2021

Modeling Continental-Scale Outdoor Environmental Sound Levels with Limited Data

Katrina Lynn Pedersen
Brigham Young University

Modeling Continental-Scale Outdoor Environmental

Sound Levels with Limited Data

Katrina Lynn Pedersen

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Mark K. Transtrum, Chair
Kent L. Gee
Sean Warnick
Ryan R. Jensen
Matthew J. Heaton

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

Modeling Continental-Scale Outdoor Environmental Sound Levels with Limited Data

Katrina Lynn Pedersen
Department of Physics and Astronomy, BYU
Doctor of Philosophy

Modeling outdoor acoustic environments is a challenging problem because outdoor acoustic environments are the combination of diverse sources and propagation effects, including barriers to propagation such as buildings or vegetation. Outdoor acoustic environments are most commonly modeled on small geographic scales (e.g., within a single city). Extending modeling efforts to continental scales is particularly challenging due to an increase in the variety of geographic environments. Furthermore, acoustic data on which to train and validate models are expensive to collect and therefore relatively limited. It is unclear how models trained on this limited acoustic data will perform across continental-scales, which likely contain unique geographic regions which are not represented in the training data.

In this dissertation, we consider the problem of continental-scale outdoor environmental sound level modeling using the contiguous United States for our area of study. We use supervised machine learning methods to produce models of various acoustic metrics and unsupervised learning methods to study the natural structures in geospatial data. We present a validation study of two continental-scale models which demonstrates that there is a need for better uncertainty quantification and tools to guide data collection. Using ensemble models, we investigate methods for quantifying uncertainty in continental-scale models. We also study methods of improving model accuracy, including dimensionality reduction, and explore the feasibility of predicting hourly spectral levels.

Keywords: sound level models, geospatial modeling, machine learning, ensemble models, uncertainty quantification, GIS, environmental noise, validation

ACKNOWLEDGMENTS

I would like to thank my advisor, Mark Transtrum, for his patience and mentorship. I deeply appreciate the time he took out of his busy schedule to discuss problems thoughtfully and review the work for this dissertation. I'm also grateful for the opportunities he helped provide for me to teach and mentor undergraduate students.

I would also like to thank my committee (Kent Gee, Sean Warnick, Ryan Jensen, and Matthew Heaton) for their time and thoughtful feedback. I especially want to thank Kent Gee for his guidance and advice both in my research and writing. I'm also grateful to Ryan Jensen for sharing his expertise in geography and collaborating on a paper. Additionally, I am grateful for the suggestions of and collaboration with Blue Ridge Research and Consulting, LLC. In particular, I'm grateful to Shane Lympany for his feedback and insight. I'm also grateful for the support and encouragement of the faculty and staff of the Department of Physics and Astronomy.

Finally, I would like to thank my family. I'm especially grateful to my Great-Grandmother Thelma, Nana Jean, Grandpa Kent, and my parents who helped instill a love of learning in me. I'm grateful also to the rest of my family for always being willing to talk on the phone, especially since the COVID-19 pandemic began. Their love and support have been invaluable to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Outdoor acoustic environments are the result of diverse sources and complex propagation effects within a geographic area. Possible acoustic sources include traffic noise, the electrical hum of generators, bird song, rivers, and human conversation, while propagation effects include barriers to propagation such as buildings or vegetation and their absorptive and reflective properties. Modeling outdoor acoustic environments is therefore challenging because it requires knowledge of all physical contributions to the acoustic environment. Additionally, complete characterization of an acoustic environment over different times and frequency bands requires further knowledge of the temporal and spectral dependence and effects of contributing factors to an acoustic environment. Modeling outdoor acoustic environments across large geographic scales (i.e., continental scales) faces further challenges due to the wide range of possible environmental conditions. However, continental-scale modeling of outdoor acoustic environments has potentially broad applications.

Accurate characterization of outdoor acoustic environments, particularly in areas with significant environmental noise (i.e., unwanted sound due to anthropogenic, or human, activity), may benefit public health studies and have implications for social justice (e.g., community noise ordinances). Environmental noise is correlated with depression and anxiety [2, 3], increased cardiovascular risk [4-8], increased risk of breast cancer [9], annoyance[10], sleep disturbance [10, 11], hearing loss [11], and more in humans. It is estimated that environmental noise exposure causes 12,000 premature deaths and contributes to 48,000 new cases of ischemic heart disease per year in the European territory [12]. Indeed, environmental noise has been called 'the new secondhand smoke' [11]. On the other hand, natural acoustic environments have been linked to physiological and psychological benefits in humans [13]. Therefore, accurate modeling of outdoor environmental sound levels may also have commercial applications for real estate and urban development [14, 15].

Environmental noise has also been linked to changes in animal behavior and species interactions, particularly in animals that utilize auditory signals. Many mammals, birds, amphibians, and insects use sound for tasks including antipredator defense [16], reproduction, and communication. Although the impact of environmental noise on wildlife is unclear, environmental noise has been indicated as a causal factor for changes in avian behavior and community diversity [17, 18], marine life[19], and anurans (i.e., frogs and toads) [20, 21]. Accurate characterization of outdoor acoustic environments, especially for areas in which environmental noise is prevalent, may improve studies of the relationship between environmental noise and changes in animal behavior and species interactions, as well as have implications for wildlife policy and management.

1.2 Continental-scale environmental sound level modeling

Remote sensing has greatly increased the amount of geospatial data available, much of which is applicable to environmental sound modeling. In a foundational study, Mennitt et al. [1, 22, 23] used machine learning to relate geospatial features and acoustic metrics to produce continental-scale sound maps. Note that the only publicly available continental-scale sound map to our knowledge is generated using methods described by Mennitt et al. and made publicly available through the National Park Service [1, 24]. Elsewhere, machine learning and/or land-use regression models have been used to map noise in urban areas on smaller geographic scales [25-27]. These modeling efforts, both for continental-scale and smaller geographic scales, utilize geospatial data to predict statistical summaries of acoustic environments, therefore circumventing the need for detailed, physical modeling through the use of data-driven approaches. As geospatial data have become more widely available and improved in accuracy and precision, many problems, such as environmental sound level modeling, have become more feasible.

In this dissertation, we build upon the foundation laid by Mennitt et al. to research continental-scale outdoor environmental sound modeling. We implement supervised machine learning models trained on geospatial data and measured acoustic data from training sites throughout the contiguous United States to create continental-scale models of outdoor environmental sound levels at a 270-m spatial resolution.

