-68.276	82	2.669	0
4	AGFO001	Summer	AGFO	42.424	-103.732	1,339	1.017	0
5	ARCH001	Summer	ARCH	38.682	-109.543	1,550	5.781	0.167

Issues found in Exploratory Analysis

Shrubland or Evergreen Forest Data

Several issues became evident in the exploratory analysis phase. First, most of the data (65%) was from just two types of landscape, as shown in table 2.

LCLUCI 52 is Shrub/Scrub; "less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in an early succession stage or trees stunted from environmental conditions." [80]

LCLUCI 42 is 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."

[80]

The rest of the LCLUCIs are 5% or lower representation. Table 2 gives the number of observations by LCLUCI. Figure 1 shows a visual of the information in Table 2.

Table 2. Top Recorded LCLUCIs (Descending)

LCLUCI	LCLUCI Labels	Count	Percent
52	Shrub/Scrub	180	38.7
42	Evergreen Forest	137	29.5
41	Deciduous Forest	30	6.5
31	Barren Land (Rock/Sand/Clay)	29	6.2
21	Developed (Low Intensity)	27	5.8
71	Grassland/Herbaceous	24	5.2
95	Emergent Herbaceous Wetland	10	2.2
90	Woody Wetlands	9	1.9
11	Open Water	6	1.3
23	Developed (High Intensity)	4	0.9
81	Pasture/Hay	4	0.9
12	Perennial Ice/Snow	2	0.4
22	Developed (Medium Intensity)	2	0.4
82	Cultivated Crops	1	0.2

Large variations in scale between variables

A second issue is variation in scale. About 65 variables describe a percentage of an area, and are between 0 and 1. The "Distance to ..." measures are several orders of magnitude larger. The 'Wilderness' variable is on average 1,943,512 meters^2, but varies from 0.0 to a maximum of 9,031,550 meters^2. Wide variations in the scale of independent variables can affect model estimates. Data transformation such as re-scaling or normalization are used to accommodate data issues due to scaling. Table 3 shows the contrast between some of the large and small variables discussed. To amend the large difference in scale that could hinder understanding the coefficients in regression, most values that were measures of distance in meters were converted

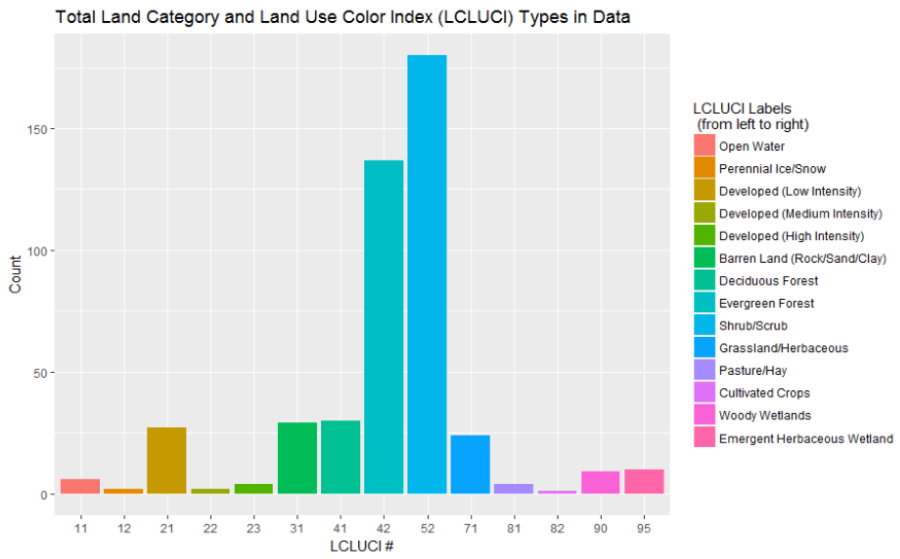

Figure 1. The National Park Service acoustic data mostly contained "Shrub" and "Evergreen" land-cover, and so the resulting model was predicted to be more reliable predicting those LCLUCIs

to kilometers. The Flight-Frequency-variable was also divided by 1000. However the MilitarySum_25Miles variable was not changed, because its values were smaller than expected from the description given–the values only ranged from 0 to 0.5.

Table 3. Partial Illustration of Differences in Scale between Variables

Statistic	Mean	St. Dev.	Min	Max
Wilderness	1,943,512.000	2,774,080.000	0	9,031,550
DistAirportsAllMotorized	18,575.000	12,778.000	201.000	67,810.000
DistWaterbody	2,312.000	2,863.000	0.218	19,061.000
Barren200m	0.059	0.139	0.000	0.833
TPIRaw	0.613	20.000	−107	136

Skewness, Kurtosis, Normality

Most of the geospatial variables have values between 0 and 1 to indicate percentage some type of land-cover in a given area. Therefore if Shrubland200m = 0.5 then the amount of shrubs in the area was 50%. However most of the geospatial percentage

variables had observations closer to zero, and less frequently the data observations fell closer to one, resulting in right-skew. The variable 'RecCon200m' was the one exception, as the majority of observations had a value of Recreational or Conserved lands around 1, which made it left-skewed. Transformations were reviewed but ultimately deemed unnecessary since the residuals of most of the linear regression fits resulted in data that was approximately normally distributed.

Impossible Values: negative distances, values greater than 100%

Many observations for variables that were supposed to be between zero and one were negative or greater than 1, indicating a value of a resource greater than 100% or less than 0% which was physically not possible. For example, the observation of the proportion of natural water in a 200m radius–'WaterNat200m'–was often recorded as less than zero, and no documentation in the NPS Standard Operating Manual on Exploratory Values described any of these observations. The majority of values that had these seeming errors were for the human-use land type characteristics. Since the errors were systemically found for human-use land characteristics it was speculated that some transformation from one measurement to another measurement in ArcGIS on the original data source may have caused this unwanted error. To correct the error, values less than zero were coded as zeros, and the values greater than one were coded as one. This is an assumption that would need to be clarified with the original source of the data, otherwise the variables that make up these observations may need to be omitted. The modified human-use variables were kept, but may be systemically incorrect.

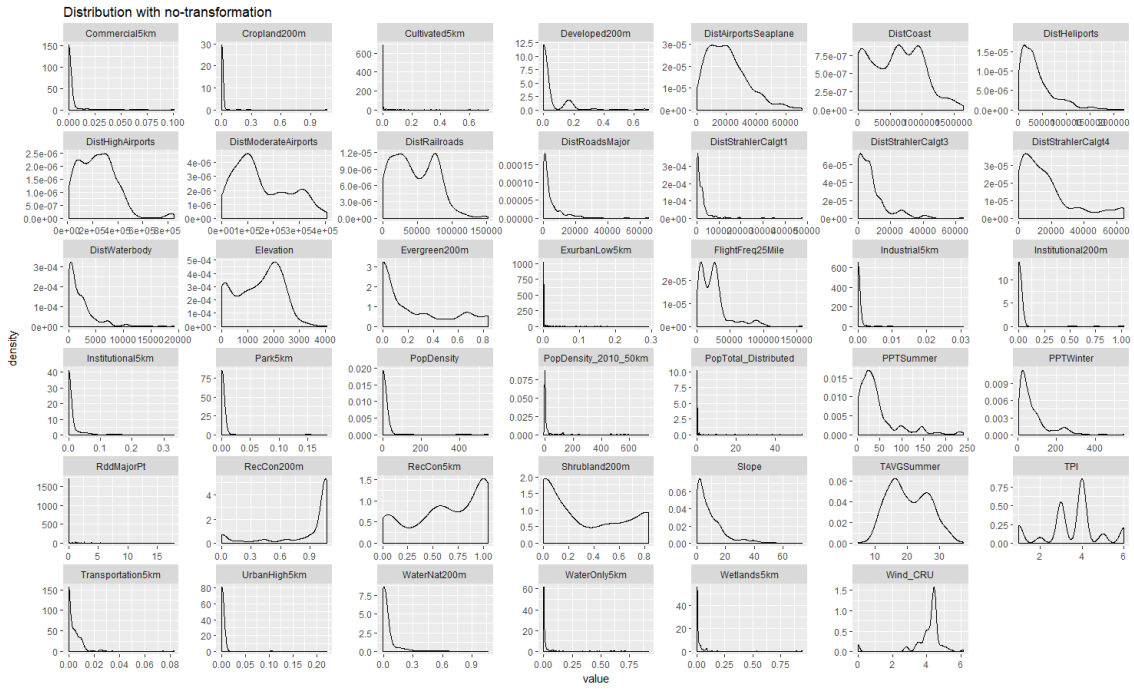

Figure 2. This figure shows the distribution of all variables in the dataset with no transformation. Most are largely right-skewed with the exception of RecCon5km and RecCon200m

Sparsity and Near Zero Variance

Many of the variables in the dataset had very few unique values relative to the number of samples. One of the methodologies to resolve this issue is called near-zero-variance function using an R package called 'caret' that examines the variables that have "few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value was very large." Most of the 200 meter land-cover variables were too sparse to remain in the dataset without causing issues in cross-validation so nineteen of the 200-meter resolution variables were eliminated. Table 4 breaks down the variables by column name in the dataset, the 'Frequency ratio' and the 'Percent Unique'. The first variable, listed in table 4, 'Snow200m', had 502 observations recorded as 0.0 and two observations of 0.5, one observation of 0.167 and one observation of 0.34. This resulted in a fre-

quency ratio between the most common and second most common values of 502 to 2, which reduces to 251 to 1, Therefore the Snow200m variable's frequency ratio is 251. The 'Percent Unique' value is the number of unique number levels divided by total observations multiplied by 100. 'Snow200m' had a total of 4 possible unique values, 0.0, 0.16667, 0.3333, and 0.5, so there are $\frac{4}{506} * 100 = 0.791\%$ unique values.

Multi-collinearity

The final issue is a majority of variables show multi-collinearity. For example, variables that measure the amount of Deciduous, Evergreen, Forest, and Mixed Forest at 200m are very similar, highly collinear, and if put in a regression model one would likely mask the importance of any another.

To correct this issue, geospatial variables that were highly correlated with each other were removed. A cutoff value of 0.75 and 0.50 were explored. When using 75% as the cutoff, 26 variables were removed. When using 50% as the cutoff, 32 variables were removed.

Furthermore, it should be noted, that the greatest single geospatial variable correlation with any of the L90 $\frac{1}{3}$-octave frequency bands is with the 16th $\frac{1}{3}$ octave band, correlated with a negative correlation with Shrubland5km of -0.56 and a positive correlation of 0.47 with 'PPTNorms' which was the average yearly precipitation. Although correlation does not mean causality, it suggests the data collected had larger L90f1 values associated with higher average precipitation values, and conversely lower L90f1 values were associated with higher amounts of Shrubland5km. These correlations were calculated on the entire data set as a general correlation-based feature approach to reduce the amount of variables in the data.

Table 4. Variables with few unique values and ratios of most common value to second most common value were very large (i.e near-zero-variance variables)

	Variables	Frequency Ratio	Percent Unique	Zero Variance	Near Zero Variance
1	Snow200m	251	0.791	FALSE	TRUE
2	Cropland200m	250	0.988	FALSE	TRUE
3	Park200m	250.000	1.190	FALSE	TRUE
4	UrbanLow200m	250.000	0.988	FALSE	TRUE
5	Commercial200m	249	1.190	FALSE	TRUE
6	ExurbanHigh200m	248.000	1.780	FALSE	TRUE
7	Suburban200m	246.000	1.780	FALSE	TRUE
8	Timber200m	246.000	2.570	FALSE	TRUE
9	ExurbanLow200m	244.000	2.370	FALSE	TRUE
10	Cropland5km	239	4.150	FALSE	TRUE
11	Wet200m	163	2.370	FALSE	TRUE
12	Pasture5km	124.000	1.780	FALSE	TRUE
13	Institutional200m	123.000	1.580	FALSE	TRUE
14	Snow5km	118.000	4.940	FALSE	TRUE
15	Mixed200m	117.000	3.560	FALSE	TRUE
16	Transportation200m	112.000	6.520	FALSE	TRUE
17	WaterHum200m	93	4.940	FALSE	TRUE
18	PopDensity	87.400	7.120	FALSE	TRUE
19	PopTotal	87.400	7.120	FALSE	TRUE
20	Cultivated200m	81.500	0.791	FALSE	TRUE
21	Park5km	66.600	3.950	FALSE	TRUE
22	Industrial5km	59	1.980	FALSE	TRUE
23	UrbanHigh5km	58.600	2.960	FALSE	TRUE
24	Institutional5km	57.400	8.890	FALSE	TRUE
25	WaterHum5km	50.100	9.880	FALSE	TRUE
26	Built200m	47.800	8.100	FALSE	TRUE
27	Extractive200m	35.400	9.490	FALSE	TRUE
28	MixedForest200m	32.500	0.791	FALSE	TRUE
29	ExurbanHigh5km	31.200	8.500	FALSE	TRUE
30	ExurbanHigh5km__1	31.200	8.500	FALSE	TRUE
31	Wet5km	27.900	7.310	FALSE	TRUE
32	Commercial5km	27.800	4.940	FALSE	TRUE
33	Timber5km	21.400	9.880	FALSE	TRUE
34	Deciduous200m	20.700	1.190	FALSE	TRUE

Zero values

While plotting the values of sound across frequencies it became apparent there were a few sites with a value of 0.0 for the two first $\frac{1}{3}$ octave frequencies. These values are most likely missing data rather than an actual level of 0.0 because the sound levels for the same site's third-$\frac{1}{3}$-octave band frequencies are much higher than 0. The sites that had zero values for these frequencies are noted in Table 5. Table 5 shows the sites with zero values for L90f1 and L90f2 are BADL, CANY, LAME, MORU, and MUWO. Although park abbreviations were not provided in the given dataset, most abbreviations were found on a website for the National Park Service. If a result came up for that abbreviation it was assumed to be the right acronym. Some of the abbreviations were not found so they were kept as acronyms and a best guess is supplied (for example BLMNV was not found, but it may be Bureau of Land Management, Nevada). The sites in Table 5 are most likely a lack of information than an actual value. Instead of throwing these 18 observations away, data was imputed: the missing values of L90f1 became L90f3 +2 decibels, and L90f2 became L90f3 +1 decibels. This was based on the general relationships in all the data showing these values were generally greater than the subsequent frequencies.

It was not until the model building process with forward stepwise regression, that other incorrect zero-values were found. The following sites had zero values for historical wind values: CAHA002, CALO001, EVER001, EVER006, GOGA001, GOGA003, GOGA004, GOGA005, GUIS001, GUIS002, GUIS003, WRBR001, WRBR002. The sites identified as WRBR were the Wright Brothers Memorial in Kitty Hawk North Carolina where Orville and Wilbur Wright began flight testing their prototype aircraft, and it is public knowledge that the winds are always blowing in this environment. A low wind value, let alone a zero value, was concerning. Further investigation

siteID	Latitude	Longitude	Elevation	Wind_CRU	DistAirportsAllMotorized	DistAirportsSeaplane	DistCoast	FlightFreq25Mile	WaterNat5km
WRBR001	36.01807	-75.66445	2	0	725.6032	725.6032	498.492	3329	0.6898793
WRBR001	36.01807	-75.66445	2	0	725.6032	725.6032	498.492	3329	0.6898793
WRBR002	36.01728	-75.67393	5	0	201.2461	201.2461	619.049	4141	0.6805943
WRBR002	36.01728	-75.67393	5	0	201.2461	201.2461	619.049	4141	0.6805943

Figure 3. The site WRBR001 and WRBR002 (WRight BRothers Memorial, Kitty Hawk North Carolina) were two of thirteen sites with zero-values for 'Wind_CRU' that were imputed with near match values. Other sites were CAHA, Cape Hatteras National Seashore, CALO, Cape Lookout National Seashore, EVER, Everglades National Park, GOGA, Golden Gate National Recreation Area, and GUIS, Gulf Islands Seashore.

revealed all the sites missing wind values were National Seashore sites. See figure 3 for a birds-eye view of the siteIDs for Wright Brothers Memorial. The coast is just visible in the graphic, as it is just under 500 meters away from site 'WRBR001'. The sites with missing wind values were imputed with a value of 4.7 based on other sites that had similar Distance to Coast values and Elevations. As these were not based on real historical data, future researchers should use a generalizable geospatial database to ensure the data is available for sites outside the United States.

Sites with multiple significantly different observations (MUWO0001)

While exploring sites that had recorded observations of 0.0 for the first octave-band L90f1, it became apparent some sites had multiple observations in the dataset with very different values. Muir Woods (MUWO) National Monument and Canyonlands (CANY) National Park were two sites with multiple observations with great differences in value. The problem this presents is the reality of the limitations of the

Table 5. Sites with zero-values for first and second $\frac{1}{3}$-octave band frequencies before imputation. BADL is Badlands National Park, South Dakota. CANY is Canyonlands National Park, Utah. LAME is Lake Mead, National Recreation Center, NV. MORU is Mount Rushmore National Memorial, South Dakota. MUWO is Muir Woods National Monument, California.

#	siteID	L90f1	L90f2	L90f3	L90f4	L90f5
1	BADL001	0	0	31	29.9	28
2	CANY001	0	0	24	23	21.2
3	CANY004	0	0	20.1	19	17.2
4	CANY004	0	0	23.2	21.9	19.8
5	CANY005	0	0	16.2	15.6	13.9
6	CANY006	0	0	19.6	18.2	17
7	CANY007	0	0	27.2	24.6	21.6
8	CANY007	0	0	19.8	18.8	18
9	CANY009	0	0	31.3	29.2	27
10	LAME001	0	0	44.6	43.8	42.2
11	LAME014	0	0	48.1	47.3	46.2
12	MORU001	0	0	38.9	40.1	40.9
13	MUWO001	0	0	30.2	30.8	32.1
14	MUWO001	0	0	26.6	26.9	27
15	MUWO002	0	0	37.4	40.5	41.3
16	MUWO004	0	0	24.9	25.7	26.2
17	MUWO004	0	0	26.4	27	27.6
18	MUWO005	0	0	26.5	27.2	27.9

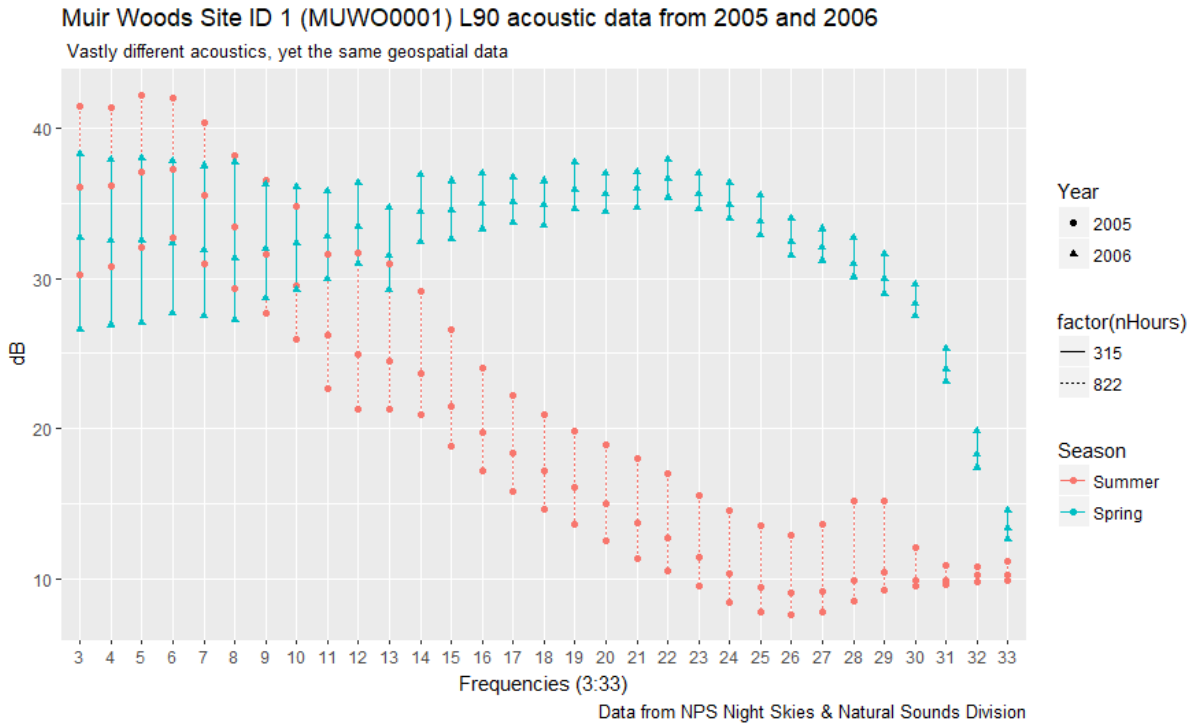

Figure 4. Muir Woods Site-1 (MUWO001) vastly different acoustics between Spring 2005 (hours = 315) and Summer 2006 (hours = 822). The top of the line represents L_{10}, the point which 10% of the observations were louder, and the bottom represents the L_{90}, level at which 90% of the observations were louder.

dataset. Despite having the exact same values across all the geospatial variables, the only thing that changed was 'year' and 'season', and the results were drastically different. For example, Muir Woods siteID 1, MUWO001, results show sound exceeded 10 decibels 90% of the time for the 22nd-$\frac{1}{3}$ octave band in summer 2005, yet exceeded 35 decibels 90% of the time in the same frequency in spring 2006. This is a difference of 25 decibels for the exact same location. So no matter how precise the model can predict using the given data, each location is prone to differences in time. Reference Figure 4 for visual of Muir Woods sound-frequency data.

Summary

Data exploration revealed a possible reason why the nearest matching algorithm [5] performed poorly for the 'shrubland' category due to that category being more sampled and thus having more variety. It revealed that the national parks are primarily of two different Land Category Land Use types (52 - Shrub and 42 - Evergreen Forest) so that when extrapolating the resulting models to areas of different Land Categories the results will most likely not be relevant (would not be advisable to a Snowy Tundra, or Barren Dessert). It revealed systematic errors like negative distances, and values greater than 1.0 and less than 0.0 for proportions, and recommended future corrections from the original data source (National Park Service) and in the absence of information made and annotated assumptions. Through exploration of the sparse variables, near-zero variance variables, and highly multi-collinear variables, the original number of geospatial variables was decreased from approximately 100 to 60 variables. It is also noted that sites did not vary much from Season to Season, except for the sites Muir Woods, CA and Canyonlands, Utah. Imputations for zero-values like the wind_CRU and first and second one-third octave band were discussed.

V. Analysis

Introduction

The analysis chapter will discuss how the models with each method were built, with detailed explanation on forward stepwise regression, backward stepwise regression, and exhaustive (maximum of eight variables) stepwise regression. It also includes conditional inference trees (a single decision tree model), and random forests methodologies to 1) see if able to replicate previous NPS findings using random forests [4, 3, 2, 1], 2) provide a reference for how well linear regression performs in comparison, 3) give sponsors a computational estimation on how random-forests performs, and 4) provide future researchers a baseline on the minimum performance level needed to improve based on random forests.