Acoustic environments are described by various acoustic metrics, including statistical time-exceeded levels (L_{NN}) which correspond to the sound pressure level exceeded NN% of the time (for a given day, night, hour, or season). Levels may also be described as a function of frequency (e.g., levels at one-third octave frequency bands), and different weightings may be applied to levels as a function of frequency. For example, the A-weighting is meant to better quantify sound as perceived by humans. In this dissertation, we present models for the summer daytime A-weighted L_{50} levels (i.e., median summer daytime A-weighted levels) and summer hourly L_{50} one-third octave band levels (i.e., median summer levels as a function of hour and frequency). The L_{50} is a reasonable statistical time-exceeded level to start with because it is a good descriptor of typical sound levels. However, future work could utilize the same methods described in this dissertation for prediction of other acoustic metrics, such as the L_{10} or L_{90}, which correspond to the sound pressure level exceeded ten percent of the time and ninety percent of the time, respectively. L_{10} levels are often used in traffic noise studies and are good indicators of loud, intermittent events while L_{90} levels describe background or ambient levels and may be more useful for studies of noise pollution.

One challenge of this data-driven approach is that acoustic training data are relatively limited (approximately 500 unique geographic sites) due to the high costs associated with collecting high-fidelity acoustic data. Many of the more dramatic successes of machine learning are known to require millions of training instances [28, 29]; therefore, it is unclear how model performance will transfer to continental-scales, given the limited training data.

In particular, traditional validation metrics estimate model accuracy for inputs that are statistically similar to the training data. However, the training data (composed of only ~500 unique geographic sites) are unlikely to be statistically similar to or representative of the 110 million sites in the contiguous United States for which model predictions are made. (Note that the number of sites in the contiguous United States is a result of sampling with a 270-m spatial resolution.) Indeed, a clustering analysis indicated the distribution of training data is significantly different from that of the contiguous United States [30]. Therefore, traditional validation metrics are likely overly optimistic and insufficient for estimating expected predictive error for predictions in the contiguous United States. This motivates the need for quantitative estimates of model transferability, i.e., ability of a model to accurately make predictions for novel inputs that are statistically different from the training set. In this dissertation, we research optimal methods for modeling continental-scale outdoor sound levels as well as improving model performance. We also research methods of estimating uncertainty for model predictions in locations atypical of training data (i.e., model transferability).

1.3 Organization

Chapter 2 is a manuscript to be submitted as an Express Letter in the Journal of Acoustical Society of America. In it, we present a validation study of two continental-scale models of the summer daytime A-weighted L_{50}, an acoustic metric that describes the median sound pressure level during an average day in summer, for the contiguous United States. More specifically, we validate the National Park Service published sound map and an ensemble model map created by us in a natural environment, in an urban environment, and on a holdout set of 25 sites. We observe significant differences between the two models' predictions as well as large errors at many holdout validation sites. These discrepancies are much larger than those estimated from statistical validation metrics such as leave-one-out. We attribute them to having limited training data which force models to extrapolate to make predictions when geospatial inputs are not similar to those in the training data. These results motivate some skepticism of the accuracy of continental-scale sound models as well as motivate further efforts in improving model predictions and their corresponding uncertainty estimates, which we explore in subsequent chapters. Indeed, we caution against using continental-scale sound level predictions without a thorough consideration for the possibly uncertainty in such predictions.

Chapter 3 is a manuscript to be submitted in the Journal of Acoustical Society of America. In Chapter 3, we investigate estimation of model transferability, or predictive error, within the contiguous United States, inspired by techniques from the uncertainty quantification community. Uncertainty quantification has existed as long as probability and statistics and is the science of identifying, quantifying, and reducing uncertainty in model predictions [31]. There are two main types of uncertainty: aleatoric and epistemic uncertainty [31]. Aleatoric uncertainty, or statistical uncertainty, is inherent to a problem, and hence cannot be reduced [31]. Epistemic, or systematic, uncertainty in a model arises from incomplete knowledge or missing physics and can be reduced through improved modeling methods [31]. Following Kennedy et al. [32], uncertainty types may be further refined into six classes: parameter uncertainty, model inadequacy, residual variability, parametric variability, observation error, and code uncertainty. Of these six quantities, we focus on quantifying model inadequacy (also called structural uncertainty) in Chapter 3.

Structural uncertainty, or uncertainty in the form of the model due to limited knowledge of the true underlying mechanisms of data generation, is only one of six classes of uncertainty; however, we anticipate that is it one of the primary sources of uncertainty for continental-scale environmental sound level modeling. Each of the six models comprising the ensemble is from a different model class, so the range of predictions within the ensemble provides some measure of the structural uncertainty. The median predicted level from the six models is used for ensemble model predictions and the standard deviation of the ensemble model's predictions as a surrogate estimate of the accuracy of ensemble model predictions. Although estimates of structural uncertainty are relatively low for much of the contiguous United States with mean and median levels of 3.5 and 2.9 dBA, respectively, 6.2% of sites have an uncertainty greater than 6 dBA. The mean uncertainty of these sites (i.e., sites with uncertainties greater than 6 dBA) is about 12.6 dBA, which corresponds to a change in perceived loudness by more than a factor of two. Additionally, structural uncertainty estimates only indicate minimum levels of uncertainty because we only quantify one source of uncertainty. Future work may include quantifying additional sources of uncertainty, such as uncertainty associated with measuring the geospatial or acoustic data.

In Chapter 4, we present the results of training two large ensemble models for the prediction of summer hourly spectra. Ensembles are composed of 100 models for a specific model class trained of different bootstrap subsets of the data. Similar to ensembles in Chapter 3, these ensembles quantify the epistemic uncertainty. However, they utilize bootstrap samples rather than different model classes to generate diversity within the ensembles. In contrast to previous chapters, this chapter examines the results of environmental sound level predictions which vary with time of day and frequency band. We show predicted and measured levels at three training data sites characterized by road traffic noise and insect and bird activity. Predictions indicate that ensemble models are able to predict physical-looking spectra, however additional data are likely required for improving model accuracy and reducing uncertainty in model predictions.

Chapter 5 is a manuscript to be submitted in the Journal of Acoustical Society of America. Chapter 5 investigates methods for feature selection to improve model accuracy, uncertainty estimates, and model interpretability. Following an initial screening, 120 geospatial features are reduced to a set of 51 features. Then, we employ four different methods for further feature reduction. Interestingly, the different methods all reduce the number of geospatial features to 15 without significant degradation in model predictive error. However, differences in feature subsets and their corresponding ensemble predictions demonstrate that results of feature selection are sensitive to details of the problem formulation. These results motivate the need for more sophisticated dimensionality reduction techniques. We explore one such method, manifold learning using diffusion maps, in Chapter 7.

Chapter 6 is a manuscript to be submitted to *Applied Geography*. In Chapter 6, we discuss the results of a clustering analysis of 51 geospatial features, selected and scaled for the purpose of modeling environmental sound levels, over the contiguous United States. Due to how the geospatial features were selected and scaled, it is possible that different clusters may correspond to different acoustic environments. We observe that largescale patterns of vegetation are influential in the cluster model, possibly indicating that different types of vegetation correspond to different acoustic environments. This indicates that, to the extent acoustic environments vary with vegetation type, they are in principle resolvable by geospatial data. Possible applications of the clustering analysis are discussed, including applications in public health and ecology and for environmental sound modeling.

Chapter 7 explores the use of diffusion maps for dimensionality reduction. Findings from Chapter 5 illustrate that feature selection results are sensitive to details of the problem formulation. However, feature reduction is still an important problem. Unlike feature selection methods, which attempt only to identify a reduced set of features, diffusion maps utilize nonlinear relationships between features to identify a reduced set of coordinates. Chapter 7 discusses the application and advantages of diffusion maps for dimensionality reduction of the CONUS geospatial data. In particular, results may aid in identifying optimal locations for acoustic data collection as well as provide a means of making environmental sound level predictions for regions outside of CONUS.

Finally, we present concluding thoughts and suggestions for future work in Chapter 8. Although continental-scale sound modeling has potentially broad applications, this dissertation demonstrates that the technology is limited by the amount of available acoustic data leading to large uncertainties in predicted sound levels. We strongly caution against using such models without accounting for uncertainty in model predictions. Future work may include identifying optimal methods for acoustic data collection as well as incorporating physics-based modeling approaches, both of which have the potential to improve model accuracy and uncertainty estimates.

Chapter 2

Validating two geospatial models of continental-scale environmental sound levels*

Katrina Pedersen,[1] Mark K. Transtrum,[1] Kent L. Gee,[1] Shane V. Lympany,[2] Michael M. James,[2] and Alexandria R. Salton[2]

[1)] Physics and Astronomy, Brigham Young University, Provo, UT, 84602
[2)] Blue Ridge Research and Consulting, LLC, Asheville, NC, 28801

2.1 Abstract

Modeling acoustic environments is a challenging problem. This paper reports on a validation study of two continental-scale machine learning models using summer daytime A-weighted L_{50} as a validation metric. The first model was developed by the National Park Service while the second was developed by the present authors. Validation errors greater than 20 dBA are observed. Large errors are attributed to limited acoustic training data. Validation environments are geospatially dissimilar to training sites, requiring models to extrapolate beyond their training sets. Results motivate further work in optimal data collection and uncertainty quantification.

*Chapter 2 is a manuscript to be submitted as an Express Letter in the Journal of Acoustical Society of America.

2.2 Introduction

Modeling environmental sound levels is challenging because ambient sound is the accumulation of many effects, including diverse sources, barriers to propagation, etc., and can vary with time of day or season. Accurate characterization of acoustic environments has potentially broad applications, and recent data-driven approaches have led to continental-scale models that circumvent the need for detailed, physical modeling. However, environmental sound level modeling on continental scales is still in its relatively early phases of development.

This paper presents a validation study of two continental-scale sound maps of the summer daytime A-weighted L_{50}. In Section 2.3 we summarize the models and present a comparison of their predictions for the Contiguous United States (CONUS). Sections 2.4 and 2.5 validate both models in a natural environment (a region in Grand Teton National Park) and an urban environment (Washington, D.C.), respectively. Section 2.6 evaluates both models' performance on a holdout validation set and Section 2.7 gives concluding remarks.

Significantly, this study demonstrates that the models may disagree both with each other and with holdout acoustic measurements by more than 20 dBA. We attribute these large errors to the limited acoustic data available for training. Indeed, both models were trained on data from fewer than 500 geographically unique sites, which are likely not representative of all geospatial environments of interest within CONUS. In other words, validation environments are often geospatially dissimilar to all available training environments. When models make predictions at these unrepresented sites, they necessarily extrapolate into unsampled domains, referred to as “extrapolation regions” below. For data-driven approaches that do not incorporate any physical insight, it is difficult to identify the relationship between geospatial features and sound levels in extrapolation regions, leading to inaccurate model predictions.

2.3 Comparison of models on CONUS-scale predictions

The National Park Service (NPS) map of predicted summer daytime A-weighted L_{50} levels is the only publicly available continental-scale sound map [1, 24] to our knowledge. It has been used in fields such as public health [11, 33-35], landscape ecology and urban planning [36-38], ecology [39, 40], and aviation [41]. The NPS model was constructed using machine learning techniques and a training data set consisting of both acoustic and geospatial data. The acoustic data are comprised of acoustic measurements from 492 unique sites, 333 of which were located in quiet, uninhabited areas within national parks and 159 of which came from urban areas [1]. These acoustic data were combined with 45 geospatial layers at a 270-m spatial resolution for CONUS in a random forest algorithm to predict the summer daytime A-weighted L_{50} [1]. The reported Leave-One-Out (LOO) cross-validation Root-Mean-Square Error (RMSE) and Median Absolute Deviation (MAD) for the seasonal A-weighted L_{50} were 4.40 and 2.29 dBA, respectively [1]. The discrepancy between these two error metrics was attributed to outliers in the training data set. Readers interested in learning more about various validation metrics (e.g., holdout validation and LOO cross-validation) are referred to [42].

Similar to the NPS sound map, we utilize a training data set of acoustic measurements and geospatial layers to develop supervised learning algorithms. We use acoustic measurements from 496 unique sites (327 natural and 169 urban) and geospatial data consisting of 51 layers at 270-m resolution that were manually scaled for environmental sound-level modeling (see Appendix A for further information). Our approach uses of an ensemble of six different supervised machine learning models (gradient boosted regression trees, neural networks, k-nearest neighbors, support vector machines, kernel ridge regression, and Gaussian process regression). This ensemble approach allows us to assess predictive performance for a broader range of machine learning models. Each model class is trained independently, and the median prediction from the ensemble is taken as the ensemble prediction. Hyperparameters are tuned to