The overall training model results are presented as standard regression tables. Forward stepwise Regression is Table 6 for the first five one-third octave frequencies and Table 7 for the sixth through 10th one-third frequencies. Backward stepwise regression training model metrics are in Table 8 for the first five one-third octave frequencies, and Table 9 for sixth through 10th one-third octave frequencies. Exhaustive regression trainig model results are in Table 13 for the first five one-third octave frequencies, and Table 14 for frequencies 6-10.

These frequency-specific training-data derived models were then applied to the hold out test data, and then again to all the data (training and test). The results of applying the training-data derived Backwards Stepwise Regression model to the test data and all data are in Table 11 for the first five frequencies and Table 12 for the second five frequencies.

The results of applying the training-data derived Exhaustive Regression model to the test data and all data are in Table 16 for the first five one-third octave frequencies and Table 17 for the second five one-third octave frequencies.

Model Building

The data were initially split using 50% training, and 50% for testing the resulting training model in just linear regression. After developing and saving the initial linear models from the 50/50 split, the data were split 75% training and 25% testing for better comparison using all models: linear regression, conditional inference trees, and random forests. The initial models formed from the 50/50 split informed on what variables to keep. The 75% and 25% split and subsequent training tailored the coefficient values of these variables. 10-fold cross validation techniques were performed using the R package 'caret' to predict how well the models would then perform on test data. The seed-values to create the random repeat cross-validation were set to a constant to help compare different models. The results of each linear regression model is presented in several standard regression tables. To ease interpretation of the models, each of the frequency-specific models to include standard deviations for confidence intervals on the coefficients are formulated for $L_{90}f_1$ through $L_{90}f_{10}$ in equations (1) through (10) using the results of backwards stepwise regression and exhaustive regression. The results of backwards stepwise regression were very similar to exhaustive regression, especially on the frequencies where eight variables or fewer were needed. At about the fifth $\frac{1}{3}$ octave band frequency, the results of backwards stepwise regression achieved better model results to exhaustive regression because exhaustive regression was capped at eight variables and backwards regression was allowed to search best combinations up to and including fifteen variables. Allowing

exhaustive regression to compute up to fifteen variables would have taken multiple years using the given resources. Future research on speeding up exhaustive regression processes would be very beneficial in combined with more generic databases and up-to-date validated comprehensive datasets.

Method for Selecting the number of Forward and Backward Stepwise Regression Variables

Table 6 and table 7 shows the variables in a regression model for the first ten L_{90} band $\frac{1}{3}$ octave band frequencies using forward-stepwise regression with a maximum of 15 variables. Then, a multiple regression model was fit to each frequency independently, so a total of 150 models were assessed for forward stepwise regression. All of this took a standard laptop computer a couple seconds to calculate. The number of predictors chosen for each regression problem was decided by taking the minimum number of variables to do either of the three tasks: maximize the $AdjR^2$, minimize the Bayesian Information Criterion BIC, or minimize the Mallows C_p. It turns out the number of variables to minimize the Bayesian Information Criterion always was the determining factor. From these statistics, the minimum number of variables for BIC was taken as the best approximation for a good model without becoming over-fit and non-generalizable for predicting. The number of geospatial variables chosen also varied for each frequency-specific model. For example, the comparison of the three model metrics for selecting the number of variables to model $L_{90}f_1$ can be seen in figure 5 as eight variables. In table 6, there are eight coefficients listed under the dependent variables $L_{90}f_1$ (not including intercept). In contrast, $L_{90}f_3$ was best approximated at 11 variables. Table 6 shows the coefficient values for each of the first five $\frac{1}{3}$ octave band frequencies and standard errors.

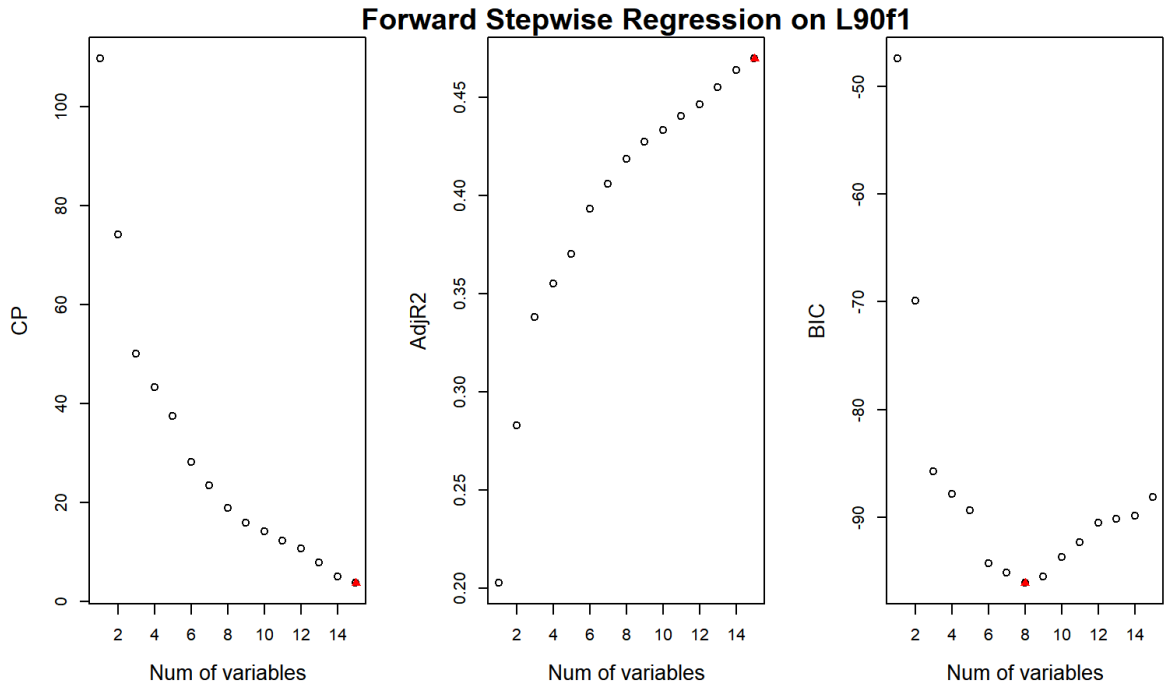

Figure 5. Method for choosing the number of variables in forward and backward stepwise regression–picking the number of variables needed to minimize the BIC statistic. This particular example is forward stepwise regression for L90f1, but the graphs are similar enough for backwards regression and all ten octave frequencies that all twenty graphs do not need to be displayed

Forward Stepwise Regression Variables

In Table 6, one can see that forward stepwise regression added some variables that were insignificant while forward stepping sequentially through the variables, for example, 'DistStrahlerCalgt3'. Forward Stepwise regression indicates the following variables were important first through tenth one-third octave band frequencies: Forest200m, Developed200m, HIHerbaceous200m, WaterNat200m, WaterOnly200m, RecCon5km, WaterOnly5km, DistStrahlerCalgt3, DistCoast, DistHeliports, FlightFreq25Mile, PopDensity_2010_50km, Elevation, SeasonSpring, Barren200m, DistAirportsSeaplane, DistRoadsMajor, RddMajor5km. Variables unique to first through fifth octave band frequencies (not in 6th through 10th): Forest200m, DistStrahlerCalgt3. Variables unique to sixth through tenth octave band frequencies (not in 1st through 5th): Elevation, SeasonSpring, Barren200m, DistAirportsSeaplane, DistRoadsMajor, RddMajor5km. Variables in both: Developed200m, HIHerbaceous200m, WaterNat200m, WaterOnly200m, RecCon5km, WaterOnly5km, DistCoast, DistHeliports, FlightFreq25Mile and PopDensity_2010_50km. Of importance to note, table 6 reveals the variables DistCoast, and PopDensity_2010_50km are not significant in some of the frequencies. DistStrahlerCalgt3 is not significant for one frequency. So to correct this issue one must 'step-back' or use a combination of forward and backward stepwise regression to get rid of the no longer significant variables. This was not done because this is only forward stepwise regression, and the exhaustive regression was used later. Furthermore, because backward stepwise regression performed better, the rest of chapter flow after table 7, focuses on just backward stepwise regression.

Table 6. Forward Stepwise Regression on the 1st through 5th 1/3 octave frequencies

	L90f1	L90f2	L90f3	L90f4	L90f5
Forest200m	-3.2^{***}	-2.9^{***}			
	(1.2)	(1.1)			
Developed200m				11.6^{***}	12.8^{***}
				(3.6)	(3.7)
HIHerbaceous200m	6.9^{**}	8.3^{***}	11.2^{***}	13.6^{***}	15.2^{***}
	(2.7)	(2.6)	(2.7)	(2.8)	(2.9)
WaterNat200m			8.6^{***}	9.6^{***}	10.0^{***}
			(3.1)	(3.2)	(3.4)
WaterOnly200m	-13.2^{***}	-15.1^{***}	-17.6^{***}	-15.7^{***}	-16.5^{***}
	(4.7)	(4.2)	(4.4)	(4.7)	(4.9)
RecCon5km	-3.6^{***}	-3.7^{***}	-3.5^{***}	-4.0^{***}	-4.4^{***}
	(1.3)	(1.2)	(1.3)	(1.3)	(1.4)
WaterOnly5km	12.2^{***}	15.3^{***}	15.5^{***}	13.6^{***}	14.1^{***}
	(3.4)	(2.6)	(2.7)	(3.1)	(3.2)
DistStrahlerCalgt3	0.1				
	(0.1)				
DistCoast				-0.001	-0.001
				(0.001)	(0.001)
DistHeliports		-0.03^{***}	-0.03^{***}	-0.03^{**}	-0.03^{**}
		(0.01)	(0.01)	(0.01)	(0.01)
FlightFreq25Mile			0.1^{***}	0.1^{***}	0.1^{***}
			(0.02)	(0.02)	(0.03)
PopDensity_2010_50km	0.02^{***}	0.02^{***}	0.01^{**}	0.003	0.01
	(0.003)	(0.003)	(0.004)	(0.005)	(0.005)
Constant	34.7^{***}	35.3^{***}	31.9^{***}	31.1^{***}	30.2^{***}
	(1.1)	(1.0)	(1.3)	(1.5)	(1.6)
Press	8525	7821	7753	8267	9116
MAE	4.32	4.1	3.97	4.1	4.37
MdAE	3.73	3.33	2.92	2.99	3.31
Adjusted R^2	0.3	0.4	0.4	0.5	0.5
Residual Std. Error	5.7	5.4	5.4	5.6	5.8
F Statistic	18.3^{***}	23.4^{***}	23.0^{***}	23.0^{***}	24.7^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7. Forward Stepwise Regression on the 6th through 10th 1/3 octave frequencies

	L90f6	L90f7	L90f8	L90f9	L90f10
Elevation			−0.001	−0.001*	
			(0.001)	(0.001)	
SeasonSpring			−2.7**	−2.8**	−2.9**
			(1.4)	(1.4)	(1.4)
Barren200m				3.5	
				(3.2)	
Developed200m	15.2***	17.2***	14.2***	13.1**	14.0***
	(3.9)	(4.1)	(5.1)	(5.1)	(5.0)
HIHerbaceous200m	15.6***	15.5***	14.2***	12.8***	11.9***
	(3.1)	(3.2)	(3.4)	(3.5)	(3.4)
WaterNat200m	11.0***	12.4***	11.8***	11.4***	11.7***
	(3.6)	(3.8)	(4.0)	(4.1)	(4.0)
WaterOnly200m	−19.0***	−22.1***	−19.2***	−18.4***	−17.4***
	(5.2)	(5.3)	(5.6)	(5.8)	(5.6)
RecCon5km	−4.2***	−4.7***	−4.7***	−4.2**	−4.7***
	(1.5)	(1.5)	(1.7)	(1.7)	(1.7)
WaterOnly5km	15.0***	16.3***	14.7***	13.3***	14.3***
	(3.4)	(3.3)	(3.8)	(4.0)	(3.7)
DistAirportsSeaplane					−0.1***
					(0.04)
DistCoast	−0.001				−0.002**
	(0.001)				(0.001)
DistHeliports	−0.03***	−0.04***	−0.04***	−0.04***	
	(0.01)	(0.01)	(0.01)	(0.01)	
DistRoadsMajor				−0.1**	
				(0.1)	
FlightFreq25Mile	0.1***	0.1***	0.1***	0.1***	0.1***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
PopDensity_2010_50km	0.01	0.01			
	(0.01)	(0.01)			
RddMajor5km			0.02**	0.02***	0.02***
			(0.01)	(0.01)	(0.01)
Constant	28.3***	26.4***	26.6***	26.0***	24.3***
	(1.7)	(1.6)	(1.9)	(1.9)	(1.9)
Press	10098	11189	12375	12744	12168
MAE	4.6	4.9	5.1	5.2	5.2
MdAE	3.6	3.8	4.3	4.1	4.3
Adjusted R^2	0.5	0.5	0.5	0.5	0.5
Residual Std. Error	6.1	6.5	6.8	6.9	6.8
F Statistic	25.5***	30.1***	25.8***	22.0***	25.6***

Note: *p<0.1; **p<0.05; ***p<0.01

Backward Stepwise Regression Variables

In tables 8 and 9, one can see that backward stepwise regression did not add any insignificant variables, unlike the forward method. The final results are also very similar to the variables chosen by exhaustive regression, displayed later in this chapter.

The results of Backward Stepwise Regression used the following 19 variables (SeasonSummer and SeasonFall counted as one variable), in their fit equations for the ten one-third octave band frequencies: Slope, Barren200m, Forest200m, Shrubland200m, WaterNat200m, WaterOnly200m, Wetlands200m, Barren5km, Transportation5km, RecCon5km, WaterOnly5km, DistCoast, DistHeliports, FlightFreq25Mile, Wind_CRU, TPI, SeasonSummer, SeasonFall, DistRailroads, RddMajorPt. Variables unique to first through fifth octave band frequencies (not in 6th through 10th): Slope, Barren5km, DistCoast, Wind_CRU, TPI. Variables unique to sixth through tenth octave band frequencies (not in 1st through 5th): SeasonSummer, SeasonFall, DistRailroads, RddMajorPt. Variables in both : Barren200m, Forest200m, Shrubland200m, WaterNat200m, WaterOnly200m, Wetlands200m, RecCon5km, Transportation5km, WaterOnly5km, DistHeliports, and FlightFreq25Mile.

Table 8. Backward Stepwise Regression on the 1st through 5th 1/3 octave frequencies

	L90f1	L90f2	L90f3	L90f4	L90f5
Slope				-0.1^{**}	-0.1^{**}
				(0.04)	(0.05)
Barren200m				-12.4^{***}	-16.6^{***}
				(3.4)	(3.4)
Forest200m	-10.7^{***}	-9.8^{***}	-9.8^{***}	-12.9^{***}	-16.4^{***}
	(1.8)	(1.7)	(1.8)	(2.2)	(2.2)
Shrubland200m	-8.1^{***}	-7.8^{***}	-8.6^{***}	-13.3^{***}	-17.7^{***}
	(1.8)	(1.8)	(1.8)	(2.2)	(2.2)
WaterNat200m		8.2^{***}	9.1^{***}	10.1^{***}	13.3^{***}
		(3.0)	(3.1)	(3.2)	(3.3)
WaterOnly200m	-17.1^{***}	-23.1^{***}	-23.7^{***}	-30.7^{***}	-36.4^{***}
	(4.3)	(4.3)	(4.4)	(5.0)	(5.2)
Wetlands200m	-11.8^{***}	-9.8^{***}	-10.2^{***}	-14.6^{***}	-16.7^{***}
	(2.8)	(2.6)	(2.7)	(3.0)	(3.2)
Barren5km	-17.2^{***}	-15.8^{***}	-15.1^{***}		
	(3.6)	(3.5)	(3.6)		
Transportation5km		97.8^{***}	103.9^{***}		
		(35.4)	(36.2)		
RecCon5km				-3.3^{**}	
				(1.3)	
WaterOnly5km	12.3^{***}	15.5^{***}	14.6^{***}	15.1^{***}	18.0^{***}
	(2.8)	(2.5)	(2.6)	(2.8)	(2.8)
DistCoast	-0.004^{***}				
	(0.001)				
DistHeliports				-0.03^{***}	-0.03^{***}
				(0.01)	(0.01)
FlightFreq25Mile	0.1^{***}	0.1^{***}	0.1^{***}	0.1^{***}	0.1^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Wind_CRU	2.5^{***}				
	(0.8)				
TPI				-0.8^{***}	-0.9^{***}
				(0.3)	(0.3)
Constant	30.2^{***}	36.1^{***}	35.3^{***}	44.1^{***}	43.7^{***}
	(3.6)	(1.3)	(1.4)	(2.0)	(2.1)
Press	7893	7362	7673	8024	9139
MAE	4.14	4.01	4.09	3.97	4.36
MdAE	3.39	2.92	3.24	2.98	3.34
Adjusted R^2	0.4	0.4	0.4	0.5	0.5
Residual Std. Error	5.4	5.2	5.4	5.5	5.8
F Statistic	18.0^{***}	21.4^{***}	21.3^{***}	20.9^{***}	22.9^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9. Backward Stepwise Regression on the 6th through 10th 1/3 octave frequencies

	L90f6	L90f7	L90f8	L90f9	L90f10
SeasonSummer			3.0***	2.9***	3.0***
			(0.9)	(1.0)	(0.9)
SeasonFall			4.1***	3.9***	3.9***
			(1.4)	(1.4)	(1.4)
Barren200m	−15.9***	−17.0***	−12.8***	−13.0***	
	(3.6)	(3.8)	(4.1)	(4.1)	
Forest200m	−15.5***	−16.6***	−13.5***	−14.2***	−7.0***
	(2.4)	(2.6)	(2.9)	(2.9)	(2.2)
Shrubland200m	−17.0***	−18.5***	−15.0***	−15.8***	−9.7***
	(2.4)	(2.5)	(2.9)	(2.9)	(2.3)
WaterNat200m	11.9***	13.3***	13.3***	13.7***	10.7***
	(3.6)	(3.8)	(4.0)	(4.0)	(3.9)
WaterOnly200m	−35.1***	−38.2***	−31.9***	−33.2***	−22.9***
	(5.6)	(5.9)	(6.3)	(6.3)	(5.6)
Wetlands200m	−17.1***	−18.4***	−14.9***	−15.1***	−9.1***
	(3.3)	(3.5)	(3.8)	(3.9)	(3.3)
RecCon5km	−4.3***	−4.9***	−5.4***	−4.7***	−5.3***
	(1.5)	(1.6)	(1.7)	(1.7)	(1.6)
Transportation5km			122.8**	119.5**	150.7***
			(47.7)	(48.4)	(47.0)
WaterOnly5km	16.2***	16.6***	16.0***	16.1***	16.3***
	(3.1)	(3.3)	(3.4)	(3.4)	(3.4)
DistHeliports	−0.04***	−0.04***		−0.04***	
	(0.01)	(0.01)		(0.01)	
DistRailroads					−0.04**
					(0.02)
FlightFreq25Mile	0.1***	0.1***	0.1***	0.1***	0.1***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
RddMajorPt			0.6**	0.7**	0.9***
			(0.3)	(0.3)	(0.2)
Constant	40.6***	40.4***	32.3***	32.8***	26.4***
	(1.9)	(2.0)	(2.3)	(2.5)	(2.1)
Press	10047	11119	12346	12514	12126
MAE	4.6	4.8	5.1	5.2	5.1
MdAE	3.6	3.6	4.4	4.7	4.3
Adjusted R^2	0.5	0.5	0.5	0.5	0.5
Residual Std. Error	6.1	6.4	6.7	6.8	6.7
F Statistic	25.9***	27.7***	23.0***	22.1***	23.3***

Note: *p<0.1; **p<0.05; ***p<0.01

Backwards Linear Regression Variables, Order of Importance

According to backward stepwise regression, the most important variables (by t-values) in the regression are listed top to bottom in order of best to worst in table 10. Reading the chart top to bottom, the most important variable for L90f1 is Barren5km, followed by Forest200m, WaterOnly5km, FlightFreq25Mile, WaterOnly200m, Wetlands200m, Shrubland200m, DistCoast, and Wind_CRU. The most important variable for L90f2 is WaterOnly5km, followed by FlightFreq25Mile, Barren5km, Forest200m, WaterOnly200m, Shrubland200m, Transportation5km, Wetlands200m, WaterNat200m. The importance of WaterOnly at 200m and 5km persists for all ten one-third octave frequencies. The most important variables per frequency are mostly in agreement with exhaustive regression: Barren5km, WaterOnly5km, FlightFreq25Mile are some of the most important variables in Backward Stepwise Regression and Exhaustive Regression for the first ten octave frequencies.

Table 10. Variables of Importance across first ten 1/3 octave frequencies with Backward Stepwise Regression

f1	f2	f3	f4	f5
Barren5km	WaterOnly5km	WaterOnly5km	FlightFreq25Mile	Shrubland200m
Forest200m	FlightFreq25Mile	FlightFreq25Mile	WaterOnly200m	Forest200m
WaterOnly5km	Barren5km	Barren5km	WaterOnly5km	FlightFreq25Mile
FlightFreq25Mile	Forest200m	Forest200m	Forest200m	WaterOnly200m
WaterOnly200m	WaterOnly200m	WaterOnly200m	Shrubland200m	WaterOnly5km
Wetlands200m	Shrubland200m	Shrubland200m	Barren200m	Barren200m
Shrubland200m	Transportation5km	Transportation5km	Wetlands200m	DistHeliports
DistCoast	Wetlands200m	Wetlands200m	DistHeliports	Wetlands200m
Wind_CRU	WaterNat200m	WaterNat200m	RecCon5km	WaterNat200m
			WaterNat200m	TPI

f6	f7	f8	f9	f10
FlightFreq25Mile	FlightFreq25Mile	FlightFreq25Mile	FlightFreq25Mile	RddMajorPt
Shrubland200m	Shrubland200m	WaterOnly5km	WaterOnly5km	WaterOnly5km
Forest200m	Forest200m	RddMajorPt	WaterOnly200m	FlightFreq25Mile
WaterOnly200m	WaterOnly200m	WaterOnly200m	RddMajorPt	WaterOnly200m
WaterOnly5km	WaterOnly5km	RecCon5km	Shrubland200m	Transportation5km
Barren200m	Barren200m	Shrubland200m	Forest200m	RecCon5km
DistHeliports	DistHeliports	Forest200m	DistHeliports	DistRailroads
Wetlands200m	Wetlands200m	Transportation5km	WaterNat200m	Shrubland200m
RecCon5km	RecCon5km	WaterNat200m	Barren200m	WaterNat200m
WaterNat200m	WaterNat200m	Barren200m	Transportation5km	Forest200m

Backwards Linear Regression, Accuracy Per Frequency

The accuracy of the backward stepwise regression model per frequency is indicated in Tables 11 and 12. In general the training set will always perform better than the test set. Reading the table from left to right is the Root Mean Squared Error (RMSE), R-squared, and Mean Absolute Error computed for the Training, Test, and All the data respectively, per frequency. In general the lower frequencies did not vary as much as the greater valued frequencies, for example, the RMSE for L90f1 is 5.196, versus RMSE of 6.325 for L90f10. This was true for the test-data results as well, for example the RMSE for L90f1 test data is 5.724, versus the RMSE of 7.022 for L90f6.

Table 11. Training, Test, and Entire Dataset using Backwards Stepwise Regression on first five one-third frequencies (L90f1-L90f5)

Metrics	Training	Test	All
		L90f1	
RMSE	5.196	5.724	5.331
Rsquared	0.419	0.229	0.372
MAE	4.000	4.642	4.158
		L90f2	
RMSE	5.015	5.616	5.170
Rsquared	0.464	0.288	0.420
MAE	3.861	4.507	4.020
		L90f3	
RMSE	5.043	5.865	5.257
Rsquared	0.471	0.290	0.424
MAE	3.932	4.727	4.127
		L90f4	
RMSE	5.231	6.021	5.436
Rsquared	0.504	0.352	0.464
MAE	3.917	4.781	4.129
		L90f5	
RMSE	5.615	6.438	5.828
Rsquared	0.497	0.374	0.463
MAE	4.356	5.179	4.558

Table 12. Training, Test, and Entire Dataset using Backwards Stepwise Regression, continued for 6th-10th one-third octave frequencies (L90f6-L90f10)

Metrics	Training	Test	All
		L90f6	
RMSE	5.820	7.022	6.137
Rsquared	0.518	0.347	0.472
MAE	4.431	5.438	4.679
		L90f7	
RMSE	6.094	7.432	6.449
Rsquared	0.534	0.361	0.487
MAE	4.631	5.806	4.921
		L90f8	
RMSE	6.223	7.664	6.607
Rsquared	0.567	0.373	0.517
MAE	4.962	6.014	5.221
		L90f9	
RMSE	6.276	7.613	6.630
Rsquared	0.579	0.386	0.530
MAE	4.993	6.015	5.245
		L90f10	
RMSE	6.325	7.398	6.605
Rsquared	0.563	0.370	0.516
MAE	5.124	5.759	5.280

Backwards Linear Regression, 10-fold Cross Validation Results

The results of using 10-fold cross-validation to provide inferences on how well the frequency-specific backward regression models performs on resampled data, and is displayed in the boxplots of figure 6. One can see how the MAE and RMSE are similar to what tables 11 and 12 specified in one training sample, but these figures provide better visuals of the confidence interval on those results. For example, L90f10 was as poor as RMSE of 8 for one of the 10-fold resamples. It was also as good as achieving a mean absolute error of 3 decibels off for L90f2 for another 10-fold resample. The black dot specifies the mean. Gray dots are outliers. The least precise frequency to predict according to mean absolute error is at the top, L90f10, and the most precise is at the bottom, L90f3.

Backwards Linear Regression, Visualizing 16 Random Samples

In addition to the training model metrics provided so far by the backward regression models, it was desired to visualize individual sites predictions versus actuals using the backward and exhaustive regression, and later conditional inference tree and random forests. A function was created to randomly pull sixteen sites from the test data, and model the predictions using a specified method. The differences in values are explicit from the different symbols used in the plot. In addition, the highest difference per site is identified in red with a text label indicating the absolute difference. If the predicted value is greater than the observed value, the blue triangle (predictions) will be louder than the black circle (observations). Figure 6 shows the results of applying the best backwards regression frequency-specific models to sixteen random observations.

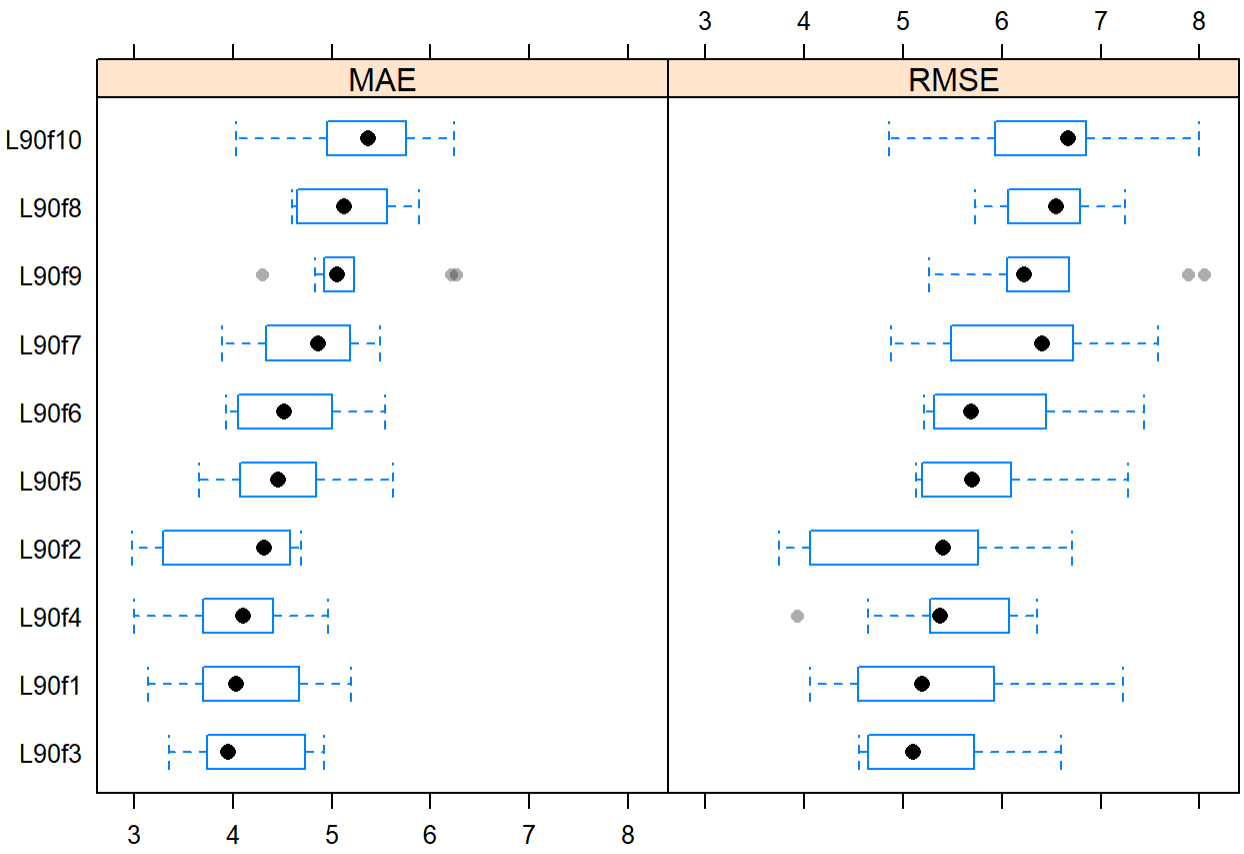

Figure 6. Resample Results for Backwards Regression Fit across all Frequencies. The least accurate frequency to predict according to mean absolute error (MAE) is at the top, L90f10, and the most accurate is at the bottom, L90f3

In Figures 7 and 10 sixteen random NPS observations are predicted using the backwards and exhaustive regression fit. The *maximum* difference over all frequencies is reported for each observation. One can see both models seem to capture the unique shape of each of the parks sound by frequency. Also important to note, is even though the models predict a mean absolute error of around 5 decibels, a majority of the random sites have at least one frequency that was missed by more than 5 decibels. In the graphic, predictions are blue circles, and the actual acoustic measurement are black triangles. A red annotation appears at the largest difference in predictions and actual values, and a text label with the difference in decibels appears over the red circle. So for example, reading the first top left figure in figure 7 is 'GLCA012' which is site 12 Glen Canyon National Park Service. This particular park was well estimated as the points are very close together, but the worst prediction was 4.7 decibels above the actual value for the first $\frac{1}{3}$ octave frequency. We therefore surmise that this park was well approximated. Another site in figure 7, YELL019, Yellowstone National Park site 19, was under-predicted by 17.4 decibels on the 7th $\frac{1}{3}$ octave frequency. This site was louder than expected given its geospatial variables. BRCA0001, site 1 of Bryce Canyon was over-predicted by 18.1 decibels on the first $\frac{1}{3}$ octave frequency. BRCA was quieter than predicted. Of particular note, BRCA was very quiet overall, as one can see the actuals were all less than 15 decibels, the lowest in the overall figure 7. This is something previous NPS studies had noted with all regression techniques. The quietest places will be overestimated, and the loudest places will be underestimated, in the "regression towards the mean" in their models using random forests as well [2, 1].

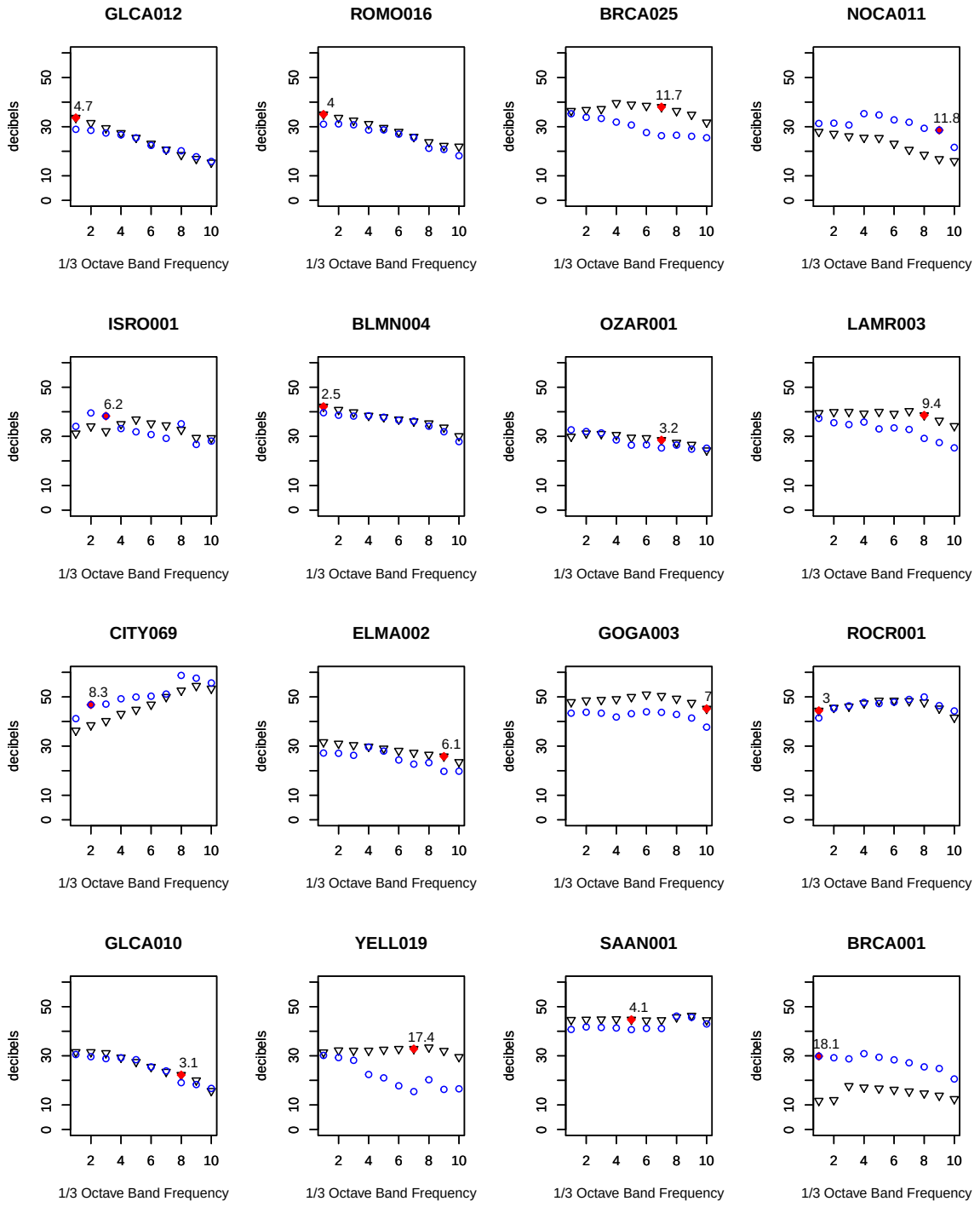

Figure 7. Random Site Predictions using Backwards Regression Fits. Predictions are blue circles, actuals are black triangles. A red annotation appears at the largest difference in predictions and actual values, and a text label with the difference in decibels appears over the red circle.

Exhaustive Linear Regression

Unlike stepwise regression, exhaustive regression is slow. Figure 8 shows how the time to compute models appears to be exponentially increasing with each additional variable. Although not recorded in the graph, over nine variables were taking hours to compute without completion.

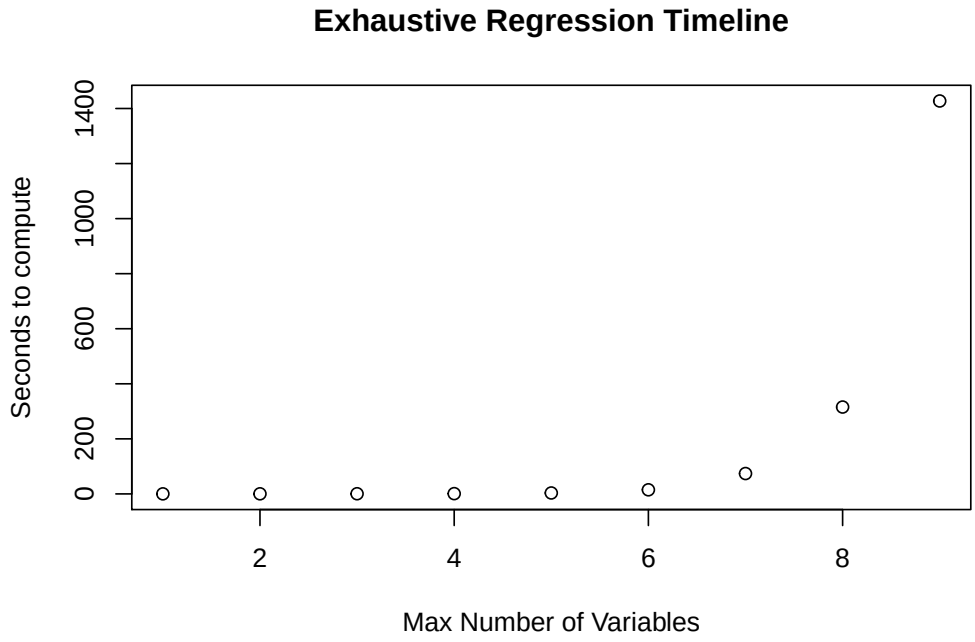

Figure 8. Exhaustive Regression Timeline. More than 9 variables took hours without completion.

Exhaustive Regression Variables

Only 14 variables were needed to model a linear regression on the first ten $\frac{1}{3}$ octave band frequencies. The results of Exhaustive Regression used the following variables in their fit equations for first through tenth one-third octave band frequencies:

Barren200m, Developed200m, HIHerbaceous200m, Barren5km, Transportation5km, WaterNat200m, WaterOnly200m, RecCon5km, WaterOnly5km, DistCoast, DistHeliports, FlightFreq25Mile, Wind_CRU, RddMajorPt. Variables unique to first through fifth octave band frequencies (not in 6th through 10th): Barren200m, Barren5km, DistCoast, Wind_CRU. Variables unique to sixth through tenth octave band frequencies (not in 1st through 5th): none. Variables in both: Developed200m, HIHerbaceous200m, WaterNat200m, WaterOnly200m, RecCon5km, Transportation5km, WaterOnly5km, DistHeliports, FlightFreq25Mile and RddMajorPt. The coefficients of each variable are located in tables 13 for first five frequencies, and 14 for the five frequencies.

In tables 13 and 14, one can see all variables chosen to be in the exhaustive regression model are significant, as one would expect. Comparing exhaustive regression model metrics in table 14 to backward stepwise model metrics in table 9 reveals backward stepwise regression outperforms exhaustive regression in the later frequencies. This may because exhaustive regression was capped at a maximum of eight variables. Exhaustive regression over eight variables should be use when computationally feasible for future research, but backwards regression appears to be a near substitute for the first five $\frac{1}{3}$ octave bands. Forward stepwise appears adds more insignificant variables in its search, and the model metrics do not appear as good as backwards regression.

Exhaustive Linear Regression Variables, Order of Importance

The most important variables in the exhaustive linear regression model are in table 15. These variables help inform a interactive data analytic application for the sponsor using a parallel coordinate visual to help select a range of values to find a best matching site, as described later in this chapter. The most imporant variables across all

Table 13. Exhaustive Regression on the 1st through 5th 1/3 octave frequencies

	L90f1	L90f2	L90f3	L90f4	L90f5
Barren200m	10.6***	9.2***			
	(3.0)	(2.9)			
Developed200m				12.3***	14.4***
				(3.3)	(3.5)
HIHerbaceous200m	12.2***	11.7***	12.2***	13.2***	15.1***
	(2.6)	(2.6)	(2.6)	(2.7)	(2.9)
Barren5km	−16.9***	−14.9***			
	(4.3)	(4.1)			
Transportation5km		102.6***			
		(34.9)			
WaterNat200m			9.1***	9.8***	10.4***
			(3.1)	(3.2)	(3.4)
WaterOnly200m			−15.7***	−16.4***	−16.4***
			(4.3)	(4.5)	(4.7)
RecCon5km			−3.0**	−3.7***	−4.2***
			(1.3)	(1.3)	(1.4)
WaterOnly5km	10.3***	11.0***	16.1***	15.2***	15.7***
	(2.2)	(2.1)	(2.7)	(2.9)	(3.0)
DistCoast	−0.004***	−0.004***			
	(0.001)	(0.001)			
DistHeliports			−0.03***	−0.03***	−0.03***
			(0.01)	(0.01)	(0.01)
FlightFreq25Mile	0.1***	0.1***	0.1***	0.1***	0.1***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Wind_CRU	3.0***	2.5***			
	(0.8)	(0.8)			
RddMajorPt			0.5***		
			(0.2)		
Constant	20.8***	21.3***	30.8***	29.9***	28.8***
	(3.1)	(3.0)	(1.3)	(1.3)	(1.4)
Press	7982	7455	7682	8242	9137
MAE	4.23	4.1	3.95	4.09	4.35
MdAE	3.38	3.34	2.91	2.91	3.05
Adjusted R^2	0.4	0.4	0.4	0.5	0.5
Residual Std. Error	5.5	5.3	5.4	5.6	5.9
F Statistic	21.4***	22.7***	23.7***	28.3***	30.2***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 14. Exhaustive Regression on the 6th through 10th 1/3 octave frequencies

	L90f6	L90f7	L90f8	L90f9	L90f10
Developed200m	16.9***	19.3***	21.8***		
	(3.7)	(3.9)	(4.1)		
HIHerbaceous200m	15.6***	16.1***	16.3***	17.5***	16.2***
	(3.0)	(3.2)	(3.4)	(3.4)	(3.3)
WaterNat200m	11.4***	12.7***	12.2***	13.4***	13.1***
	(3.6)	(3.8)	(4.0)	(4.0)	(3.9)
WaterOnly200m	−18.7***	−20.5***	−20.1***	−22.9***	−21.7***
	(5.0)	(5.2)	(5.6)	(5.6)	(5.5)
RecCon5km	−4.0***	−4.7***	−5.3***		
	(1.5)	(1.5)	(1.6)		
Transportation5km				156.1***	163.8***
				(48.6)	(47.3)
WaterOnly5km	16.4***	16.5***	16.7***	24.0***	24.2***
	(3.2)	(3.3)	(3.6)	(3.3)	(3.2)
DistHeliports	−0.04***	−0.04***	−0.04***	−0.05***	−0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
FlightFreq25Mile	0.1***	0.1***	0.1***	0.1***	0.1***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
RddMajorPt				1.3***	1.3***
				(0.3)	(0.2)
Constant	27.1***	25.9***	25.0***	19.2***	17.4***
	(1.5)	(1.5)	(1.6)	(1.0)	(1.0)
Press	10113	11212	12740	13488	12816
MAE	4.6	4.9	5.2	5.5	5.3
MdAE	3.6	3.8	4.2	4.8	4.8
Adjusted R^2	0.5	0.5	0.5	0.5	0.5
Residual Std. Error	6.2	6.5	6.9	7.1	6.9
F Statistic	31.3***	33.4***	32.9***	31.7***	32.5***

Note: *p<0.1; **p<0.05; ***p<0.01

ten frequencies of study appear to be FlightFreq25Mile, WaterOnly5km, Barren5km, Developed200m, and RddMajorPt. This is a mix of the aircraft flights taking place, the water and barren amount of space in 5 kilometers, the developed land within 200 meters, and the major road density within a certain area. These variables were consistently in the first three variables of importance across all ten frequencies. Its possible a 3, 4, or 5-variable model with just these variables could be formed at the cost of less accuracy on the training model, but perhaps more accuracy on global points.

Exhaustive Linear Regression, Accuracy per Frequency

The accuracy of the exhaustive linear regression model per frequency is indicated in Tables 16 and 17. This description is similar to that already described in 'Backwards Linear Regression, Accuracy per Frequency', with the exception that exhaustive performed a little bit better than backwards regression for some of the frequencies and a little worse than backwards regression for the frequencies where backwards had more variables in its model—namely the later frequencies. But it did not perform significantly better or worse. The models explain about 50% of the variance when built with the training data, and about 20-30% when applied to the test data.