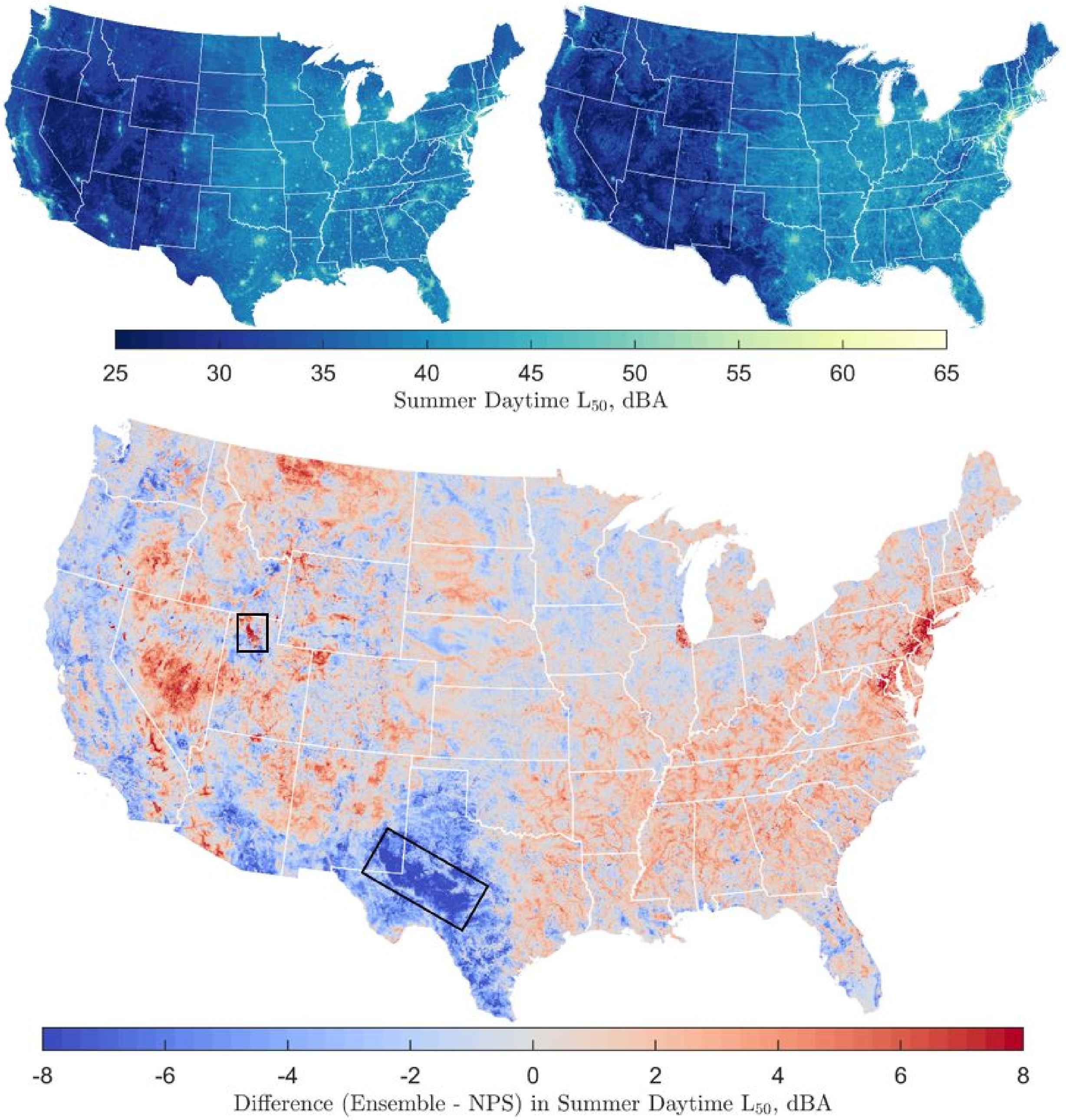
25
30
35
40
45
50
55
60
65
Summer Daytime L_{50}, dBA
-8
-6
-4
-2
0
2
4
6
8
Difference (Ensemble - NPS) in Summer Daytime L_{50}, dBA

Figure 2.1 CONUS summer daytime A-weighted L_{50} sound level predictions from the NPS model (top left) and the ensemble model (top right). The difference between the two CONUS map predictions (i.e., ensemble model predictions minus the NPS predictions) is on the bottom. Two areas of large differences are outlined. The color axis was constrained to +/- 8 dBA on the difference map, although differences were as large as 21 dBA.

minimize the LOO cross-validation MAD using the tree-structure Parzen estimator approach implemented in the Python library hyperopt [43, 44]. The LOO cross-validation RMSE and MAD for the ensemble model are 5.79 and 3.49 dBA, respectively. Due to differences in the training data (including acoustic data, geospatial layers, and scaling methods), these values are not directly comparable to those of the NPS model.

Figure 2.1 shows summer daytime A-weighted L_{50} CONUS predictions for the NPS model (top left) and the ensemble model (top right), and the difference between the two model predictions (bottom) in decibels. Despite large differences in certain areas (e.g., western Texas and New York City), the mean and median absolute differences are 2.07 and 1.64 dBA, respectively, indicating relative agreement between the two models for a large portion of CONUS. Indeed, the mean and median differences between the two models is smaller than the LOO RMSE and MAD cross-validation errors for either model. However, the maximum difference between the two maps is 21.30 dBA, which corresponds to a factor of more than four in increased loudness.

Areas of large differences likely correspond to extrapolation regions for either one or both of the models. For example, the Great Salt Lake in Utah (outlined on the difference map), which is a geographically unique region, distinct from any training site for the ensemble model (i.e., in an extrapolation region), is an area of large discrepancy between the two models. Much of western Texas extending into eastern New Mexico (outlined on the difference map) may also be in the extrapolation region of both models as neither model contains training data from that area. Considering smaller geographic regions in CONUS, particularly where training data are available, provides additional insight regarding model similarities, difference, and overall validation, as we show in subsequent sections.

2.4 Natural environment validation: Grand Teton National Park

Figure 2.2 shows the NPS predictions (2.2a) and ensemble model predictions (2.2b) for the summer daytime A-weighted L_{50} in an area of Grand Teton National Park. Training sites for the ensemble model (24 in this region) are marked by small circles and colored according to measured levels. The difference between the two maps (ensemble minus the NPS map) is shown in the Figure 2.2c. Figure 2.2d shows a histogram of residuals (observed sound level minus model predictions) overlaid with Probability Density Functions (PDFs) for training sites estimated using Matlab's ksdensity function. Note that model residuals are calculated using measured levels at each of the 24 available training sites and predicted values at the nearest 270-m raster site. We note that both the NPS and ensemble models have significant training data (at least 27 sites) within Teton County (i.e., the county containing this region), suggesting sites within this region are less likely to be in extrapolation regions of either model.

Although overall predicted levels are similar in this region, there are areas in which the difference map is saturated and the models differ by more than 8 dBA (i.e., up to 13.4 dBA). We note that the appearance of plus marks on the ensemble map (and therefore on the difference map) is a result of including land cover geospatial layers as features (i.e., the land cover maps also have plus-mark shapes). Interestingly, although both maps use geospatial layers with the same 270-m spatial resolution, the ensemble map shows greater variability over smaller spatial scales while the NPS map appears smoother; however, it is unknown how smooth the 'true' sound map is. Additionally, the NPS predictions are significantly higher than measured levels in the northern half of the map. This is evident in the residuals and is reflected in the summary error metrics at measured sites described in Table 2.1.