Table 15. Variables of Importance across first five frequencies with exhaustive regression

f1	f2	f3	f4	f5
Barren5km	WaterOnly5km	WaterOnly5km	FlightFreq25Mile	FlightFreq25Mile
WaterOnly5km	FlightFreq25Mile	FlightFreq25Mile	WaterOnly5km	Developed200m
FlightFreq25Mile	Barren5km	RddMajorPt	Developed200m	WaterOnly5km
Wind_CRU	DistCoast	WaterOnly200m	WaterOnly200m	HIHerbaceous200m
DistCoast	Wind_CRU	HIHerbaceous200m	HIHerbaceous200m	RecCon5km
HIHerbaceous200m	HIHerbaceous200m	WaterNat200m	RecCon5km	WaterOnly200m
Barren200m	Transportation5km	RecCon5km	DistHeliports	DistHeliports
	Barren200m	DistHeliports	WaterNat200m	WaterNat200m

f6	f7	f8	f9	f10
FlightFreq25Mile	FlightFreq25Mile	Developed200m	WaterOnly5km	WaterOnly5km
Developed200m	Developed200m	FlightFreq25Mile	RddMajorPt	RddMajorPt
WaterOnly5km	WaterOnly5km	WaterOnly5km	FlightFreq25Mile	FlightFreq25Mile
HIHerbaceous200m	HIHerbaceous200m	RecCon5km	HIHerbaceous200m	HIHerbaceous200m
DistHeliports	DistHeliports	DistHeliports	DistHeliports	WaterNat200m
RecCon5km	RecCon5km	HIHerbaceous200m	WaterNat200m	DistHeliports
WaterOnly200m	WaterOnly200m	WaterNat200m	WaterOnly200m	Transportation5km
WaterNat200m	WaterNat200m	WaterOnly200m	Transportation5km	WaterOnly200m

Table 16. Exhaustive Regression (Maximum of 8-variables) applied to the Training, Test, and Entire Dataset using for the first five L90 $\frac{1}{3}$ octave frequencies— all measurements in decibels

Metrics	Training	Test	All
		L90f1	
RMSE	5.349	5.889	5.487
Rsquared	0.385	0.195	0.335
MAE	4.142	4.631	4.262
		L90f2	
RMSE	5.139	5.745	5.295
Rsquared	0.438	0.264	0.392
MAE	3.992	4.530	4.125
		L90f3	
RMSE	5.071	5.953	5.301
Rsquared	0.465	0.269	0.414
MAE	3.821	4.652	4.026
		L90f4	
RMSE	5.252	6.190	5.497
Rsquared	0.500	0.320	0.452
MAE	3.918	4.832	4.143
		L90f5	
RMSE	5.509	6.591	5.794
Rsquared	0.516	0.349	0.470
MAE	4.159	5.134	4.399

Table 17. Exhaustive Regression Maximum of 8-variables applied to Training, Test, and Entire Dataset continued for the second five L90 $\frac{1}{3}$ octave frequencies (6th through 10th)— all measurements in decibels

Metrics	Training	Test	All
		L90f6	
RMSE	5.756	6.985	6.082
Rsquared	0.529	0.354	0.481
MAE	4.401	5.363	4.637
		L90f7	
RMSE	6.015	7.383	6.379
Rsquared	0.546	0.368	0.498
MAE	4.586	5.686	4.857
		L90f8	
RMSE	6.342	7.720	6.707
Rsquared	0.550	0.369	0.502
MAE	4.904	5.932	5.157
		L90f9	
Metrics	Training	Test	All
RMSE	6.474	8.002	6.881
Rsquared	0.552	0.324	0.494
MAE	5.222	6.303	5.488
		L90f10	
RMSE	6.409	7.616	6.726
Rsquared	0.552	0.337	0.499
MAE	5.179	5.965	5.373

Exhaustive Linear Regression, 10-fold Cross Validation

The results of using 10-fold cross validation to provide inferences on how well the frequency-specific exhaustive regression performs on the resampled data, is displayed in the boxplots of figure 9. The MAE and RMSE found per frequency are similar to the tables 16 and 17 explained previously. Similar to backwards regression, the least accurate estimates according to mean absolute error are at the top, L90f10, and the most accurate estimates are located at the bottom, L90f3.

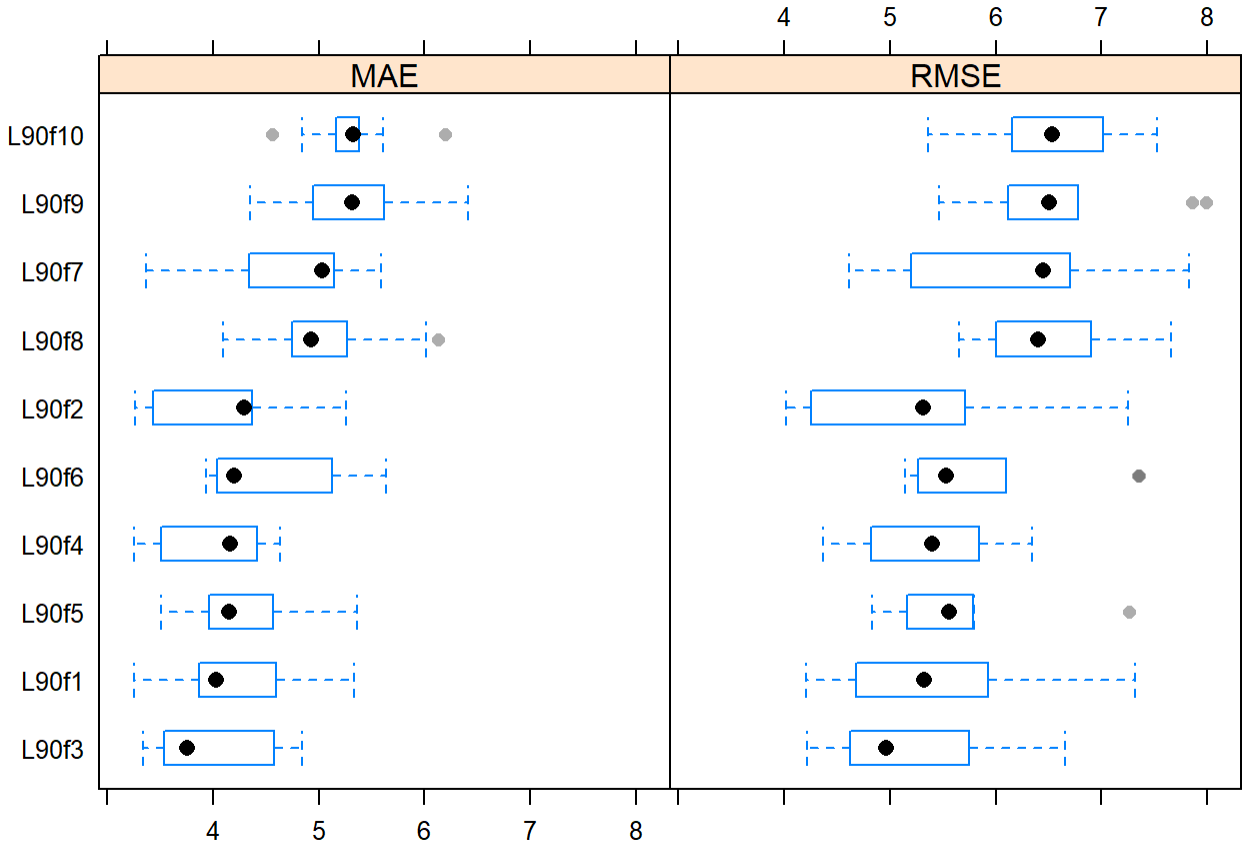

Figure 9. Resample results from Exhaustive Regression across all frequencies. Similar to backwards regression, the least accurate estimates according to mean absolute error are at the top, L90f10, and the most accurate estimates are located at the bottom, L90f3—all measurements are in decibels

Exhaustive Linear Regression, Visualizing 16 Random Samples

Similar to the explanation given for visualizing 16 random samples in the backward regression model, figure 10 shows the *maximum* difference over all frequencies is reported for each observation. Once again, even though the MAE is about 5 per frequency for exhaustive regression, a majority of the random sites–11 out of 16—-have at least one frequency that was missed by more than 5 decibels. In the graphic, predictions are blue circles, and the actual acoustic measurement are black triangles. A red annotation appears at the largest difference in predictions and actual values, and a text label with the difference in decibels appears over the red circle. The first top left figure in figure 10 is 'GLCA012' which is site 12 of Glen Canyon National Park Service. This park was also well estimated as it was in backwards regression, but exhaustive regression gives the worst prediction 6 decibels above the actual value for the first $\frac{1}{3}$ octave frequency. We therefore surmise that this park was well approximated. Another site in figure 10, YELL019, Yellowstone National Park site 19, was under-predicted by 19.2 decibels on the 8th $\frac{1}{3}$ octave frequency. This site was louder than expected given its geospatial variables. BRCA0001, site 1 of Bryce Canyon was over-predicted by 18.9 decibels on the first $\frac{1}{3}$ octave frequency.

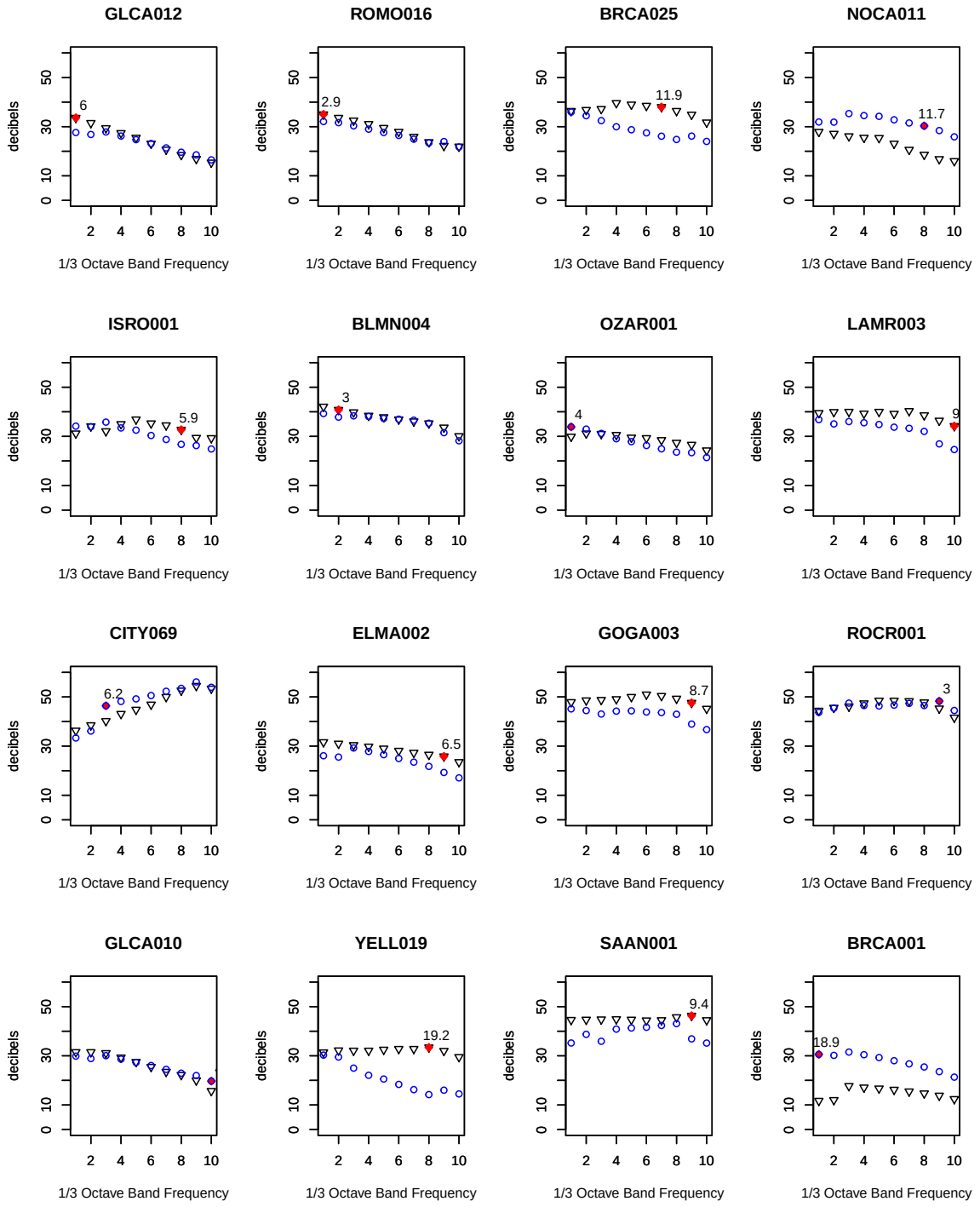

Figure 10. Random Site Predictions using Exhaustive Regression Fits. Predictions are blue circles, actuals are black triangles. A red annotation appears at the largest difference in predictions and actual values, and a text label with the difference in decibels appears over the red circle.

Discussion of Forward, Backward and Exhaustive Regression Results

For the first five one-third octave frequencies, both forward and backward stepwise regression indicated the following variables as important: Forest200m, WaterNat200m, WaterOnly200m, RecCon5km, WaterOnly5km, DistCoast, DistHeliports, and FlightFreq25Mile. Exhaustive regression also shared the following seven variables of importance with forward and backward regression: WaterNat200m, WaterOnly200m, RecCon5km, WaterOnly5km, DistCoast, DistHeliports, and FlightFreq25Mile. Most of these variables seem to be measures of water in an area, which is in agreement with other National Park Studies [3] which concluded water was an important variable: "This stems from the presence of many acoustic sources with low frequency content (e.g., wind, water, and transportation noise) and that the low frequency energy of any source propagates farther than high frequencies due to air absorption, diffraction, ground effects, etc." [2].

Variables Not Included in Either Regression Model

Two metrics in the literature review that stood out as not being included were distance to streams and population metrics. No regression models found 'DistStrahlerCalgt3'– the distance to streams with order three as significant. Other studies have shown that L_{90} or ambient background noise is influenced by the presence of rivers, waterfalls, water in general etc. One possible speculation which would need more time and research, is the Strahler Order isn't as important as the power and/or slope of the streams, which were in previous datasets [31] but not in the one studied. At least upon initial inspection, the observations that were most under-predicted seem to be related to parks that were likely close to a waterfall–such as NOCA, and PIRO.

In the case of site NOCA008–North Cascades National Park, Washington—a web search of the latitude/longitude revealed the site is called Ladder Creek Falls, which when explored using the Washington State Trails Association says this: "The falls and surrounding gardens quickly became a tourist attraction, illuminated at night with colored lights and livened up by music that was piped in on Friday evenings for visitors who came from Seattle for a night out," [81]. Therefore, the NOCA008 location was louder most likely because its main intent was to attract and entertain people with lights and music from dawn to dusk. The site is also directly behind an electric hydro-powered dam and on a pathway to a waterfall. These may contribute to the persistent overall louder background noise than other sites located in very similar conditions.

The absence of population density or any population metrics in any of the regression models also seemed odd, since other land-use regression studies have stated population density or population total are important measures [49]. However, the population density of most national parks is most likely going to be low any way and may be why this measure was not important to this study but may be still important for future modeling. In addition, the National Parks are excluded from census calculations, as stated "Protected areas (national parks, wilderness, and GAP status areas) are excluded from population calculations [...]" [30].

Cautions on Heliport Variable

Distance to heliports, or 'DistHeliports' was important to all three regression methods, however, it is suspected that the distance to heliports may be something explicitly tied to national parks. Since many National Parks offer beautiful views, like that of the Grand Canyon and Mount Rushmore, some heliports may exist specifically to enable

tourists air tours of the sites. Further vetting of the model, more random observations outside of the national parks, and more generic geospatial data would be needed before concluding that distance to heliports is truly the best measure to use. This advice generally applies to all variables in all models, but distance to heliports stood out as a metric that may be correlated with national parks. An important note here to future researchers, Sherrill [31] and Benson [5] use Distance to Heliports Only—not airports with heliports—whereas in Nelson 2015 [30], it appears Heliports are potentially at all airports in Hawaii or only heliport locations were available in Hawaii. "Airport point locations were extracted from the National Transportation Atlas Database (2012) for each modeled area. Public use and military airport locations were available for all modeled areas while heliport locations were available only for Hawaii."[30] Further research could help find some finer precision for these geospatial variables, for out whether distance to nearest airports with heliports would be a sufficient metric instead of heliports-only, since the source of the noise would presumably be similar.

Conditional Inference Trees

A conditional inference tree is a single decision tree that splits the data in various ways to build a predictive model. Using R package 'caret' and the 'ctree' model, a predictive model was found to be similar to backwards linear regression predictive performance results for mean absolute error and root mean square error.

Figure 11 shows the hyper-parameters necessary to build this particular Conditional Inference Tree for the NPS sites. It shows the root-mean-square-error is minimized to a value of approximately 5.50 when the p-value threshold is approximately 0.45.

However, at any value between zero and one the root mean square error only varies from 5.6 to 5.8 so optimization doesn't seem to make that big of a difference. The takeaway from the conditional inference model is with one decision tree, one can get a root mean square error of approximately 5.5. This helps put the random forest results in perspective, just as a null model helps put the multiple linear regression performance results in perspective. If one can get a RMSE value of 5.5 with just one tree, is it 'worth' the complexity of one-thousand, or five-hundred, or even ten trees to model each dependent value one wants to predict? This is a question the sponsor or decision maker would need to consider.

Figure 12 shows a visual representation of the decision tree. The variables used are nodes and depicted as ovals, and the data is further split into left and right nodes according to the less-than/greater-than specifications given on each arrow. The box plots at the bottom of the tree represent the range of decibel values, in case of figure 12 this is just for $L_{90}f1$. Another nine conditional inference trees were created but are not shown. The first oval in this decision tree is 'RecCon5km' so the amount of reserved or conserved land in the five kilometer radius is the first discriminating variable, and thus the most important variable, for classifying the variance in the model. This seems to coincide with the latest NPS research which found the first split, and thus the most important variable, was the one that split the data into 'urban' and more 'rural' data [4]—in that study, the variable that did that best was the Visible Infrared data (VIIRS), but that variable was not in the dataset of study. The RecCon5km variable requires all observations with less than 0.32 proportion–or 32% percent of the five kilometers of land–go to the left of the decision tree, and the ones with more than 0.322 go to the right. The second oval in figure 12, labeled with a boxed 2, encountered by all observations with less than 0.322 of 'RecCon5km', is 'DistStrahlerCalgt3', the distance to the nearest stream of Strahler order 3. It's interesting to note that none of

the Distance to Strahler Category Stream variables were in any of the multiple linear stepwise regression models but are important in this method. The conditional tree shows that all observations greater than 26 (kilometers) away from a Strahler Calgt 3 stream go on to a Wetlands5km variable splitting criteria, and then finally one can see the box-plot at the bottom of the Wetlands5km splitting criteria, decibels range from 40-50 on the right and 45-55 on the left. Unlike what NPS researchers had predicted [31], however, it appears from reviewing the boxplots in this conditional inference tree, that the sites located 26 kilometers or closer to an order 3 stream were more quiet on average than sites farther than 26 kilometers away from an order 3 stream—at least for this specific frequency.

Table 18 shows the results of applying the conditional inference tree model for L90f1 to the training data, the test data, and all of the data. Results across all the resampled training folds are in figure 14

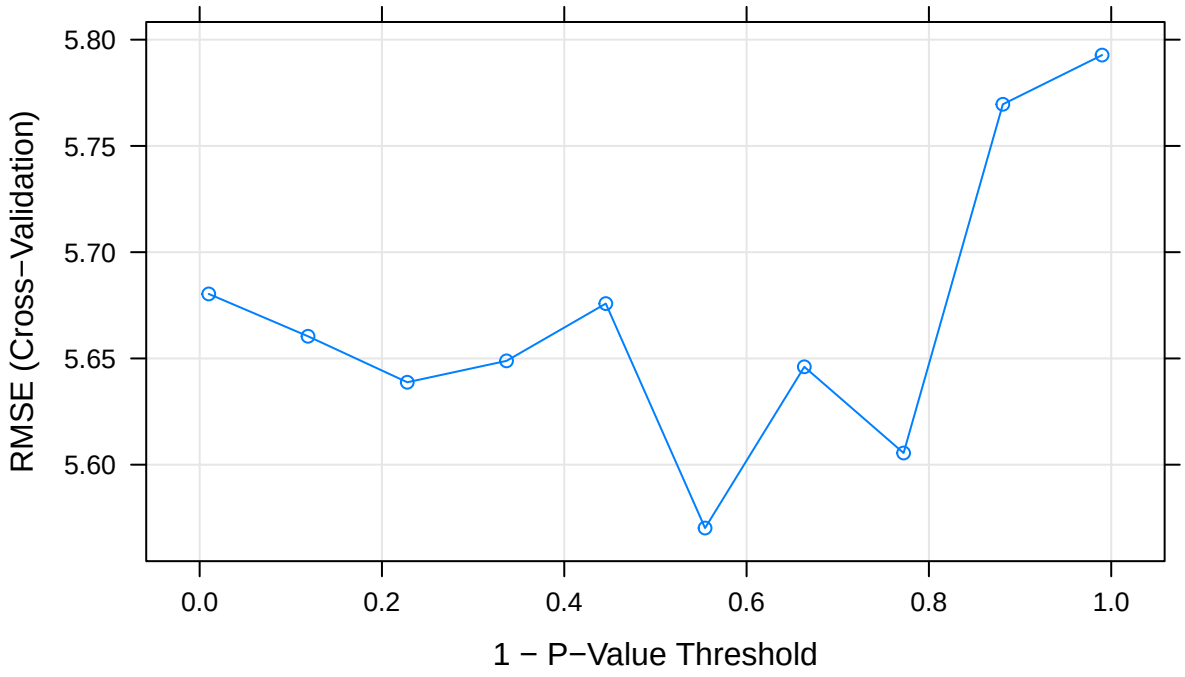

Figure 11. Hyperparameter tuning for finding the best fit one-tree model. The root mean square error varies between 5.5 and 5.8 over ten resampled trials. The optimized value is 5.5 when the p-value is set to approximately 0.45.

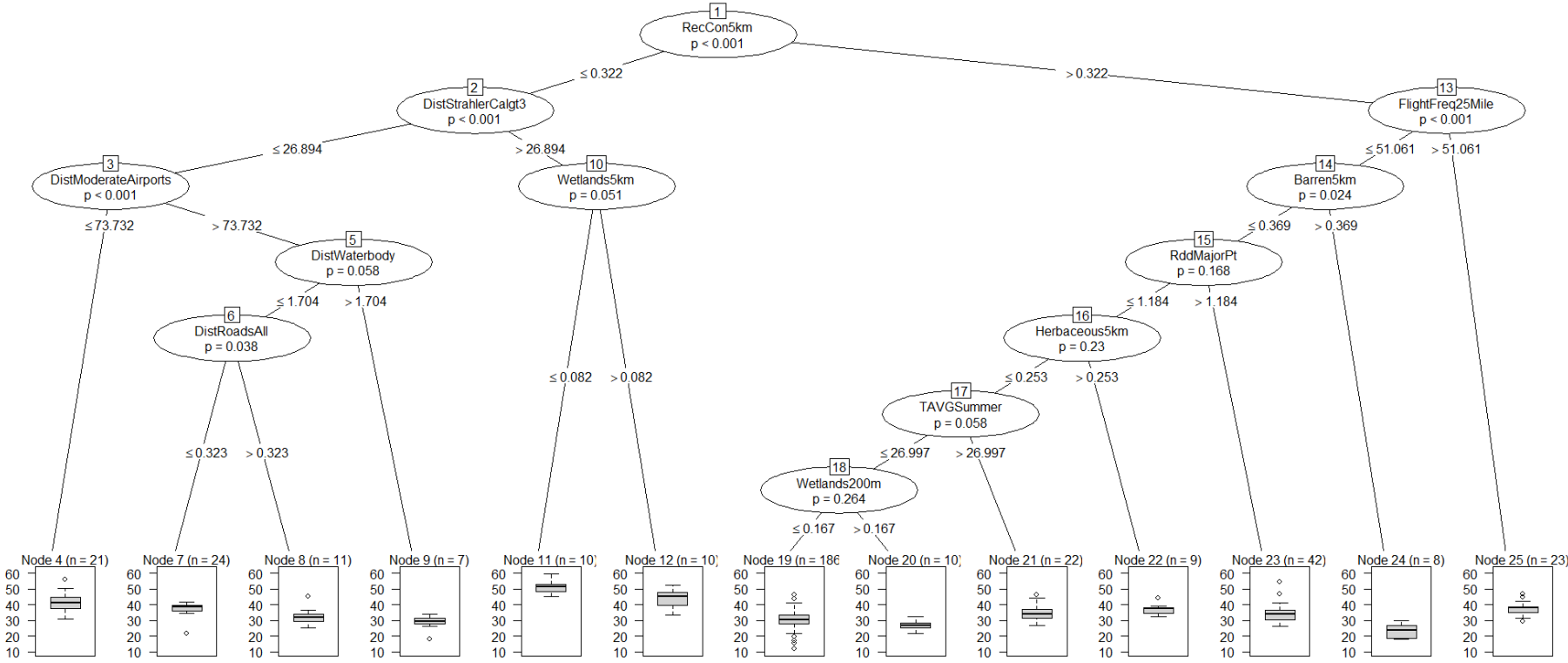

Figure 12. Graphical Representation of the best-tuned conditional inference tree for predicting $L90f1$. Ovals are nodes where information is split according to different given criteria. The bottom are boxplots indicating the range of decibels. When applying this model to predict acoustic values on new data, the mean of the boxplot is applied.

Table 18. Applying Model to Training, Test, and Entire Dataset for L90f1 using Conditional Inference Tree

Metrics	Training	Test	All
RMSE	4.649	5.696	4.927
Rsquared	0.535	0.266	0.465
MAE	3.450	4.232	3.643

The figure 13 shows the results of predictive conditional inference tree for each of the sixteen sites randomly chosen. It shows the worst results for MIMA002 (under predicting by 15 decibels), GLAC007 (under predicting by 10 decibels), GRCA018 (over predicting by 14.6 decibels), MORA001 (over predicting by 11.3 decibels), PIRO002 (under predicting by 18 decibels). MIMA is the Minute Man National Historical Park in Massachusetts, which as the website describes, "At Minute Man National Historical Park the opening battle of the Revolution is brought to life as visitors explore the battlefields and structures associated with April 19, 1775, and witness the American revolutionary spirit through the writings of the Concord authors," [82] so it is assumed to not necessarily enforce 'natural quiet' as much as the other sites since it is known for using loud cannons in its battle re-enactments.

Random Forests

The results of random forest were significantly better than the linear regression models. Table 19 shows the results of applying the best random forest model to the training, test, and entire dataset resulted in an R-Squared of 0.924, a RMSE of 2.17, and a MAE of 1.538.

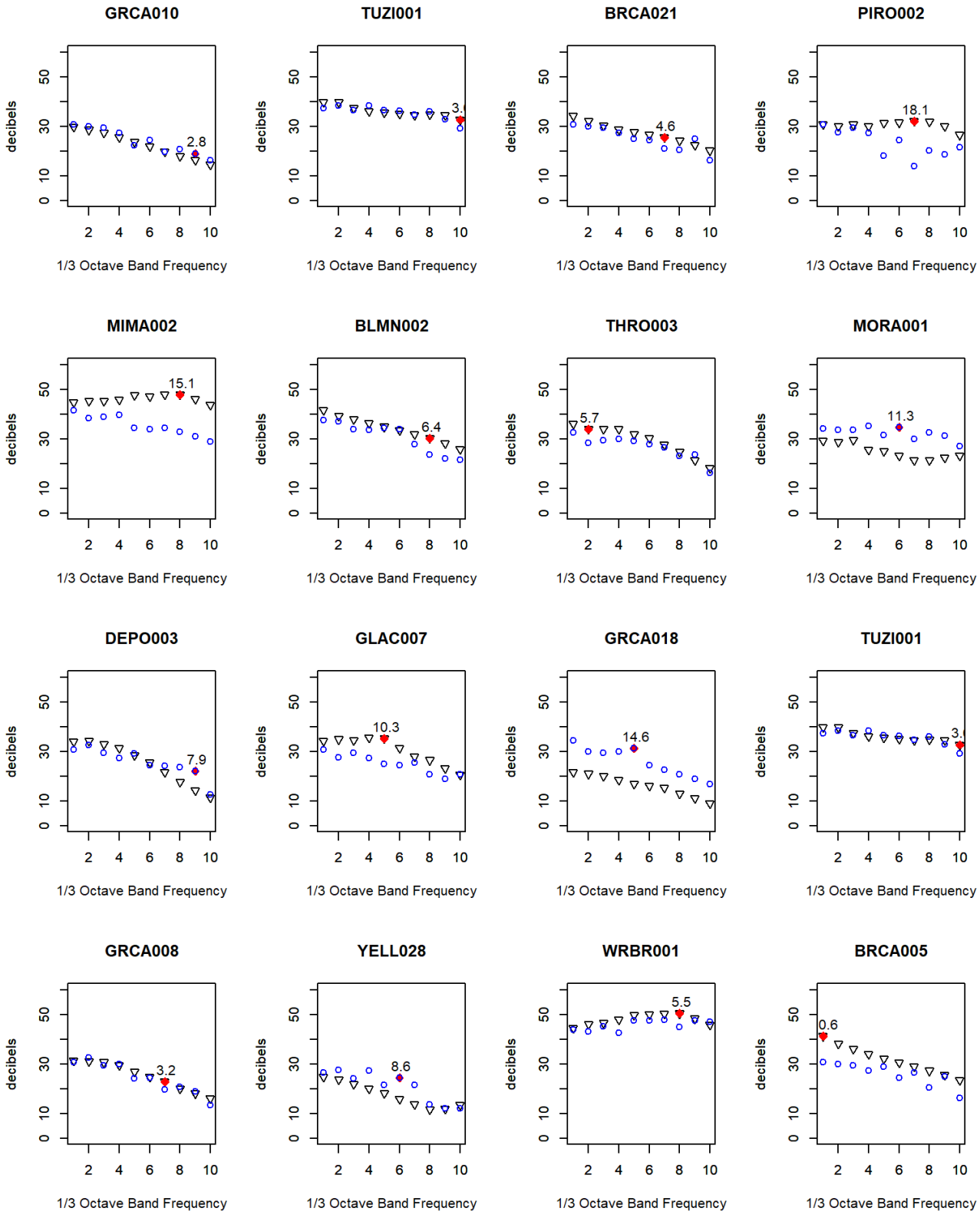

Figure 13. Results for Conditional Inference Tree on Random Sites

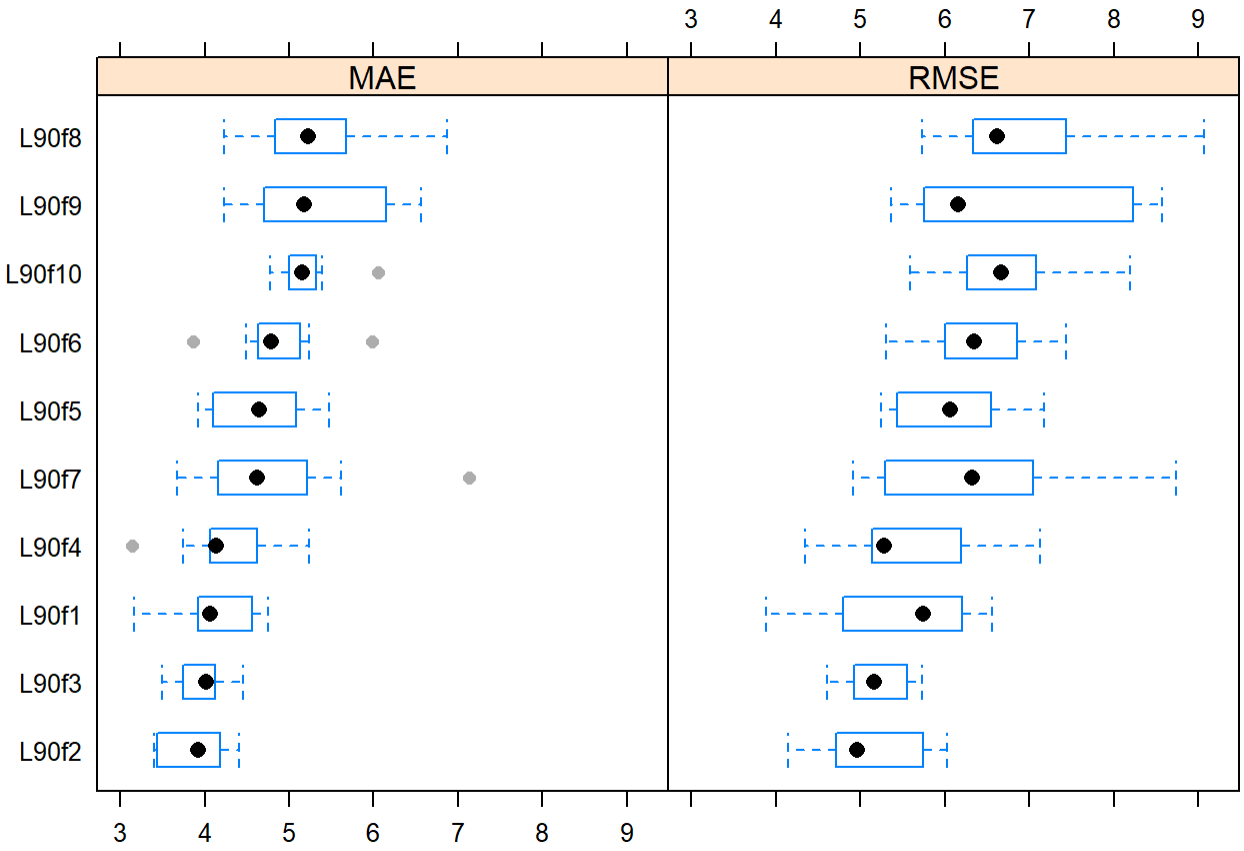

Figure 14. Resample Results for Conditional Inference Tree across all ten frequencies. MAE is Mean Absolute Error. RMSE is Root Mean Squared Error (RMSE is more influenced by outliers than MAE). The x-axis is in decibels. The x-axis is displayed under the MAE and above the RMSE to indicate different axis-scales in these side-by-side plots. Similar to linear regression, predictions are not as accurate for the greater numbered frequencies (eighth, ninth, tenth one-third octave frequency) and are more accurate for the lesser numbered frequencies (second, third, first frequency). The variance increases with the larger frequencies.

Table 19. Applying Model to Training, Test, and Entire Dataset for L90f1 using Random Forests (ranger)

Metrics	Training	Test	All
RMSE	2.176	4.566	2.949
Rsquared	0.924	0.510	0.831
MAE	1.538	3.555	2.034

Table 20. Variables of Importance across first five frequencies with Random Forests (ranger)

f1	f2	f3	f4	f5
RecCon5km	RecCon5km	RecCon5km	RecCon5km	RecCon5km
Elevation	Elevation	Elevation	Elevation	Elevation
WaterOnly5km	WaterOnly5km	FlightFreq25Mile	RddMajorPt	RddMajorPt
DistModerateAirports	FlightFreq25Mile	WaterOnly5km	FlightFreq25Mile	RddMajor5km
Forest200m	DistModerateAirports	RecCon200m	RddMajor5km	FlightFreq25Mile
FlightFreq25Mile	PopTotal_Distributed	RddMajorPt	WaterOnly5km	PopTotal_Distributed
DistHighAirports	DistStrahlerCalgt3	DistStrahlerCalgt3	PopTotal_Distributed	DistRailroads
Shrubland200m	DistCoast	RddMajor5km	DistStrahlerCalgt3	PopDensity_2010_50km
DistStrahlerCalgt3	RddMajorPt	DistModerateAirports	DistRailroads	DistStrahlerCalgt3
DistCoast	Barren5km	PopTotal_Distributed	PopDensity_2010_50km	Developed200m

f6	f7	f8	f9	f10
RecCon5km	RecCon5km	RecCon5km	RecCon5km	RecCon5km
Elevation	Elevation	Elevation	Elevation	RddMajorPt
RddMajorPt	RddMajorPt	RddMajorPt	RddMajorPt	RddMajor5km
RddMajor5km	RddMajor5km	RddMajor5km	RddMajor5km	RecCon200m
PopDensity_2010_50km	FlightFreq25Mile	PopTotal_Distributed	RecCon200m	Elevation
PopTotal_Distributed	PopTotal_Distributed	PopDensity_2010_50km	PopDensity_2010_50km	PopTotal_Distributed
FlightFreq25Mile	PopDensity_2010_50km	FlightFreq25Mile	PopTotal_Distributed	Developed200m
WaterOnly5km	RecCon200m	Developed200m	Developed200m	PopDensity_2010_50km
DistRailroads	DistRailroads	RecCon200m	FlightFreq25Mile	FlightFreq25Mile
DistStrahlerCalgt3	Developed200m	DistRailroads	DistRailroads	Transportation5km

Future predictions using best models

The predictions using linear regression, a single decision tree, and multiple decision trees in random forest, are as expected in order of better estimates. However the better the model is at predicting the national parks acoustics across frequencies is not necessarily the performance on sites outside the United States, or even sites in the United States that are not national parks. The results are only interpretable for the national parks and not readily interpretable for outside points. It is predicted that the better the model is at predicting the noise level of the national parks, the worse it will be at predicting other sites outside of the model because of overfitting. There may be a balance between using a model somewhere between the null model (means only), and the more precise random forests, but this research is not able to initially determine where. Using the data in the Philippines would help at least provide an initial estimate on how good or bad the model can be when applied outsite the national park service.

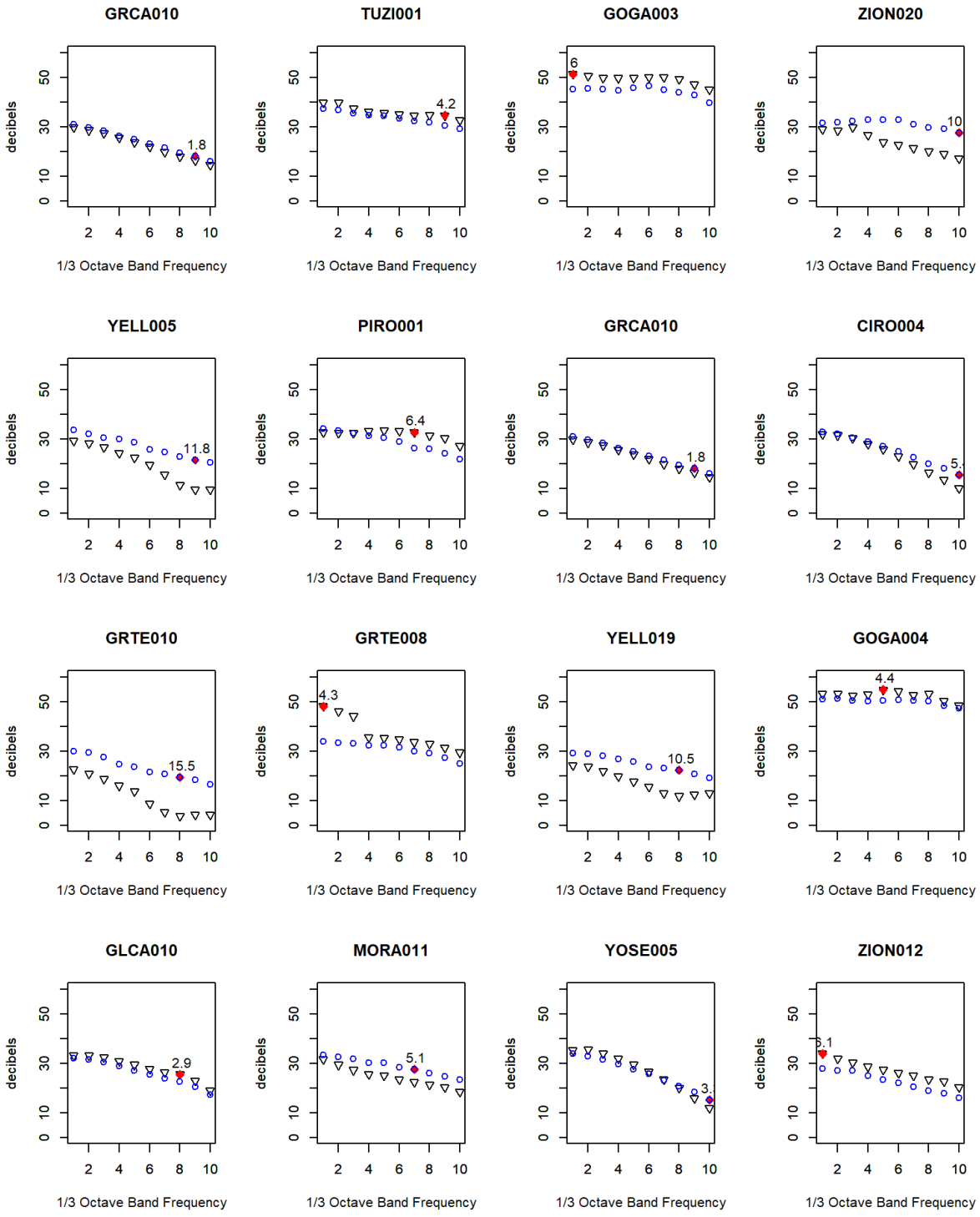

Figure 15. Random Site Predictions using Random Forests. Predictions are blue circles, actuals are black triangles. A red annotation appears at the largest difference in predictions and actual values, and a text label with the difference in decibels appears over the red circle.

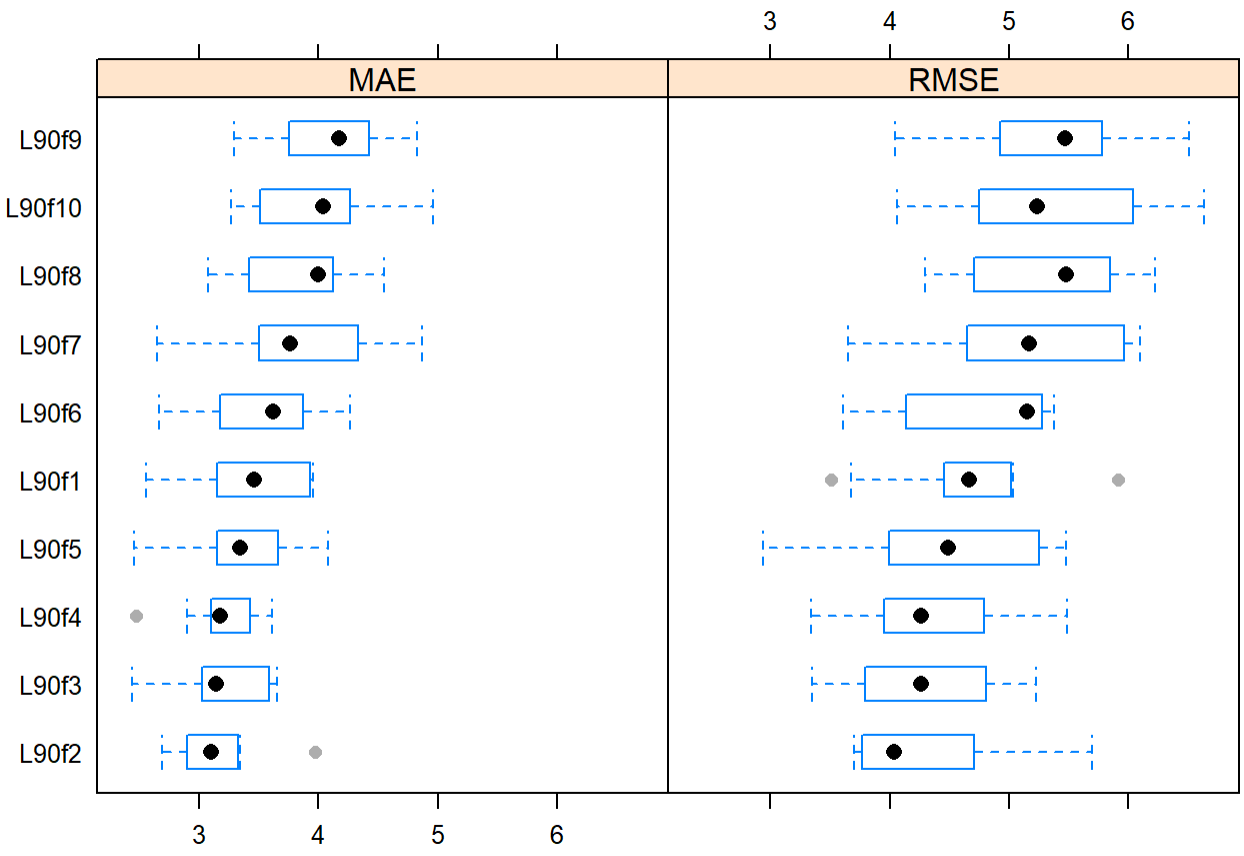

Figure 16. Resample results from Random Forest across the first ten frequencies. Similar to other methods, the least precise estimates are L90f9, the more precise estimates are L90f2

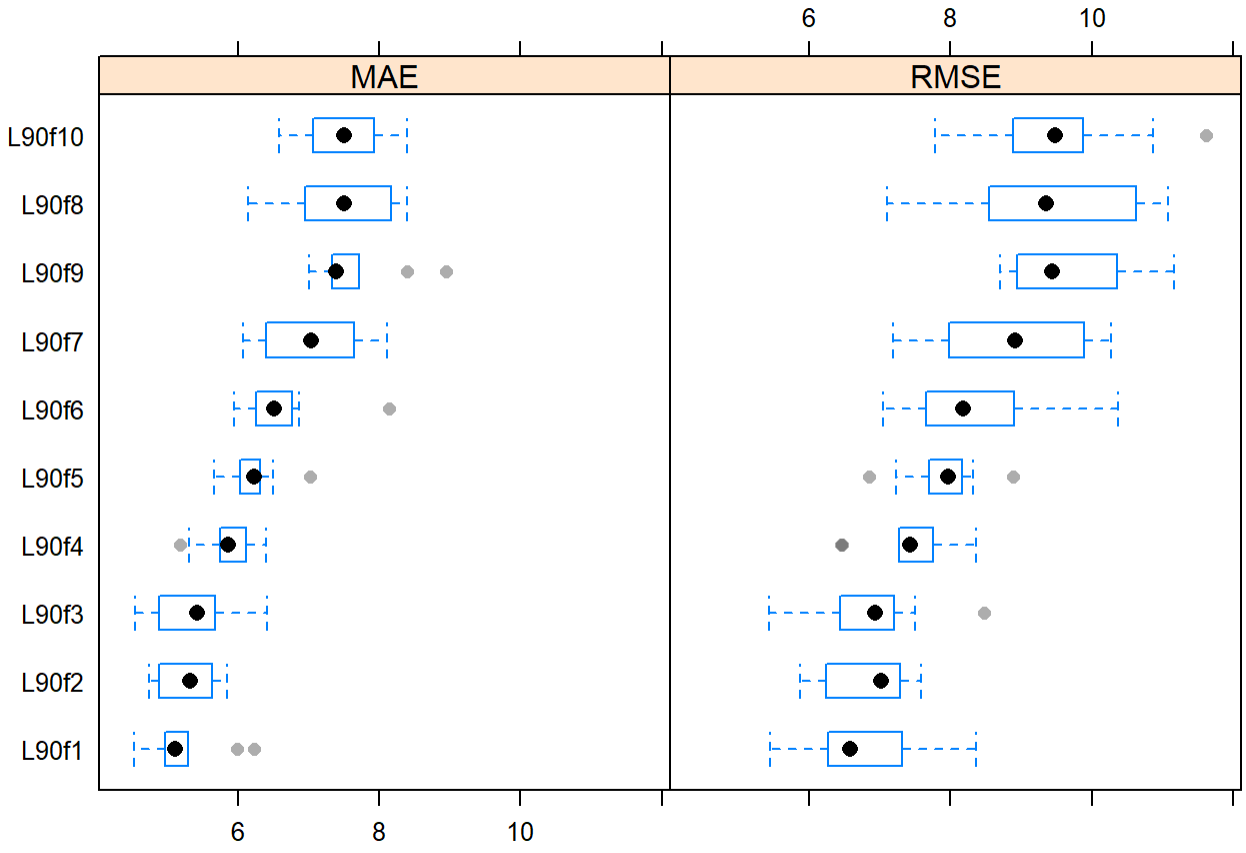

Figure 17. Resample results across the first ten frequencies using the null model, just the mean of the frequency. Similar to other results the best estimates appear to be L90f1, and the worst are L90f10