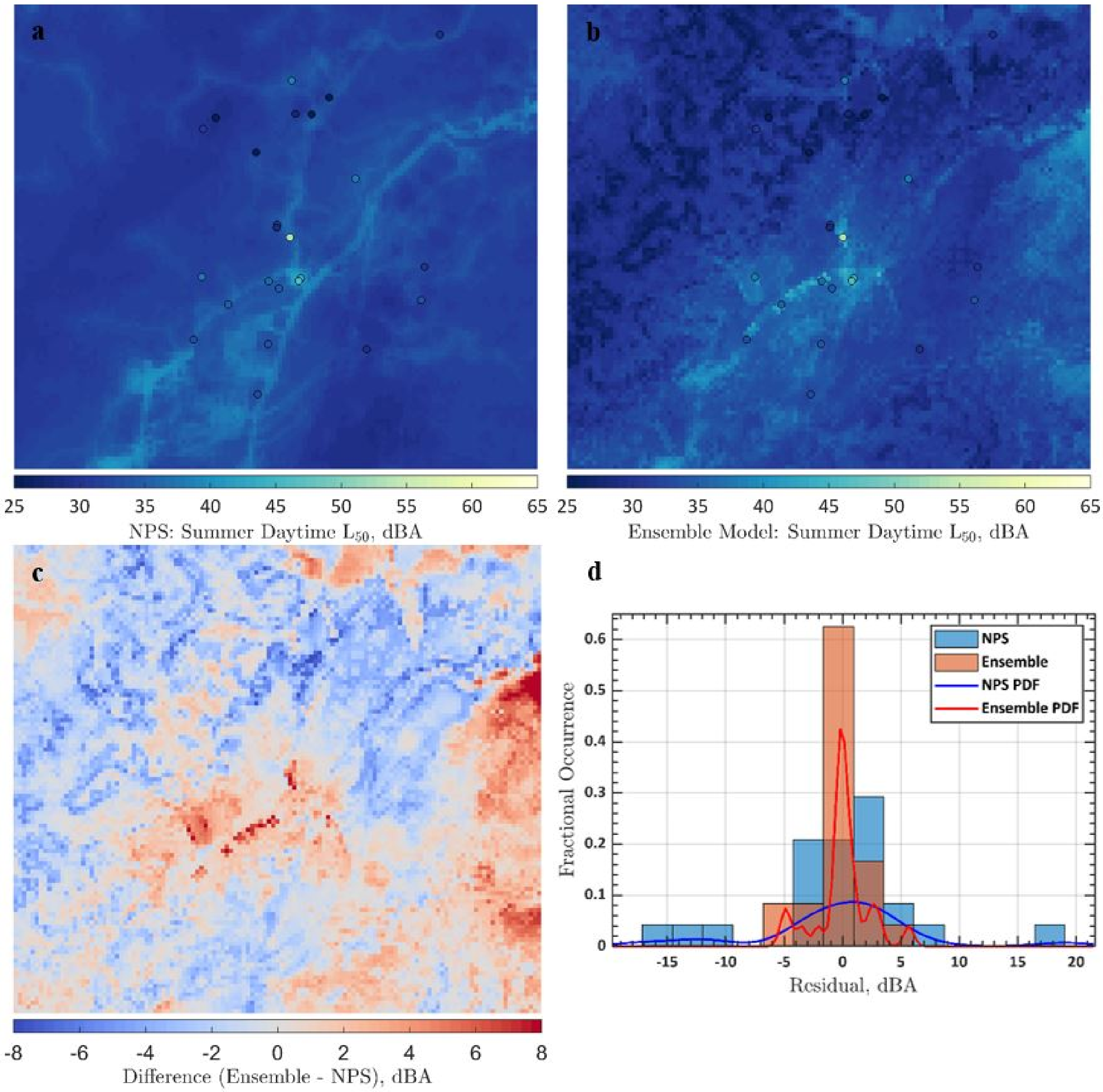

Figure 2.2 Summer daytime A-weighted L_{50} sound level predictions from the NPS model (a) and the ensemble model (b) for an area in Grand Teton National Park. Training sites for the ensemble model are marked by small circles and colored according to measured levels. The difference between the two map predictions (i.e., ensemble model predictions minus the NPS predictions) is shown in (c). A histogram and estimated PDFs of residuals at training sites included in the ensemble model's training data set are shown in (d).

For this region, residuals for both models are centered around zero with the standard deviations of the ensemble model residuals being lower. Note that the errors reported for the ensemble model are fitting errors; therefore, we expect these to be small. The NPS errors are larger, but still relatively small. Given that most training sites (333 of 492) for the NPS model are from national parks and at least 27 training sites are within Teton County, we expect that these validation sites are not in extrapolation regions. Indeed, it is possible that some of these sites were used as training sites in the NPS model.

Table 2.1 RMSE and MAD values for the NPS model and the ensemble model for sites in Grand Teton National Park indicated by the circles in Figures 2.2a and 2.2b.

Model	RMSE, dBA	MAD, dBA
NPS Model	6.8	3.0
Ensemble Model	2.3	0.6

2.5 Urban environment validation: Washington, D.C.

One of the most striking areas on the CONUS difference map (Figure 2.1, bottom) is the Washington, D.C. area. It is also an area from which a significant fraction of the urban training data (30 sites) for the ensemble model was drawn. This makes it a useful area for validation of the NPS model as well as to study how nearby training data affects model accuracy. Figure 2.3 shows the NPS predictions (2.3a) and ensemble model predictions (2.3b). Training sites for the ensemble model are marked by small circles and colored according to measured sound levels. The scale on the difference map (Figure 2.3c) was extended from that used in Figure 2.1 to provide more detail. A histogram of residuals overlaid with estimated PDFs for 30 training sites used in the ensemble model from this area is shown in Figure 2.3d.