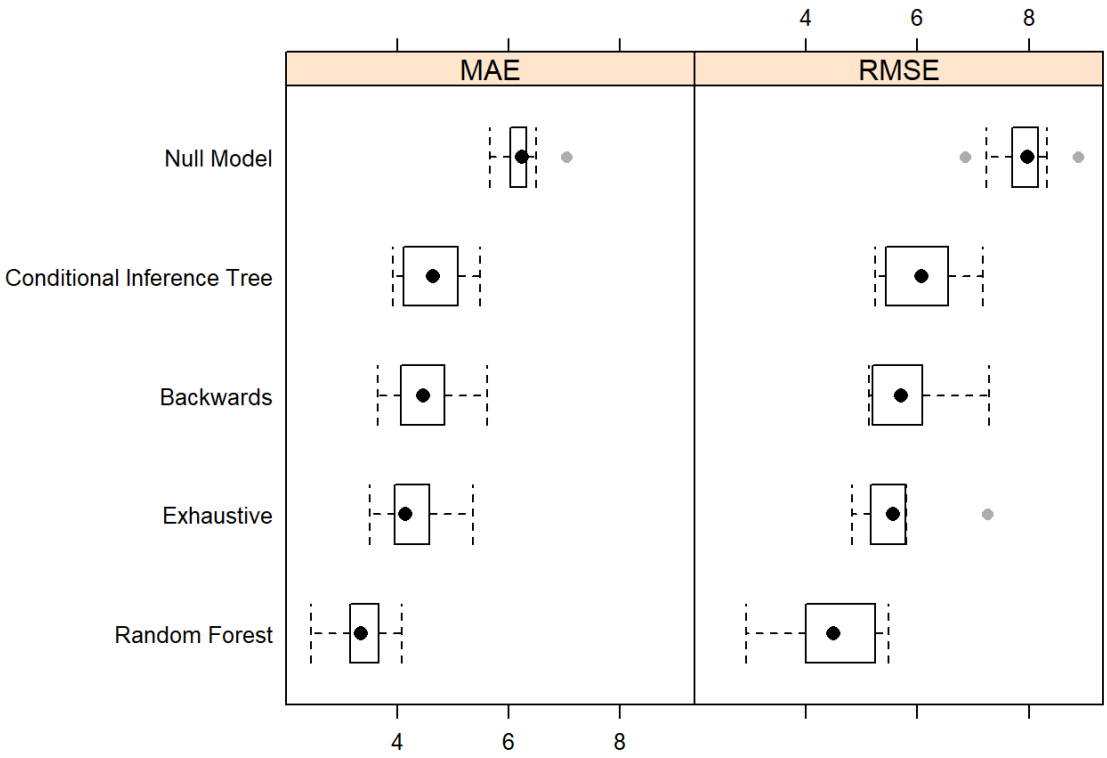

Figure 18. Comparison of the four best models using four statistical techniques for predicting the L90f1 value– the null model is for comparison

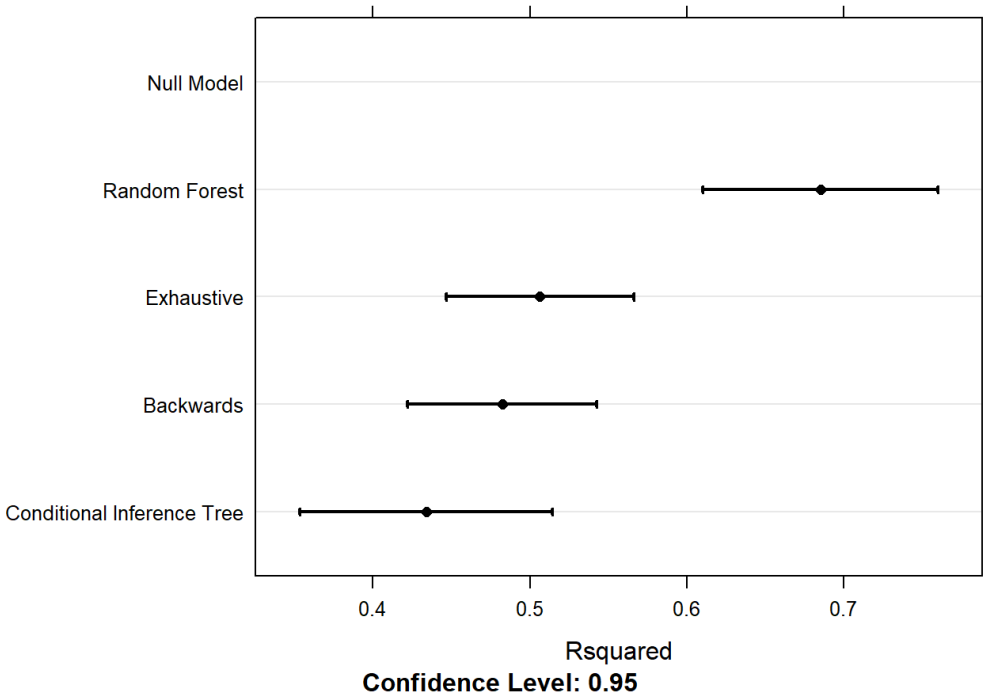

Figure 19. Comparison of the four best models across the first 1/3 octave frequency. The null model will never have an R-squared value by mathematical definition

Sponsor's Hold-Out Points

Unfortunately the hold-out points received in the Philippines did not contain any geospatial data. Attempts were made to work with ArcGIS experts available starting in October 2017, but not enough data was collected in a standardized manner to be of use in any of the developed mathematical models. In January 2018, data collection efforts were refueled with the findings that only eight variables were necessary to predict most of the frequecies. However even specifying this limited number of variables, not enough data was available through open-sources. The data sets and standardized automated processes used by the National Park Service and their collaborators to collect geospatial data were national databases unique only to the United States, not readily available for other countries like the Philippines.

Simply zooming in on the Philippine locations in Google maps allowed qualitative observations which revealed some areas resembled heavily congested traffic locations, four sites were clustered together on a peninsula near a resort with one of those four in the water approximately a hundred meters from shore, one site was further in from the coastline, and one in open water seemingly unconnected to any land sources. These qualitative observations were matched with ArcGIS collected data to help test an interactive parallel coordinates analytic to help inform the sponsor on a best matching record from the National Park Service using ranges of available information since enough point data was not available to compute with model. The parallel coordinate analytic is simple code used from existing sources, about 20 lines of code max. The code is shared in the appendix. It is also partially viewable in figures 20, and 21.

Of note, looking at the Philippines acoustic data also revealed what is most likely a

data-entry error: the L10 values were lower number values than the L90 values so these numbers were most likely switched. To predict the L90 values, we assumed the actual values to compare to were the L10 values.

Data Collection for Philippines

Two people with experience using ArcGIS tools were used to collect the data. They collected approximately 4 out of the 8 needed variables to predict L90f1 over approximately 3 hours each from what they could find in open geospatial databases available and using many assumptions. Person A collected information on type of land category and distances to national roads, distances to airports, and because of a communication error, distance to seaports when distance to airports with seaplanes was actually needed. Person B collected road density, distance to national roads, amount of herbaceous in 200meters, amount of recreational/conserved land in the area, and distance to heliports. Person B was unable to find a geospatial database for the Philippines that contained heliports, so they collected distance to airports. This resulted in two of the same measures which was not intended, yet provided useful information on how far off computations could be.

When Person A and Person B's results for the distance to airports were compared some of them were approximately 1000 meters off. It may be to differing skills/-experiences of ArcGIS analysts, non-standardized open-source geospatial databases available, and the need to improvise when certain databases are not available, or other reasons for the data ranging between different data collectors. Another reason for differences in values would be time. If the measurements were performed from a database updated in 2018, yet the acoustic observations from the Philippines were collected in 2010, any number of natural events could have occurred and changed the

data—for example, a tsunami could have changed the coastline, a sand-bar may be present where the site now appears to be open ocean, or a water body may have dried up and is no longer the closest stream to a location.

In addition, the acoustic data for the Philippines locations did not include any metrics on how many hours were recorded. Based on the literature available, most studies recommend between 10 to 25 days to get a prediction on the ambient background noise of an environment. The provided acoustic summary data may not have spanned this minimum amount of time, and thus may be louder or more quiet than it really would have been had a longer sample been used. This would lead to a similar phenomenon as speculated with the MUWO001 acoustic summary data that varied greatly between 300 hours and 800 hours of data collection in different years and different seasons.

Best Match

Since the geospatial information needed for each of the Philippines locations was imprecise or did not exist, rather than a precise point as was once assumed, a range of possible values for the variables that were found informed a 'best-match' to the observations in the original NPS dataset. The variables from an exhaustive linear regression to predict L90f1 were used in a Parallel Coordinate plot, which is a great visualization to interactively explore highly dimensional data. We found for the one point that appeared to be in an open ocean, using a limited range on WaterOnly200m, and distances to major roads, and small road density, resulted in a best match with GOGA which is the Golden Gate Bridge in San Francisco, CA, as shown in figure 21. Other points that would match the data in the Philippines are in figure 20 with limited ranges on RecCon5km, DistCoast, DistHeliports, DistRoadsMajor, and RddMajorPt. One can see in 20, there are a wide number of possible park matches.

Table 21. The data collected on the Philippines from two ArcGIS experienced persons

SiteID	Latitude	Longitude	HIHerbaceous200m	WaterOnly200m	RecCon5km	DistHeliports	DistHighAirports	DistRoadsMajor	RddMajorPt
1	10.330	124.039	0.662		0.124	-	8,147.462	106	0.0001
2	10.332	124.039	0	1	0.098	-	7,996.768	371	0.0001
3	10.329	124.039	1		0.142	-	8,040.949	525	0.0001
4	10.330	124.039	0.695		0.119	-	8,268.966	3,604	0.0001
5	10.237	123.994	0	1	0.082	-	9,823.605	445	0.00003
6	10.297	123.904	0		0	-	3,678.922	401	0.0005
7	10.339	123.908	1		0	-	3,057.874	641	0.001
8	10.338	123.911	1		0	-	3,282.505	452	0.001

These values were not collected in accordance with previously defined NPS procedures. For example, HIHerbaceous200m appears to be a measurement of *any* vegetation in the area. WaterOnly200m was collected only on the obvious open-ocean locations.

Table 22. Qualitative Data observed on Philippines

SiteID	Latitude	Longitude	LCLU	Observations
1	10.330	124.039	cultivated area mixed with brushland/grassland	Beach Resort
2	10.332	124.039	ocean	near Beach Resort
3	10.329	124.039	cultivated area mixed with brushland/grassland	Beach Resort
4	10.330	124.039	cultivated area mixed with brushland/grassland	Beach Resort
5	10.237	123.994	ocean	middle of Channel
6	10.297	123.904	built up area	street corner
7	10.339	123.908	crop land mixed with coconut plantation	busy street corner
8	10.338	123.911	crop land mixed with coconut plantation	courtyard of nice hotel

Another axis, "L90f1", shows the possible values of L90f1 given all the parameters. One can also visually differentiate the types of landscapes because they are colored by LCLU and the first axis in the parallel coordinate plot. As more data is added into the database, like from Alaska and Hawaii, the analytic may become more useful. It is our hypothesis for future work that the Philippines would probably best match locations available from the National Park Service for Hawaii since they are closer in terms of Longitude and Latitude and share many similar geographical features.

Furthermore, the best way to visualize a parallel coordinate plot is interactively, one is encouraged to use the code provided in appendix c to interact with the data.

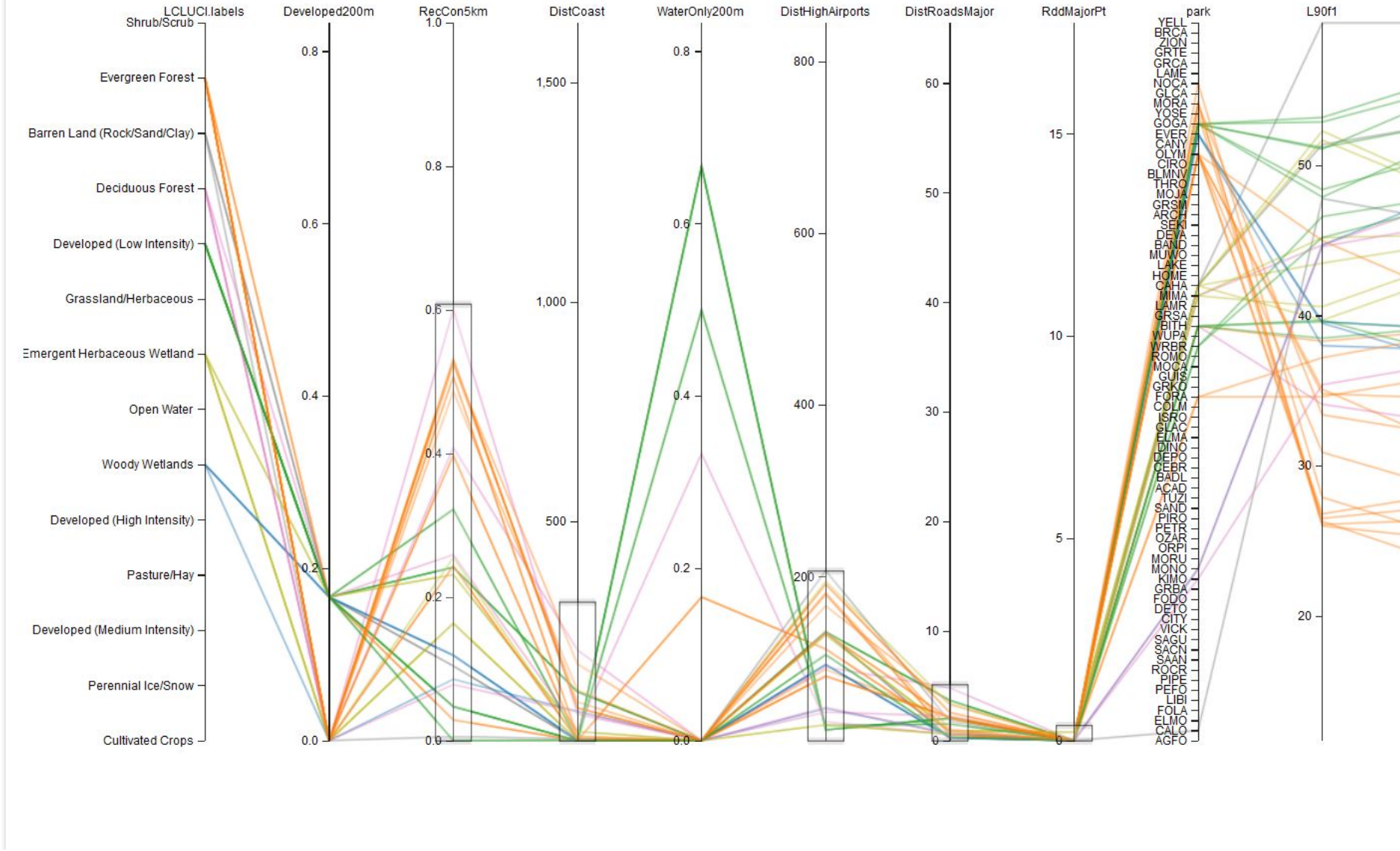

Figure 20. Parallel Coordinates Plot with Variables: LCLUCI by Label, Developed200m, RecCon5km, DistCoast, WaterOnly200m, DistHeliports, DistRoadsMajor, RddMajorPt, and Parks by acronym. All Data is restricted to 0.6 RecCon5km or lower, DistHeliports 20km or lower, DistRoadsMajor 5km or lower, and RddMajorPt as approximately 0.001. There are a lot of potential matches.

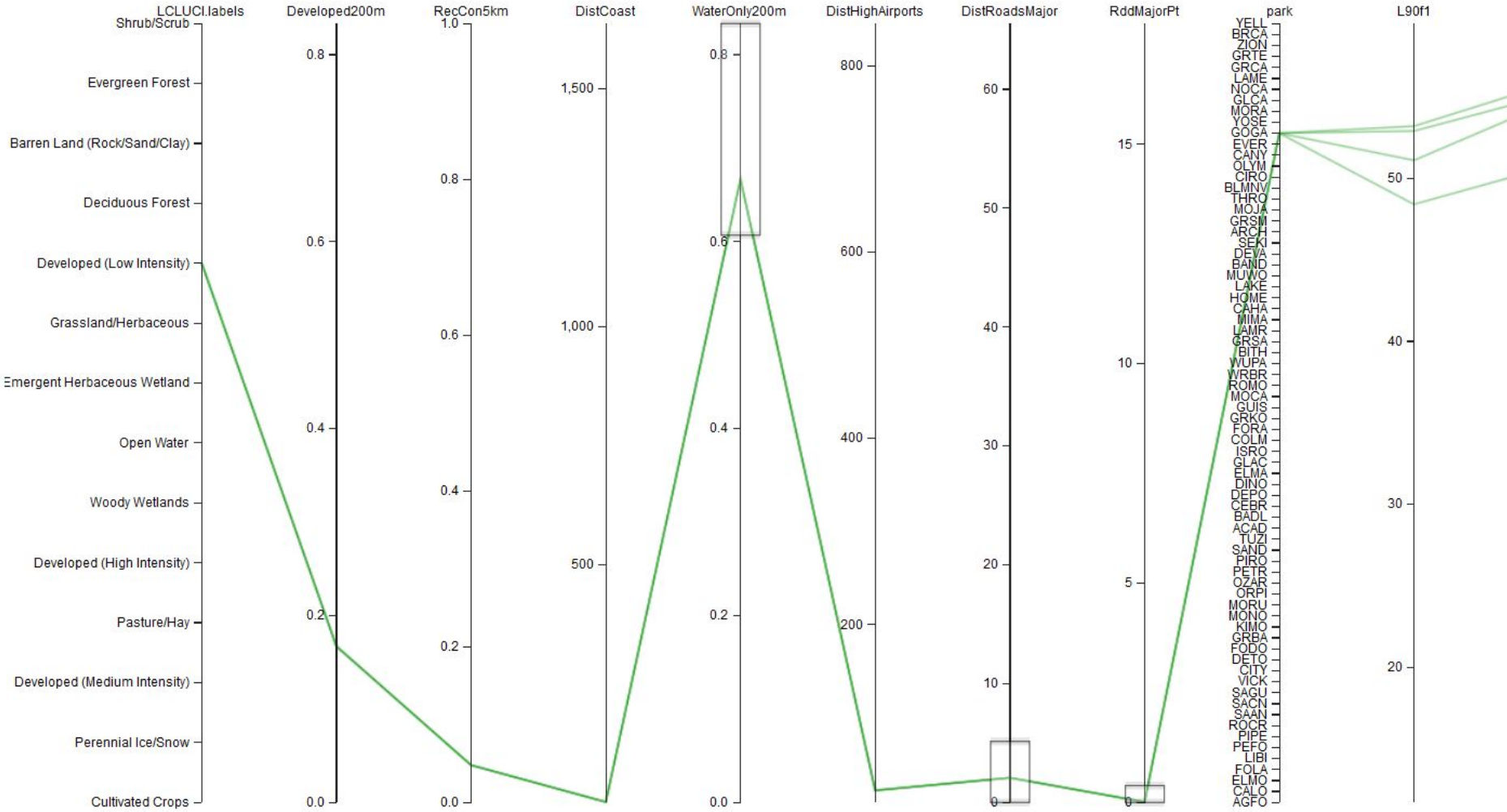

Figure 21. fig:Parallel Coordinates Plot with Variables: LCLUCI by Label, Developed200m, RecCon5km, DistCoast, WaterOnly200m, DistHeliports, DistRoadsMajor, RddMajorPt, and Parks by acronym. All Data is restricted to 0.6 WaterOnly200m or higher, DistRoadsMajor 5km or lower, and RddMajorPt as approximately 0.001. The best match is GOGA —Golden Gates, San Francisco, CA

VI. Conclusion

Summary

This research identified the best subsets of geospatial variables needed for creating predictive models for the first ten of the thirty-three one-third octave frequencies of the National Park acoustic summary data. The analysis used and compared linear regression, conditional inference trees and random forests in 10 fold cross-validation test sets and reported the MAE, MdAE, RMSE, AdjR^2, and PRESS when possible. With backwards-stepwise linear regression a predictive model can be generated within seconds for all the frequencies and through the visual examination of random observation predictions, the model had the unexpected result of capturing the unique acoustic shape of each site using just 15 variables per frequency-model. Data exploration highlighted several limitations with the data source and provided a method for dealing with them. Several assumptions would need to be investigated with the original source of the data–National Park Service Night Skies and Natural Sounds division–before deploying the model for use. For example, how to appropriately fix the negative distances and negative areas in several human-use category geospatial variables. Using backwards and exhaustive regression identified a handful of important variables that potentially can model the first ten one-third octave frequencies. This would be a 90% reduction on the number of variables required from random forests. However random forests is significantly more accurate so should be used when data collection is not an issue.

Future research on speeding up exhaustive regression processes would be very beneficial in combined with more globally recognized geospatial databases and up-to-date validated comprehensive datasets.

Deliverables

This research is also offered to the sponsor in a shortened R Markdown document that can be recompiled to create PDF, HTML or Word Document for easy "what-if" analysis; for example excluding the site ID NOCA008 or MUWO001 and forming a new predictive model. Some variations to the R code or amended code would allow the sponsor to review the other 23 frequencies if they remained of interest, or the L10 and L50 bands. The report used the R package 'caret' to pre-process, train, and validate model performance, and allows a repeatable methodology to continue to add to the analysis if other methods are of interest. The research also created the framework for and examples of a 'best matching' scheme to allow the sponsor to pick one of five hundred different National Park Service sites for its use. The files for this were previously mentioned in the Methodology Chapter. The code is not listed for these files because it is over 40 pages long, but can be requested.

There is also a web-graphical user interface application built using 'Shiny', that will allow the sponsor to explore the original dataset and validate the report findings– for example the wind values of zero. At this moment in time, another tab is being built into the Shiny application to improve upon the best matching algorithm, to use a select number of 'best' predictive variables to intuitively visualize the range of acoustic data possible given a range of geospatial data. This is expected to help alleviate the need for obtaining all the necessary and exact geospatial data for a site of interest when in reality a general range is good enough to narrow down the best matching observation. This will also help future researchers or the sponsor provide a best-match range for the Philippines data using approximations and best guesses on the range of values possible. The code for a parallel coordinate plot that can easily be implemented is included in the appendix.

Future Research/Work

A couple recommendations are offered as suggested future work and also future skills/training needed:

Enlist ArcGIS expertise

At least 40 hours of tailored training on ArcGIS software tools is expected to enable a future researcher the ability to extract multiple databases of information for multiple sites of interest. This is needed to gather data from each variable indicated in at least one of the three models (linear regression, conditional inference tree, or random forests) on the sponsor's hold-out point of interest from the Philippines. This is needed in conjunction with an understanding of what databases to use as a best substitute. Extensive documentation on how the geospatial data were obtained is available from [31] and [30] but most databases used are national databases not readily available for the Philippines. The methodology of importing the data would need to replicate the NPS methodology to ensure the same scale and same time frame. There were variables of great importance according to previous NPS published work like the Visible Infrared Imaging Radiometer Suite (VIIRS) data [3] , specifically, the nighttime lights information, that was not in this research's data set. An original datasource was found that had 996 observations that also included hundreds of airport acoustic summaries. However that data did not have frequency specific data so it was not usable within the scope of this research. A lot of the original information appears to be available on the open government data website repository, www.data.gov, but the datasets available for each site are very large (the represent hundreds of hours of audio recordings per site) and require at least a basic knowledge

of python programming once received to extract and summarize.

Narrow down locations of interest

A better understanding of the operational locations of interest for using a predictive acoustic model are desirable to ensure enough relevant data points are in the model. The data did not contain many urban sites with large L_{90} values so it is anticipated to under estimate loud urban sites. A meta-analysis of traffic-noise studies would be advisable to gain insights on what variables are important for city environments– for example distance to nearest bus stop is not an available geospatial variable in this research database, but would be applicable to improving the predictions in this research and city-studies [2], [16].

Future methodologies

Future methodologies that may be of interest are using each frequency model as a predictive model for other frequencies, or using the probability density function of each frequency as a independent variables or "dependent functions". One methodology that seemed relevant but beyond the scope of this research was canonical correlation analysis [83], which would find an overall model that could account for a majority of the variance in all ten frequencies, and then a frequency specific model to specify the influence of variables on each frequency.

Principal component regression may also provide insights on whether a small number of individual components can provide enough information on the variance to be preferred over any of the three methods studied here, but preliminary experimentation revealed only two components would explain about 25% of the variance, and still use

up to eighty components for creating a model. Since a component contains all the variables in the dataset, this would result in a very large model per frequency and did not seem as beneficial to use as the other methods presented in this research.

Another methodology of potential benefit is using the exhaustive linear results to feed into a random forest and vice versa to reduce the number of variables in the model. If linear regression identified ten variables, what would the random forest look like if built on just those variables? This would allow interactions to be evaluated.

Appendix A. Original Variables

Table 23. National Park Service Night Skies Natural Sounds Division Variables

Variable	Description	Unit
siteID	Unique Identifier	ABC####
Season	Fall, Spring, Winter, or Summer	
region	All data was CONUS	CONUS
park	Location of site	
type	Park or City	
firstYear	Year recording began	YYYY
nYears	Years spanning recording	Integer
nHours	Recorded hours	Integer
Latitude		Degrees
Longitude		Degrees
Elevation	Distance above sea level	Feet
Slope	Rate of change of elevation	Degrees
Land Cover Variables	Proportion of Area of Analysis (AOA)	Resolution
Barren 200m	barren land	0.2 km AOA
Cultivated 200m	cultivated	0.2 km AOA
Developed 200m	developed land	0.2 km AOA
Forest 200m	forest land cover	0.2 km AOA
Deciduous 200m	deciduous forest land cover	0.2 km AOA
Evergreen 200m	evergreen forest land cover	0.2 km AOA
Mixed 200m	mixed forest land cover	0.2 km AOA
Herbaceous 200m	herbaceous land cover	0.2 km AOA
Shrubland 200m	shrubland land cover	0.2 km AOA
Snow 200m	snow land cover	0.2 km AOA
Wetlands 200m	wetlands land	0.2 km AOA
WaterOnly 200m	water land cover	0.2 km AOA
Land Use Variables	Proportion of Area of Analysis (AOA)	Resolution
WaterNat 200m	landuse natural water	0.2 km AOA
WaterHum 200m	landuse human modified water	0.2 km AOA
Wet 200m	landuse wetlands	0.2 km AOA.
RecCon 200m	landuse recreation and conservation	0.2 km AOA
Timber 200m	extractive land use timber harvesting	0.2 km AOA
Grazing 200m	extractive land use livestock grazing	0.2 km AOA
Cropland 200m	extractive land use cropland	0.2 km AOA
Suburban 200m	built land use residential suburban	0.2 km AOA
Commercial 200m	built land commercial	0.2 km AOA
Industrial 200m	built land use industrial	0.2 km AOA
Institutional 200m	land use institutional	0.2 km AOA
Transportation 200m	land transportation	0.2 km AOA
Extractive 200m	extractive land use class	0.2 km AOA
Built 200m	built land use class	0.2 km AOA
Other Environmental Factors		
DistAirportsAllMotorized	Distance to motorized airports	Meters
DistAirportsSeaplane	Distance to all airports and seaplane bases only	Meters
DistCoast	Distance to National Hydrology Dataset (10 mile AOA)	Meters
DistHeliports	Distance to heliports only	Meters
DistHighAirports	Distance to airport with over 1M enplanements	Meters
DistLowAirports	Distance to airports with greater than 5K enplanements	Meters
DistMilitary	Distance to nearest military flight path (25 mile AOA)	Meters

Continued on next page

DistModerateAirports	Distance to all airports with greater than 250K enplanements	Meters
DistRailroads	Euclidean distance to National Atlas 2012 GIS data	Meters
DistRoadsAll	Distance to nearest road (all roads) m	Meters
DistRoadsMajor	Point Distance to nearest road (major roads)	Meters
DistStrahlerCalgt1	Distance to NHD Plus flowline with a SC stream order greater than 1 .	Meters
DistStrahlerCalgt3	Distance to NHD Plus flowline with a SC stream order greater than 3.	Meters
DistStrahlerCalgt4	Distance to NHD Plus flowline with a SC stream order greater than 4.	Meters
DistStreamsAny	Distance to closest stream	Meters
DistWaterbody	Distance to waterbody (10 mile AOA)	Meters
FlightFreq25Mile	Sum of weekly flight observations (25 mile AOA)	Meters
MilitarySum_25miles	Sum of designated military flight paths (25 mile AOA)	Meters
RddAll5km*	Sum of road density all roads (5 mile AOA)	Meters
RddAllPt	Sum of road density all roads (5 km AOA)	Meters
RddMajor5km	Sum of road density all roads(5 km AOA)	Meters
RddMajorPt	Sum of road density major roads (5 km AOA)	Meters
RddWeighted5km	Sum of road density weighted roads (5 km AOA)	Meters
RddWeightedPt	Sum of road density weighted roads (5 km AOA)	Meters
PPTNorms	Average yearly precipitation	Millimeters * 100
PPTSummer	Average summer precipitation	(Point - millimeters times 100)
TAVGNorms*	Average yearly temperature	C
TAVGSummer	Average summer yearly temperature	C
TDEWAvgSummer*	Average summer dew point temperature	C
TDEWNorms*	Average yearly dew point temperature	C
TPI	Ordinal bin of TPIRaw	1-6
TPIRaw	Topographic Position raw value.	Continuous
Wilderness	Sum of designated wilderness	$Meters^2$
Wind_CRU	Wind power class potential density (50m AOA)	$\frac{W}{Meters^2}$

Appendix B. Equations

Equations for Backward-Stepwise Regression:

$$\begin{aligned} L90f1 = {} & (30.188 \pm 3.565) \times \text{(Intercept)} + \\ & (-10.66 \pm 1.829) \times \text{Forest200m} + \\ & (-8.063 \pm 1.783) \times \text{Shrubland200m} + \\ & (-17.094 \pm 4.325) \times \text{WaterOnly200m} + \\ & (-11.783 \pm 2.793) \times \text{Wetlands200m} + \\ & (-17.182 \pm 3.622) \times \text{Barren5km} + \\ & (12.301 \pm 2.819) \times \text{WaterOnly5km} + \\ & (-0.004 \pm 0.001) \times \text{DistCoast} + \\ & (0.062 \pm 0.018) \times \text{FlightFreq25Mile} + \\ & (2.549 \pm 0.831) \times \text{Wind\_CRU} \end{aligned} \tag{1}$$

$$\begin{aligned} L90f2 = {} & (36.065 \pm 1.35) \times \text{(Intercept)} + \\ & (-9.75 \pm 1.748) \times \text{Forest200m} + \\ & (-7.838 \pm 1.777) \times \text{Shrubland200m} + \\ & (8.247 \pm 3.001) \times \text{WaterNat200m} + \\ & (-23.149 \pm 4.342) \times \text{WaterOnly200m} + \\ & (-9.751 \pm 2.648) \times \text{Wetlands200m} + \\ & (-15.844 \pm 3.542) \times \text{Barren5km} + \\ & (97.761 \pm 35.428) \times \text{Transportation5km} + \\ & (15.547 \pm 2.503) \times \text{WaterOnly5km} + \\ & (0.091 \pm 0.017) \times \text{FlightFreq25Mile} \end{aligned} \tag{2}$$

$$
\begin{aligned}
L90f3 = (35.348 \pm 1.379) \times \text{(Intercept)} + \\
(-9.795 \pm 1.787) \times \text{Forest200m} + \\
(-8.633 \pm 1.816) \times \text{Shrubland200m} + \\
(9.13 \pm 3.068) \times \text{WaterNat200m} + \\
(-23.658 \pm 4.438) \times \text{WaterOnly200m} + \\
(-10.152 \pm 2.706) \times \text{Wetlands200m} + \\
(-15.115 \pm 3.621) \times \text{Barren5km} + \\
(103.938 \pm 36.213) \times \text{Transportation5km} + \\
(14.648 \pm 2.558) \times \text{WaterOnly5km} + \\
(0.101 \pm 0.017) \times \text{FlightFreq25Mile}
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
L90f4 = (44.106 \pm 2.019) \times \text{(Intercept)} + \\
(-0.109 \pm 0.043) \times \text{Slope} + \\
(-12.428 \pm 3.36) \times \text{Barren200m} + \\
(-12.868 \pm 2.221) \times \text{Forest200m} + \\
(-13.271 \pm 2.156) \times \text{Shrubland200m} + \\
(10.094 \pm 3.233) \times \text{WaterNat200m} + \\
(-30.749 \pm 5.044) \times \text{WaterOnly200m} + \\
(-14.595 \pm 2.958) \times \text{Wetlands200m} + \\
(-3.309 \pm 1.341) \times \text{RecCon5km} + \\
(15.114 \pm 2.813) \times \text{WaterOnly5km} + \\
(-0.03 \pm 0.01) \times \text{DistHeliports} + \\
(0.111 \pm 0.018) \times \text{FlightFreq25Mile} + \\
(-0.789 \pm 0.287) \times \text{TPI}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
L90f5 = (43.659 \pm 2.132) \times (\text{Intercept}) + \\
(-0.115 \pm 0.046) \times \text{Slope} + \\
(-16.603 \pm 3.436) \times \text{Barren200m} + \\
(-16.434 \pm 2.231) \times \text{Forest200m} + \\
(-17.727 \pm 2.156) \times \text{Shrubland200m} + \\
(13.263 \pm 3.335) \times \text{WaterNat200m} + \\
(-36.378 \pm 5.175) \times \text{WaterOnly200m} + \\
(-16.662 \pm 3.151) \times \text{Wetlands200m} + \\
(17.981 \pm 2.819) \times \text{WaterOnly5km} + \\
(-0.035 \pm 0.01) \times \text{DistHeliports} + \\
(0.131 \pm 0.019) \times \text{FlightFreq25Mile} + \\
(-0.895 \pm 0.301) \times \text{TPI}
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
L90f6 = (40.635 \pm 1.911) \times (\text{Intercept}) + \\
(-15.886 \pm 3.57) \times \text{Barren200m} + \\
(-15.461 \pm 2.442) \times \text{Forest200m} + \\
(-16.978 \pm 2.408) \times \text{Shrubland200m} + \\
(11.925 \pm 3.62) \times \text{WaterNat200m} + \\
(-35.141 \pm 5.58) \times \text{WaterOnly200m} + \\
(-17.132 \pm 3.317) \times \text{Wetlands200m} + \\
(-4.29 \pm 1.485) \times \text{RecCon5km} + \\
(16.227 \pm 3.107) \times \text{WaterOnly5km} + \\
(-0.035 \pm 0.011) \times \text{DistHeliports} + \\
(0.128 \pm 0.02) \times \text{FlightFreq25Mile}
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
L90f7 = {} & (40.443 \pm 2.01) \times \text{(Intercept)} + \\
& (-17.041 \pm 3.755) \times \text{Barren200m} + \\
& (-16.632 \pm 2.569) \times \text{Forest200m} + \\
& (-18.474 \pm 2.533) \times \text{Shrubland200m} + \\
& (13.252 \pm 3.807) \times \text{WaterNat200m} + \\
& (-38.237 \pm 5.87) \times \text{WaterOnly200m} + \\
& (-18.433 \pm 3.489) \times \text{Wetlands200m} + \\
& (-4.911 \pm 1.562) \times \text{RecCon5km} + \\
& (16.635 \pm 3.268) \times \text{WaterOnly5km} + \\
& (-0.038 \pm 0.012) \times \text{DistHeliports} + \\
& (0.142 \pm 0.021) \times \text{FlightFreq25Mile}
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
L90f8 = (32.325 \pm 2.333) \times \text{(Intercept)} + \\
(3.033 \pm 0.948) \times \text{SeasonSummer} + \\
(4.097 \pm 1.388) \times \text{SeasonFall} + \\
(-12.76 \pm 4.052) \times \text{Barren200m} + \\
(-13.543 \pm 2.852) \times \text{Forest200m} + \\
(-15.046 \pm 2.882) \times \text{Shrubland200m} + \\
(13.283 \pm 3.983) \times \text{WaterNat200m} + \\
(-31.909 \pm 6.253) \times \text{WaterOnly200m} + \\
(-14.917 \pm 3.841) \times \text{Wetlands200m} + \\
(-5.427 \pm 1.653) \times \text{RecCon5km} + \\
(122.775 \pm 47.714) \times \text{Transportation5km} + \\
(15.983 \pm 3.423) \times \text{WaterOnly5km} + \\
(0.145 \pm 0.022) \times \text{FlightFreq25Mile} + \\
(0.637 \pm 0.254) \times \text{RddMajorPt}
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
L90f9 = {} & (32.816 \pm 2.524) \times \text{(Intercept)} + \\
& (2.869 \pm 0.961) \times \text{SeasonSummer} + \\
& (3.901 \pm 1.397) \times \text{SeasonFall} + \\
& (-13.029 \pm 4.133) \times \text{Barren200m} + \\
& (-14.198 \pm 2.905) \times \text{Forest200m} + \\
& (-15.849 \pm 2.929) \times \text{Shrubland200m} + \\
& (13.732 \pm 4.013) \times \text{WaterNat200m} + \\
& (-33.191 \pm 6.322) \times \text{WaterOnly200m} + \\
& (-15.094 \pm 3.88) \times \text{Wetlands200m} + \\
& (-4.71 \pm 1.674) \times \text{RecCon5km} + \\
& (119.484 \pm 48.434) \times \text{Transportation5km} + \\
& (16.063 \pm 3.447) \times \text{WaterOnly5km} + \\
& (-0.037 \pm 0.013) \times \text{DistHeliports} + \\
& (0.123 \pm 0.023) \times \text{FlightFreq25Mile} + \\
& (0.65 \pm 0.258) \times \text{RddMajorPt}
\end{aligned} \tag{9}
$$

$$
\begin{aligned}
L90f10 = (26.41 \pm 2.134) \times \text{(Intercept)} + \\
(2.966 \pm 0.939) \times \text{SeasonSummer} + \\
(3.903 \pm 1.377) \times \text{SeasonFall} + \\
(-6.963 \pm 2.19) \times \text{Forest200m} + \\
(-9.678 \pm 2.279) \times \text{Shrubland200m} + \\
(10.678 \pm 3.93) \times \text{WaterNat200m} + \\
(-22.903 \pm 5.616) \times \text{WaterOnly200m} + \\
(-9.085 \pm 3.34) \times \text{Wetlands200m} + \\
(-5.263 \pm 1.621) \times \text{RecCon5km} + \\
(150.724 \pm 46.993) \times \text{Transportation5km} + \\
(16.323 \pm 3.386) \times \text{WaterOnly5km} + \\
(-0.04 \pm 0.016) \times \text{DistRailroads} + \\
(0.099 \pm 0.023) \times \text{FlightFreq25Mile} + \\
(0.95 \pm 0.246) \times \text{RddMajorPt}
\end{aligned}
\tag{10}
$$

Equations for Exhaustive Regression:

$$\begin{aligned} L90f1 = {} & (20.843 \pm 3.088) \times \text{(Intercept)} + \\ & (10.572 \pm 3.024) \times \text{Barren200m} + \\ & (12.151 \pm 2.605) \times \text{HIHerbaceous200m} + \\ & (-16.862 \pm 4.278) \times \text{Barren5km} + \\ & (10.287 \pm 2.191) \times \text{WaterOnly5km} + \\ & (-0.004 \pm 0.001) \times \text{DistCoast} + \\ & (0.07 \pm 0.018) \times \text{FlightFreq25Mile} + \\ & (2.952 \pm 0.812) \times \text{Wind\_CRU} \end{aligned} \tag{11}$$

$$\begin{aligned} L90f2 = {} & (21.307 \pm 2.987) \times \text{(Intercept)} + \\ & (9.185 \pm 2.92) \times \text{Barren200m} + \\ & (11.687 \pm 2.55) \times \text{HIHerbaceous200m} + \\ & (-14.94 \pm 4.137) \times \text{Barren5km} + \\ & (102.609 \pm 34.943) \times \text{Transportation5km} + \\ & (11.02 \pm 2.123) \times \text{WaterOnly5km} + \\ & (-0.004 \pm 0.001) \times \text{DistCoast} + \\ & (0.069 \pm 0.018) \times \text{FlightFreq25Mile} + \\ & (2.495 \pm 0.786) \times \text{Wind\_CRU} \end{aligned} \tag{12}$$

$$
\begin{aligned}
L90f3 = (30.794 \pm 1.274) \times \text{(Intercept)} + \\
(12.205 \pm 2.641) \times \text{HIHerbaceous200m} + \\
(9.1 \pm 3.099) \times \text{WaterNat200m} + \\
(-15.713 \pm 4.318) \times \text{WaterOnly200m} + \\
(-3.01 \pm 1.267) \times \text{RecCon5km} + \\
(16.061 \pm 2.703) \times \text{WaterOnly5km} + \\
(-0.027 \pm 0.01) \times \text{DistHeliports} + \\
(0.094 \pm 0.017) \times \text{FlightFreq25Mile} + \\
(0.5 \pm 0.187) \times \text{RddMajorPt}
\end{aligned} \tag{13}
$$

$$
\begin{aligned}
L90f4 = (29.851 \pm 1.323) \times \text{(Intercept)} + \\
(12.347 \pm 3.338) \times \text{Developed200m} + \\
(13.225 \pm 2.741) \times \text{HIHerbaceous200m} + \\
(9.829 \pm 3.246) \times \text{WaterNat200m} + \\
(-16.439 \pm 4.486) \times \text{WaterOnly200m} + \\
(-3.733 \pm 1.315) \times \text{RecCon5km} + \\
(15.18 \pm 2.875) \times \text{WaterOnly5km} + \\
(-0.031 \pm 0.01) \times \text{DistHeliports} + \\
(0.106 \pm 0.018) \times \text{FlightFreq25Mile}
\end{aligned} \tag{14}
$$

$$
\begin{aligned}
L90f5 = (28.826 \pm 1.39) \times \text{(Intercept)} + \\
(14.354 \pm 3.507) \times \text{Developed200m} + \\
(15.126 \pm 2.88) \times \text{HIHerbaceous200m} + \\
(10.377 \pm 3.411) \times \text{WaterNat200m} + \\
(-16.45 \pm 4.714) \times \text{WaterOnly200m} + \\
(-4.21 \pm 1.382) \times \text{RecCon5km} + \\
(15.652 \pm 3.022) \times \text{WaterOnly5km} + \\
(-0.032 \pm 0.011) \times \text{DistHeliports} + \\
(0.111 \pm 0.019) \times \text{FlightFreq25Mile}
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
L90f6 = (27.099 \pm 1.462) \times \text{(Intercept)} + \\
(16.897 \pm 3.69) \times \text{Developed200m} + \\
(15.648 \pm 3.03) \times \text{HIHerbaceous200m} + \\
(11.435 \pm 3.588) \times \text{WaterNat200m} + \\
(-18.727 \pm 4.959) \times \text{WaterOnly200m} + \\
(-4.028 \pm 1.454) \times \text{RecCon5km} + \\
(16.381 \pm 3.179) \times \text{WaterOnly5km} + \\
(-0.036 \pm 0.011) \times \text{DistHeliports} + \\
(0.124 \pm 0.02) \times \text{FlightFreq25Mile}
\end{aligned} \tag{16}
$$

$$
\begin{aligned}
L90f7 = (25.872 \pm 1.54) \times \text{(Intercept)} + \\
(19.345 \pm 3.886) \times \text{Developed200m} + \\
(16.109 \pm 3.191) \times \text{HIHerbaceous200m} + \\
(12.743 \pm 3.779) \times \text{WaterNat200m} + \\
(-20.474 \pm 5.222) \times \text{WaterOnly200m} + \\
(-4.693 \pm 1.531) \times \text{RecCon5km} + \\
(16.495 \pm 3.348) \times \text{WaterOnly5km} + \\
(-0.038 \pm 0.012) \times \text{DistHeliports} + \\
(0.136 \pm 0.021) \times \text{FlightFreq25Mile}
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
L90f8 = (24.986 \pm 1.643) \times \text{(Intercept)} + \\
(21.824 \pm 4.146) \times \text{Developed200m} + \\
(16.34 \pm 3.405) \times \text{HIHerbaceous200m} + \\
(12.24 \pm 4.032) \times \text{WaterNat200m} + \\
(-20.074 \pm 5.572) \times \text{WaterOnly200m} + \\
(-5.269 \pm 1.633) \times \text{RecCon5km} + \\
(16.67 \pm 3.572) \times \text{WaterOnly5km} + \\
(-0.043 \pm 0.013) \times \text{DistHeliports} + \\
(0.136 \pm 0.022) \times \text{FlightFreq25Mile}
\end{aligned} \tag{18}
$$

$$
\begin{aligned}
L90f9 = {} & (19.243 \pm 1.016) \times \text{(Intercept)} + \\
& (17.514 \pm 3.368) \times \text{HIHerbaceous200m} + \\
& (13.433 \pm 3.984) \times \text{WaterNat200m} + \\
& (-22.902 \pm 5.641) \times \text{WaterOnly200m} + \\
& (156.139 \pm 48.601) \times \text{Transportation5km} + \\
& (23.986 \pm 3.257) \times \text{WaterOnly5km} + \\
& (-0.047 \pm 0.013) \times \text{DistHeliports} + \\
& (0.125 \pm 0.023) \times \text{FlightFreq25Mile} + \\
& (1.273 \pm 0.25) \times \text{RddMajorPt}
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
L90f10 = {} & (17.375 \pm 0.99) \times \text{(Intercept)} + \\
& (16.162 \pm 3.28) \times \text{HIHerbaceous200m} + \\
& (13.104 \pm 3.88) \times \text{WaterNat200m} + \\
& (-21.727 \pm 5.494) \times \text{WaterOnly200m} + \\
& (163.822 \pm 47.333) \times \text{Transportation5km} + \\
& (24.217 \pm 3.172) \times \text{WaterOnly5km} + \\
& (-0.045 \pm 0.013) \times \text{DistHeliports} + \\
& (0.11 \pm 0.023) \times \text{FlightFreq25Mile} + \\
& (1.3 \pm 0.244) \times \text{RddMajorPt}
\end{aligned} \tag{20}
$$