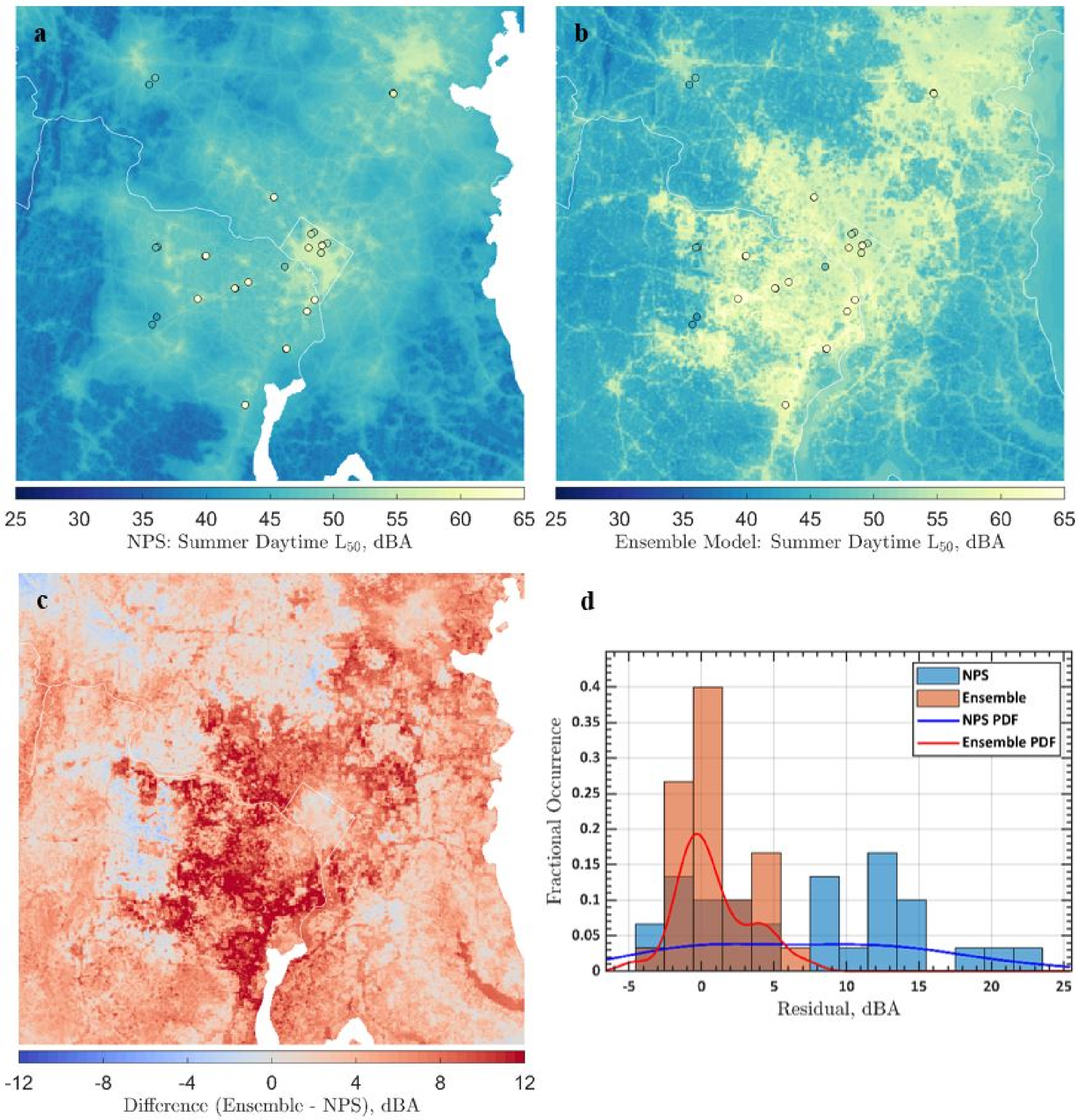

Figure 2.3 Summer daytime A-weighted L_{50} sound level predictions from the NPS model (a) and the ensemble model (b) in the Washington, D.C. area. Training sites for the ensemble model are marked by small circles and colored according to measured levels. The difference between the two map predictions (i.e., ensemble model predictions minus the NPS predictions) is shown in (c). A histogram and estimated PDFs of residuals at training sites included only in the ensemble model's training data set are shown in (d).

Table 2.2 summarizes the model errors on the 30 measured sites. Importantly, the residuals for the ensemble model are *fitting* errors (because the 30 measured sites were included in the ensemble model training data) while those of the NPS model are holdout *validation* errors (because the 30 measured sites were not included in the NPS training data). As such, residuals for the ensemble model are relatively low, centered around zero, and comparable to the LOO validation estimates. However, holdout errors for the NPS model are much larger than its reported LOO metrics. Indeed, it is clear from Figure 2.3 that the NPS map systematically underpredicts much of the D.C. area, predicting levels more than 20 dBA lower than measured values at some sites.

Table 2.2 Holdout validation RMSE and MAD for the NPS model and fit RMSE and MAD for the ensemble model in Washington, D.C. indicated by circles in Figure 2.3.

Model	RMSE, dBA	MAD, dBA
NPS Model	10.4	7.8
Ensemble Model	2.6	1.3

We attribute the large holdout errors in the NPS model to the limited (159 sites) urban data on which it was trained. Although the ensemble model only included training data from 169 urban sites; these included 30 sites from the D.C. area. Thus while much of the D.C. area is an extrapolation region for the NPS model, that is not the case for ensemble model. One would reasonably expect similar errors in the ensemble model for urban environments that are uncharacteristic of its training set. In any case, the contrast in performance of the NPS model between the Grand Teton National Park and D.C. area, together with the better agreement of the two models in Grand Teton National Park highlights how more representative samples in training data improve prediction accuracy.

2.6 Final holdout validation for both models

Lastly, we investigate the performance of both models on a holdout validation set of 25 sites. This differs from the analyses in Sections 2.4 and 2.5 because neither models' training data contained these sites. Twelve validation sites were near Los Angeles International Airport (LAX), five were near Long Beach Airport (LGB), three were close to rivers in Utah (UT), and five were in a rural region of Wyoming (WY). The training data for the ensemble model does not include sites near airports or close to rivers, so we expect many of these sites to be in the extrapolation regime of one or both models. Summary holdout validation errors for each region and all regions combined are reported in Table 2.3. The number in parentheses next to each region indicates the number of validation sites for that region.

Table 2.3 Holdout validation RMSE and MAD for the NPS model and the ensemble model at 25 validation sites (12 near LAX, 5 near LGB, 3 in UT, and 5 in WY) for the summer daytime A-weighted L_{50}.

	RMSE/MAD, dBA				
Model	All (25)	LAX (12)	LGB (5)	UT (3)	WY (5)
NPS Model	10.6/6.1	6.7/6.3	8.5/4.6	24.6/29.2	4.0/3.8
Ensemble Model	10.4/4.4	6.4/3.7	7.9/5.9	24.8/28.2	4.0/2.8

The RMSE and MAD holdout validation errors using all 25 sites are significantly greater than LOO errors for both models, with the exception of the ensemble model LOO MAD, which is relatively close to the ensemble model holdout MAD. Both models perform well at validation sites in Wyoming, which are likely similar to training data from natural environments. On the other hand, both models perform poorly at validation sites near rivers in Utah where validation errors are close to 30 dBA at some sites. This helps account for the difference between the holdout RMSE and MAD when using all 25 sites. It is interesting that the worst case error in this relatively small validation sample, is larger than the worst case discrepancy between the two models in all of CONUS. Indeed, the strong agreement between the two models over much of CONUS may be a result of shared training data (particularly within national parks), and not indicative of a higher confidence in predictions.

2.7 Conclusion

In this paper, we have reported a validation study of predictions for the summer daytime A-weighted L_{50} from the NPS published sound map [1, 24] and an ensemble model described in this paper. Both models were produced using supervised machine learning techniques; however, differences in the training data between the two models allowed us to assess how nearby training data affect model accuracy. A comparison of model predictions for the contiguous United States, validation studies in a natural environment (part of Grand Teton National Park) and an urban environment (the Washington, D.C. area), and a holdout validation on both models were performed. Results demonstrate that there are significant differences (greater than 20 dBA) between the two models' predictions and even larger errors at many holdout validation sites. These discrepancies are much larger than the leave-one-out statistical metrics reported for either model.

We attribute large errors to the limited acoustic data available for training that sparsely sample the space of geospatial features. In cases of large errors, we have argued that the validation environments are geospatially dissimilar to all training environments used to create the model. Indeed, both models were trained on data from fewer than 500 unique geographic sites; it is unlikely that these sites are representative of all geographic environments of interest within CONUS. Without incorporating any a priori physics, data-driven approaches are only reliable when they interpolate observed data. Yet, when models make predictions at these unrepresented sites they necessarily venture into unsampled, extrapolation regions, leading to inaccurate model predictions. This situation results in dramatically different predictions between the two models as well as large errors in predicted levels at many validation sites.