Appendix C. Code

```
#Import Data
cleandata <- readRDS('cleandata')

 #Install R-Packages
 library(tidyverse)
 library(stargazer)
 devtools::install_github("timelyportfolio/parcoords")

## Parallel coordinates
# This function was adopted from  'timelyportfolio' on github.

myparacoords <- function(dataSource, colorBy = "LCLUCI.labels", width =
    1200, height = 600){
  parcoords::parcoords(
  dataSource
  ,reorderable = T
  ,rownames = FALSE
  ,alpha=0.5
   ,axisDots = 0
  ,mode = "queue"
  ,rate = 1
  # ,autoresize = TRUE
  ,width = width
  ,height = height
  ,brushMode = "1d-axes"
  ,color = list(colorScale = htmlwidgets::JS('d3.scale.category10()'),
                    colorBy = colorBy)

)
}

```

```
31 ## Example of Future Use:
32 #These are the variables that seem important from exhaustive linear
       regression for one specific frequency: "HIHerbaceous200m"  "
       RecCon5km"  "DistHeliports"  "DistRoadsMajor"  "FlightFreq25Mile"
       "RddMajorPt". However these are the points we have data for, and
       they are of potentially poor quality; Variables with information:
       Herbaceous200m  RecCon5km  _DistAirports_  DistRoadsMajor
       RddMajorPt
33
34 myparacoords(cleandata[, c("LCLUCI.labels", "HIHerbaceous200m", "
       DistModerateAirports", "DistHighAirports", "RecCon5km", "
       DistHeliports", "DistRoadsMajor", "FlightFreq25Mile", "RddMajorPt
       ","PopDensity_2010_50km", "park", "L90f1")])
```

Listing C.1. Parallel Coordinates Plot

Bibliography

1. D. J. Mennitt, K. Fristrup, K. R. Sherrill, and L. Nelson, "Mapping sound pressure levels on continental scales using a geospatial sound model," in *inter.noise*, 2013, pp. 1–12.

2. D. J. Mennitt, K. R. Sherrill, and K. Fristrup, "A geospatial model of ambient sound pressure levels in the contiguous United States," *The Journal of the Acoustical Society of America*, vol. 135, pp. 2746–2764, 2014. [Online]. Available: http://dx.doi.org/10.1121/1.4870481

3. D. J. Mennitt and K. M. Fristrup, "Influential factors and spatiotemporal patterns of environmental sound levels," in *inter.noise*, 2015.

4. D. J. Mennitt and K. M. Fristrup, "Influence factors and spatiotemporal patterns of environmental sound levels in the contiguous United States," *Noise Control Engineering Journal*, vol. 64, pp. 342–353, 5 2016.

5. B. Benson, "An Initial Ambient Noise Database Based on National Park Service Data." Master's thesis, Air Force Institute of Technology, Wright Patterson Air Force Base, OH, 2017.

6. T. Lumley, "leaps: Regression Subset Selection," 2017. [Online]. Available: https://cran.r-project.org/package=leaps

7. D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ: John Wiley & Sons, 2012.

8. P. Gaski, "Characterizing and Classifying Acoustical Ambient Sound Profiles," Master's thesis, Air Force Institute of Technology, Wright Patterson Air Force Base, OH, 2015.

9. B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti, "Soundscape Ecology: The Science of

Sound in the Landscape," *BioScience*, vol. 61, pp. 203–216, 3 2011. [Online]. Available: https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2011.61.3.6

10. K. F. Fluitt, T. J. Mermagen, and S. Letowski, "Auditory Perception in an Open Space : Detection and Recognition," US Army Research Lab, Tech. Rep. June, 2015.

11. DoD, "Defense Industrial Base Capabilities Study: Force Application," Tech. Rep., 2004. [Online]. Available: http://www.acq.osd.mil/ip

12. I. Waitz, S. Lukachko, and J. Lee, "Military Aviation and the Environment: Historical Trends and Comparison to Civil Aviation," in *AIAA International Air and Space Symposium and Exposition: The Next 100 Years*, 2003. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2003-2620

13. H. H. Hubbard, *Aeroacoustics of Flight Vehicles - Theory and Practice*, 1991, vol. 2, no. Noise Control. [Online]. Available: http://www.dtic.mil/dtic/tr/fulltext/u2/a241141.pdf

14. E. Murphy and E. A. King, "Chapter 4 - Strategic Noise Mapping," in *Environmental Noise Pollution*, 2014, pp. 81–121. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124115958000045

15. S. Goudreau, C. Plante, M. Fournier, A. Brand, Y. Roche, and A. Smargiassi, "Estimation of Spatial Variations in Urban Noise Levels with a Land Use Regression Model," *Environment and Pollution*, vol. 3, pp. 48–58, 9 2014. [Online]. Available: http://www.ccsenet.org/journal/index.php/ep/article/view/37895

16. E. D. Walker, J. E. Hart, P. Koutrakis, J. M. Cavallari, T. VoPham, M. Luna, and F. Laden, "Spatial and temporal determinants of A-weighted and frequency specific sound levelsAn elastic net approach," *Environmental Research*, vol. 159, pp. 491–499, 11 2017. [Online]. Available: http://dx.doi.org/10.1016/j.envres.2017.08.034

17. V.-S. Wang, E.-W. Lo, C.-H. Liang, K.-P. Chao, B.-Y. Bao, and T.-Y. Chang, "Temporal and spatial variations in road traffic noise for different frequency components in metropolitan Taichung, Taiwan," *Environmental Pollution*, vol. 219, pp. 174–181, 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0269749116318280

18. K. Hamad, M. Ali Khalil, and A. Shanableh, "Modeling roadway traffic noise in a hot climate using artificial neural networks," *Transportation Research Part D: Transport and Environment*, vol. 53, pp. 161–177, 2017.

19. N. D. Merchant, K. M. Fristrup, M. P. Johnson, P. L. Tyack, M. J. Witt, P. Blondel, and S. E. Parks, "Measuring acoustic habitats," *Methods in Ecology and Evolution*, vol. 6, pp. 257–265, 2015.

20. T. C. Mullet, J. M. Morton, S. H. Gage, and F. Huettmann, "Acoustic Footprint of Snowmobile Noise and Natural Quiet Refugia in an Alaskan Wilderness," *Natural Areas Journal*, vol. 37, pp. 332–349, 7 2017. [Online]. Available: http://www.bioone.org/doi/10.3375/043.037.0308

21. N. Koper, L. Leston, T. M. Baker, C. Curry, and P. Rosa, "Effects of ambient noise on detectability and localization of avian songs and tones by observers in grasslands," *Ecology and Evolution*, vol. 6, pp. 245–255, 1 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26811789

22. G. Shannon, M. F. McKenna, L. M. Angeloni, K. R. Crooks, K. M. Fristrup, E. Brown, K. A. Warner, M. D. Nelson, C. White, J. Briggs, S. McFarland, and G. Wittemyer, "A synthesis of two decades of research documenting the effects of noise on wildlife," *Biological Reviews*, vol. 91, pp. 982–1005, 2016.

23. E. Hoglund, D. Brungart, N. Iyer, J. Hamil, F. Mobley, and J. Hall, "Auditory acuity for aircraft in real-world ambient environments." *The Journal of the*

Acoustical Society of America, vol. 128, pp. 164–171, 2010. [Online]. Available: http://dx.doi.org/10.1121/1.3438480

24. M. McDaniel and Z. Hall, "Acoustic Situational Awareness for Survivability," *Aircraft Surviability*, pp. 18–19, 2013.

25. S. Ambrose and C. Florian, "Sound Levels and Audibility of Common Sounds in Front-country and Transitional Areas in Grand Canyon National Park , 2007-2008; NPS Report No. GRCA-08-04," National Park Service, Castle Valley, UT, Tech. Rep., 2008.

26. M. Buchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *Eurasip Journal on Applied Signal Processing*, vol. 2005, pp. 2991–3002, 2005.

27. Z. Azkorra, G. Pérez, J. Coma, L. F. Cabeza, S. Bures, J. E. Álvaro, A. Erkoreka, and M. Urrestarazu, "Evaluation of green walls as a passive acoustic insulation system for buildings," *Applied Acoustics*, vol. 89, pp. 46–56, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.apacoust.2014.09.010

28. D. D. Ballweg, "The effect of neonatal intensive care unit noise on the habituation of neonatal chicks," Master's thesis, Air Force Institute of Technology, 1991.

29. K. D. Kryter, "Acoustical model and theory for predicting effects of environmental noise on people," *The Journal of the Acoustical Society of America*, vol. 125, pp. 3707–3721, 2009. [Online]. Available: http://dx.doi.org/10.1121/1.3125320

30. L. Nelson, M. Kinseth, and T. Flowe, "Explanatory variable generation for geospatial sound modeling standard operating procedure. Natural Resource Report NPS/NRSS/NRR 2015/936," National Park Service Inventory, Fort Collins, Colorado, Tech. Rep., 2015.

31. K. R. Sherrill, "GIS Metrics - Soundscape Modeling Standard Operating Procedure. Sherrill, K. R. 2012. GIS metrics - soundscape modeling: Standard

operating procedure. Natural Resource Report NPS/NRSS/IMD/NRR," National Park Service, Fort Collins, Colorado, Tech. Rep., 2012. [Online]. Available: http://www.nature.nps.gov/publications/nrpm

32. J. A. Casey, R. Morello-Frosch, D. J. Mennitt, K. Fristrup, E. L. Ogburn, and P. James, "Race/ethnicity , socioeconomic status , residential segregation , and spatial variation in noise exposure in the contiguous United States," *Environmental Health Perspectives*, vol. 125, pp. 1–10, 2017.

33. J. Popovich, "A Model of Ambient Noise Caused by Wind Flow," Master's thesis, Air Force Institute of Technology, Wright Patterson Air Force Base, OH, 2016.

34. T. M. Leung, C. Chau, and S. Tang, "Developing a multivariate model for predicting the noise annoyance responses to the acoustic environment containing both water and road traffic sounds," *Applied Acoustics*, vol. 127, pp. 284–291, 2017.

35. M. S. Alam, L. Corcoran, E. A. King, A. McNabola, and F. Pilla, "Modelling of intra-urban variability of prevailing ambient noise at different temporal resolution," *Noise Mapping*, vol. 4, pp. 20–44, 2017. [Online]. Available: http://www.degruyter.com/view/j/noise.2017.4.issue-1/noise-2017-0002/noise-2017-0002.xml

36. M. Haines, "Ambient neighbourhood noise and children's mental health," *Occupational and Environmental Medicine*, vol. 60, pp. 146–146, 2003. [Online]. Available: http://oem.bmj.com/cgi/doi/10.1136/oem.60.2.146

37. G. W. Evans and S. J. Lepore, "Nonauditory effects of noise on children: A critical review," *Children's Environments*, vol. 10, pp. 31–51, 1993.

38. C. Sieber, M. S. Ragettli, M. Brink, O. Toyib, R. Baatjies, A. Saucy, N. Probst-Hensch, M. A. Dalvie, and M. Röösli, "Land Use Regression Modeling of Outdoor Noise Exposure in Informal Settlements in Western Cape, South Africa," pp. 1–10, 2017. [Online]. Available: https://www.preprints.org/manuscript/201708.0035/v1

39. A. J. Torija and D. P. Ruiz, "A general procedure to generate models for urban environmental-noise pollution using feature selection and machine learning methods," *Science of the Total Environment*, vol. 505, pp. 680–693, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.scitotenv.2014.08.060

40. M. S. Ragettli, S. Goudreau, C. Plante, M. Fournier, M. Hatzopoulou, S. Perron, and A. Smargiassi, "Statistical modeling of the spatial variability of environmental noise levels in Montreal, Canada, using noise measurements and land use characteristics," *Journal of Exposure Science and Environmental Epidemiology*, vol. 26, pp. 597–605, 11 2016. [Online]. Available: http://www.nature.com/doifinder/10.1038/jes.2015.82

41. U.S. Navy, "Sonar." [Online]. Available: http://www.public.navy.mil/usff/environmental/Pages/Sonar.aspx

42. S. Amoser and F. Ladich, "Year-round variability of ambient noise in temperate freshwater habitats and its implications for fishes," *Aquatic Sciences*, vol. 72, pp. 371–378, 6 2010. [Online]. Available: http://link.springer.com/10.1007/s00027-010-0136-9

43. A. Pyzdek, "10 The World Through Sound: Refraction - Acoustics Today." [Online]. Available: http://acousticstoday.org/10-world-sound-reflection-refraction-principle-least-time/

44. A. M. von Benda-Beckmann, P. J. Wensveen, F. I. P. Samarra, S. P. Beerens, and P. J. O. Miller, "Correction: Separating underwater ambient noise from flow noise recorded on stereo acoustic tags attached to marine mammals," *The Journal of Experimental Biology*, vol. 219, pp. 2774–2774, 9 2016. [Online]. Available: http://jeb.biologists.org/lookup/doi/10.1242/jeb.148197

45. M. J. Buckingham, B. V. Berkhout, and S. a. L. Glegg, "Acoustic daylight: Imaging the ocean with ambient noise." *The Journal of the Acoustical Society of America*, vol. 91, 1992.

46. C. Erbe, C. Reichmuth, K. Cunningham, K. Lucke, and R. Dooling, "Communication masking in marine mammals: A review and research strategy," *Marine Pollution Bulletin*, vol. 103, pp. 15–38, 2016.

47. H. F. Boersma, "Characterization of the natural ambient sound environment: Measurements in open agricultural grassland," *The Journal of the Acoustical Society of America*, vol. 101, pp. 2104–2110, 4 1997. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.418141

48. N. Garg, A. Sinha, V. Gandhi, R. Bhardwaj, and A. Akolkar, "A pilot study on the establishment of national ambient noise monitoring network across the major cities of India," *Applied Acoustics*, vol. 103, pp. 20–29, 2 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0003682X15002546

49. H. Ryu, I. K. Park, B. S. Chun, and S. I. Chang, "Spatial statistical analysis of the effects of urban form indicators on road-traffic noise exposure of a city in South Korea," *Applied Acoustics*, vol. 115, pp. 93–100, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.apacoust.2016.08.025

50. D. Bormpoudakis, J. Sueur, and J. D. Pantis, "Spatial heterogeneity of ambient sound at the habitat type level: ecological implications and applications," *Landscape Ecology*, vol. 28, pp. 495–506, 3 2013. [Online]. Available: http://link.springer.com/10.1007/s10980-013-9849-1

51. Congress, "Public Law 93-620: Grand Canyon National Park Enlargement Act," pp. 2089–2093, 1975. [Online]. Available: https://www.gpo.gov/fdsys/pkg/STATUTE-88/pdf/STATUTE-88-Pg2089.pdf

52. A. C. Keyel, S. E. Reed, M. F. McKenna, and G. Wittemyer, "Modeling anthropogenic noise propagation using the Sound Mapping Tools ArcGIS toolbox," *Environmental Modelling & Software*, vol. 97, pp. 56–60, 2017. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1364815217301500

53. G. Becker and A. Güdesen, "Passive sensing with acoustics on the battlefield," *Applied Acoustics*, vol. 59, pp. 149–178, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003682X99000237

54. D. Vergun, "New tools developed by Army engineers map safer routes from shore to inland objectives — Article — The United States Army," 2017. [Online]. Available: https://www.army.mil/article/198630/new_tools_developed_by_army_engineers_map_safer_routes_from_shore_to_inland_objectives

55. U.S. Army, "Acoustic Research Complex (ARC)," 2016. [Online]. Available: http://www.wsmr.army.mil/testcenter/TE/testing/landf/Pages/AcousticResearchComplex(ARC).aspx

56. J. Maxime, X. Alameda-Pineda, L. Girin, and R. Horaud, "Sound representation and classification benchmark for domestic robots," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2014, pp. 6285–6292.

57. R. Serizel, V. Bisot, S. Essid, and G. Richard, "Machine listening techniques as a complement to video image analysis in forensics," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 9 2016, pp. 948–952. [Online]. Available: http://ieeexplore.ieee.org/document/7532497/

58. A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 8 2016, pp. 1128–1132. [Online]. Available: http://ieeexplore.ieee.org/document/7760424/

59. J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2015-Augus. IEEE, 4 2015, pp. 171–175. [Online]. Available: http://ieeexplore.ieee.org/document/7177954/

60. Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3 2012, pp. 333–336. [Online]. Available: http://ieeexplore.ieee.org/document/6287884/

61. I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 540–552, 3 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7003973/

62. J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 10 2013, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/document/6701857/

63. J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, pp. 367–377, 2013.

64. J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using Gabor filterbank features," in *2015 23rd European Signal Processing Conference, EUSIPCO 2015*, 2015, pp. 714–718.

65. J. George, A. Cyril, B. I. Koshy, and L. Mary, "Exploring Sound Signature for Vehicle Detection and Classification Using ANN," *International Journal on Soft Computing*, vol. 4, pp. 29–36, 2013. [Online]. Available: http://www.airccse.org/journal/ijsc/papers/4213ijsc03.pdf

66. M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, pp. 2895–2907, 2003.

67. S. Chachada and C. C. J. Kuo, "Environmental sound recognition: A survey," in

2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013, no. December 2014, 2013.

68. R. V. Sharan and T. J. Moir, "Robust acoustic event classification using deep neural networks," *Information Sciences*, vol. 396, pp. 24–32, 2017.

69. Q. Nguyen and J. S. Choi, "Matching pursuit based robust acoustic event classification for surveillance systems," *Computers and Electrical Engineering*, vol. 57, pp. 43–54, 1 2017. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0045790616307157

70. Y. Li, Q. Wang, X. Li, X. Zhang, Y. Zhang, A. Chen, Q. He, and Q. Huang, "Unsupervised detection of acoustic events using information bottleneck principle," *Digital Signal Processing: A Review Journal*, vol. 63, pp. 123–134, 2017.

71. Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282. [Online]. Available: http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdfhttp://ieeexplore.ieee.org/document/598994/

72. R Core Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2017. [Online]. Available: https://www.r-project.org/

73. J. Wing and M. Kuhn, "caret: Classification and Regression Training," 2017. [Online]. Available: https://cran.r-project.org/package=caret

74. T. Wei and V. Simko, "R package "corrplot": Visualization of a Correlation Matrix," 2017. [Online]. Available: https://github.com/taiyun/corrplot

75. H. Wickham, "tidyverse: Easily Install and Load the 'Tidyverse'," 2017. [Online]. Available: https://cran.r-project.org/package=tidyverse

76. F. E. Harrell Jr, with contributions from Charles Dupont, and many others., "Hmisc: Harrell Miscellaneous," 2017. [Online]. Available: https://cran.r-project.org/package=Hmisc

77. M. Hlavac, "stargazer: Well-Formatted Regression and Summary Statistics Tables," Cambridge, USA, 2015. [Online]. Available: http://cran.r-project.org/package=stargazer

78. M. N. Wright and A. Ziegler, "{ranger}: A Fast Implementation of Random Forests for High Dimensional Data in {C++} and {R}," *Journal of Statistical Software*, vol. 77, pp. 1–17, 2017.

79. L. Komsta and F. Novomestky, "moments: Moments, cumulants, skewness, kurtosis and related tests," 2015. [Online]. Available: https://cran.r-project.org/package=moments

80. NOAA Office for Coastal Management, "C-CAP Regional Land Cover and Change." [Online]. Available: https://www.coast.noaa.gov/digitalcoast/data/ccapregional

81. "Ladder Creek Falls Washington Trails Association." [Online]. Available: https://www.wta.org/go-hiking/hikes/ladder-creek-falls

82. "Minute Man National Historical Park (U.S. National Park Service)." [Online]. Available: https://www.nps.gov/mima/index.htm

83. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2009.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
22-03-2018	Master's Thesis	Sept 2016-Mar 2018

4. TITLE AND SUBTITLE

CHARACTERIZATION OF AMBIENT NOISE

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Rachel C. Ramirez, Maj, U.S. Air Force

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-18-M-155

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

United States Air Force Research Lab
711th Human Performance Wing
2610 Seventh Street, Bldg. 441
Wright-Patterson AFB, OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)

711 HPW

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT An Air Force sponsor is interested in improving an acoustic detection model by providing better estimates to characterize the background noise of various environments. This would help inform decision makers on the probability of acoustic detection of different systems of interest given different levels of noise. Data mining and statistical learning techniques are applied to a National Park Service acoustic summary data set to find overall trends over varying environments. Linear regression, conditional inference trees, and random forest techniques are discussed. Findings indicate only sixteen geospatial variables at different resolutions are necessary to characterize the first ten 1/3 octave band frequencies of the L90 band using just the linear regression. The accuracy of the regression model is within 2 to 6 decibels and depends on the frequency of interest. This research is the first of its kind to apply linear regression to the national park service acoustic dataset, and second to apply random forests for predicting noise levels. Future research is needed to determine the accuracy of the model when applied outside of the national park service in intended Air Force operational environments.

15. SUBJECT TERMS

linear regression, geospatial variables, ambient noise

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Ray Hill, AFIT/ENS
U	U	U	U	151	19b. TELEPHONE NUMBER *(include area code)* (937) 255-6565, 7469; Rhill@afit.edu