These results motivate several areas of future research. First, additional acoustic data, particularly at more diverse training sites, would improve the predictive performance of models. We have found that validation errors are significantly smaller when they are made at sites characteristic of training data. However, acoustic data are expensive to collect, requiring several days of recording for each site, so optimal data collection strategies should be used to help mitigate these costs. Second, there is a need for better uncertainty quantification of environmental sound level predictions, particularly in regions that are atypical of the training data. Uncertainty estimates of continental-scale environmental sound level predictions should be acknowledged and clearly communicated. Because the distribution of training data is non-representative of CONUS, we have seen that traditional statistical metrics, such as leave-one-out errors, are typically poor estimates of actual model error. Therefore, more sophisticated uncertainty quantification techniques are necessary to assess the reliability of a model for a particular application. At present, we caution against using any continental-scale sound level predictions or research dependent upon such predictions without a thorough consideration of the possible uncertainty in a region of interest.

Acknowledgements

This research was supported by a U.S. Army Small Business Innovation Research (SBIR) contract to Blue Ridge Research and Consulting, LLC.

Chapter 3

Uncertainty quantification of an ensemble of geospatial acoustic machine learning models in the limited-data regime*

Katrina Pedersen,[1] Mark K. Transtrum,[1] Kent L. Gee,[1] Brooks A. Butler,[1] Shane V. Lympany,[2] Michael M. James,[2] and Alexandria R. Salton[2]

[1)] Physics and Astronomy, Brigham Young University, Provo, UT, 84602

[2)] Blue Ridge Research and Consulting, LLC, Asheville, NC, 28801

3.1 Abstract

An ensemble of six machine learning models has been used previously to predict ambient sound levels throughout the contiguous United States [Pedersen et al., 2018]. The ensemble model was generated from 117 geospatial features and a training data set of 596 geographically unique sites. Various acoustic metrics, such as overall daytime L_{50} levels and one-third octave frequency band levels, were obtained at the training sites. Maps showing the ensemble model predictions and ensemble standard deviations, which are a surrogate for the structural uncertainty, highlight the advantages of an ensemble model as well as some challenges of machine learning in the limited-

*Chapter 3 is a manuscript to be submitted in the Journal of Acoustical Society of America.

data regime. Results motivate a need for accurate uncertainty quantification techniques for environmental sound level modeling. The statistical advantages and limitations of using an ensemble of machine learning models, particularly for limited data sets, are discussed.

3.2 Introduction

3.2.1 Environmental sound level modeling

An outdoor acoustic environment is a collection of all the ambient sounds in a geographic area. Modeling outdoor acoustic environments is a challenging problem because ambient sound is the accumulation of many effects, including diverse sources, barriers to propagation, etc., and can vary with time of day or season. However, accurate characterization of a region's acoustic environment has potentially broad applications.

The study of natural acoustic environments has been led by the National Park Service (NPS) through the Natural Sounds and Night Skies Division, charged with the protection and restoration of the national parks [45, 46]. The natural acoustic environment is an important part of a visitor's experience [13, 46, 47] that may be negatively impacted by unnatural contributions such as recreational motorized noise [48].

Accurate characterization of outdoor acoustic environments is also important for fields such as ecology and public health. Acoustic environments play an integral role in a species' habitat, particularly for animals that respond to sound such as birds [17, 49-51], marine life [52-56], and anurans (i.e., frogs and toads) [57]. Additionally, in areas dominated by anthropogenic sources, sound levels correlate with depression and anxiety [2], as well as hypertension [6-8]. An accurate environmental sound level model may also hold commercial applications for real estate and urban development and have implications for social justice [15, 58, 59]. However, accurate environmental sound level modeling for any location requires physical knowledge of the possible sound sources and barriers to sound propagation.

Remote sensing has greatly increased the amount of geospatial data available. These data may be useful in predicting characteristics of outdoor acoustic environments through methods such as land-use regression models [27] or machine learning [1, 22, 23]. However, there are various sources of uncertainty in the modeling process, including uncertainty in geospatial and acoustic measurements and uncertainty in the appropriate model structure or form. Geospatial and acoustic data will always have some error or uncertainty, which will propagate through the modeling process and affect the uncertainty in model predictions. Additionally, it is generally not clear what the relationship between individual geospatial layers and acoustic outputs should be, creating uncertainty in determining an appropriate model structure. Uncertainty in the model structure is magnified by the fact that the availability of high-quality acoustic data is low, putting outdoor environmental sound level modeling in the limited-data regime. Hence, although remote sensing has made environmental sound level modeling possible on larger geographic scales, caution should be taken to understand possible sources of model uncertainty.

3.2.2 Previous work

Previously, Mennitt et al. [1, 22, 23] used machine learning to relate geospatial features and acoustic metrics to produce continental-scale sound maps. Elsewhere, linear and nonlinear land-use regression models have been used to map urban environmental noise on smaller geographic scales [27]. Both of these efforts utilize geospatial data to predict information characterizing outdoor acoustic environments. As geospatial data have become more widely available and improved in accuracy and precision, many problems, such as environmental sound level modeling, have become more feasible. However, environmental sound level modeling on continental scales is still in its relatively early phases of development.

Much of the pioneering work in continental-scale environmental sound level modeling has been performed by Mennitt et al. [1, 22, 23] Their data set contains both acoustic and geospatial data. The acoustic data consists of acoustic measurements from 492 unique sites, 333 of which were located in quiet uninhabited areas within national parks and 159 of which came from urban areas. Note that acoustic data acquisition is a time consuming and expensive process. The acoustic data were collected by the NPS and airport noise monitoring systems [60] and they provide summary acoustic metrics, such as the summer daytime L_{50}, for all 492 training sites. In addition to acoustic data at specified training sites, data for 115 geospatial layers at a 270-m spatial resolution for the Contiguous United States (CONUS) were used [1].

Mennitt et al. applied random forest algorithms to the acoustic and geospatial data to predict various summary acoustic metrics, such as NN% time-exceeded levels (L_{NN} ; e.g., L_{50} and L_{90}) and the equivalent sound level (L_{eq}) [1, 22, 23]. Their methods for generating predictions of the summer daytime L_{50} are summarized here. Note that their predictions are publicly available [24]. All available L_{50} measurements from all seasons (995 measurements in total) were included in the training data set and model inputs were used to indicate the corresponding season of the measurement. After construction of the training data set, feature reduction was performed to obtain a reduced set of 45 features. The Leave-One-Out (LOO) cross-validation error was then calculated by leaving out all data for a given site, training a random forest model on the remaining data, and calculating the residual for the site that was removed. The reported LOO cross-validation Root-Mean-Square Error (RMSE) and Median Absolute Deviation (MAD) for the seasonal A-weighted L_{50} were 4.40 and 2.29 dB, respectively

[1]. The discrepancy between these two error metrics was attributed to outliers in the training data set. Additionally, the authors caution that sound levels at the quietest and loudest sites are over- and underestimated respectively, which is likely a result of how random forest models are constructed [1, 23]. Accuracy of predictions is also lower for extreme or unusual acoustic environments [1].

Traditional validation metrics, such as LOO, estimate model accuracy for inputs that are statistically similar to the training set. Compared to other validation methods (such as 10-fold or holdout validation), LOO cross-validation is better suited to limited data sets because it makes maximal use of the available data. LOO has a computational disadvantage in that it requires training the model many times (i.e., once per data point); however, when data are limited this is a reasonable trade-off. Additionally, for limited data sets, as is presently the case for environmental sound level modeling, each data point may contain unique information about the input/output relationship. Randomly omitting instances from the training set to use for testing results in a high probability of leaving out information that is important for a predictive model. In this case, the testing error will vary greatly depending on the random subset selected for training and testing.

Estimating expected predictive error becomes more challenging when the model is to make predictions on data that are statistically different from the training set. In this case, validation methods, including LOO cross-validation, will give overly optimistic error measures for sites that are geospatially (and therefore statistically) different from those in the training data set. The training data set used in previous work [1] consists of 492 unique sites, and it is unclear to what extent these sites are representative of the geospatial conditions throughout CONUS. Furthermore, with so few training sites, it is probable that the training data are insufficient to represent much of CONUS. Hence, the LOO cross-validation measures reported in [1] likely provide overly optimistic error estimates. Indeed, we saw this to be true in Chapter 2 of this dissertation. This motivates the need for a more quantitative estimate of the model transferability, i.e., ability of a model to accurately make predictions for novel inputs that are statistically different from the training set. Transferability estimation is an important, open problem in machine learning [61]. In this paper, we approach the problem of estimating transferability of machine-learned environmental sound level models using an ensemble of different learning algorithms.

3.2.3 Paper overview

In this study, a single function was chosen from each of six classes of machine learning models. Models from each model class were trained to predict ambient sound levels throughout CONUS. The best model from each class was selected (as measured by the LOO cross-validation error) and their acoustic predictions on new spatial, temporal, and frequency domains were considered. Results presented in this paper are limited to varying spatial domains for the summer daytime L_{50}.

The range of predictions from the ensemble of models serves as a surrogate for the accuracy of a single model. Because our ensemble consists of the single best model from each machine learning class, we interpret this range as a measure of the structural uncertainty, i.e., the uncertainty due exclusively to differences in functional forms. Predictions for which the range of ensemble predictions is large correspond to greater structural uncertainty. The ensemble not only provides a method of quantifying structural uncertainty, but also aids in directing further data collection and management efforts.

A qualitative analysis of ensemble model predictions and structural uncertainty for CONUS, Asheville (North Carolina), and Iowa manifests both the importance of uncertainty quantification estimates and advantages of an ensemble model in different types of geospatial regions. These analyses demonstrate that continent-scale environmental sound level models are currently in the limited-data regime and, as such, have limited transferability. Hence, as demonstrated here, it is critical that sound level predictions be accompanied with careful uncertainty estimates that go beyond the traditional LOO metric. One substantial method to predict structural uncertainty is documented below.

3.3 Methods

3.3.1 Data sets

We developed an ensemble of machine learning models using a database of geospatial and acoustic data points. The geospatial database contains CONUS layers for 117 geospatial features (see Table A.1 in Appendix A; removing AviationNoise, PopDensity, and RoadNoise from the list of 120 geospatial features) from the NPS Natural Sounds and Night Skies and Inventory and Monitoring Divisions database [62, 63]. All 117 geospatial features were used in this paper. The geospatial features can be classified into six categories: topography, climate, land cover, hydrology, anthropogenic, and position. Many of these layers are the same as those considered or used by Mennitt et al. Our modeling efforts are ongoing, both because the theory of data-driven predictive modeling with limited data is immature and acoustic data are continuously collected. Hence, the modeling results presented here are subject to future refinements.

The acoustic database contains measurements from 596 distinct training sites, which are compiled from multiple sources including Blue Ridge Research and Consulting, LLC’s internal acoustic data, the NPS database [62, 63] , and a 1974 Environmental Protection Agency study

[64] . (Note that the number of training sites differs from the number reported previously in [65] due to ambiguities in counting the same site at various times.) The acoustic measurements from each site are summarized on a seasonal basis using several statistical measures including the L_{50}, L_{90}, and L_{eq}. Note, measurement data during all four seasons are not available at every site. Therefore, model results presented here are for summer acoustic metrics, as the largest number of acoustic training sites (496 for time-exceeded levels) contain summer data. Hence, training data were only composed of summer data. Data from other seasons may be utilized in future work.

The raw acoustic measurement data are the result of considerable effort in terms of number of recorded hours. However, the summary statistics that compose the acoustic database are actually a very limited data set. Our training set consists of only a few hundred instances, while many of the more dramatic successes of machine learning are known to require millions of training instances [28, 29]. Although the formalism of machine learning can be applied to our data set, it is unclear whether the predictions will be accurate away from training sites. In this limited-data regime, statistical validation measures (such as LOO) do not reflect the actual confidence we have in model predictions, and more sophisticated model-driven, physics-guided uncertainty quantification techniques are needed.

3.3.2 Machine learning overview

We implemented a computational pipeline to facilitate the machine learning process [65]. This pipeline enables us to explore different validation metrics and a wide variety of machine learning models. We explored six machine learning model classes: gradient boosted regression trees, neural networks, k-nearest neighbors, support vector machines, kernel ridge regression, and Gaussian process regression. To compare initial model performance, LOO cross-validation was used to calculate the RMSE and MAD for each model. Hyperparameters were optimized for each model class using a grid search to minimize the LOO MAD cross-validation error.

A comparison of the prediction errors from the six model classes showed that the residuals given by the LOO cross-validation errors are non-Gaussian and the MAD is typically about half the value of the RMSE [65]. The difference between the MAD and RMSE is explained by the existence of large outliers in LOO cross-validation residuals. Similar results are reported by Mennitt et al. [1, 23] Because the MAD and RMSE of the LOO cross-validation errors are comparable between model classes and the difference in errors is statistically indistinguishable [65], each of the model classes gives an adequate fit to the available training data.

Using an ensemble of the models, defined as the median predicted value of the six optimized models (one from each model class), reduces the model's sensitivity to outliers. Additionally, the ensemble model predictions appear more physically accurate in the extrapolation regimes. However, without any way of estimating model transferability in these extrapolation regimes, prediction accuracy is undetermined.

3.3.3 Uncertainty quantification

A central question in machine learning is the problem of model validation. (We direct the reader *to Evaluating learning algorithms: A classification perspective* [42] for further explanation of validation methods in machine learning than what is provided in this paper.) Traditional validation metrics estimate model accuracy for inputs that are statistically similar to the training set. Most often, validation is performed on a subset of the available data using methods such as LOO or k-fold cross-validation, or holdout validation. A broader question, and our primary focus, is that of transferability, or the ability of a model to accurately make predictions for novel inputs that are statistically different from the training set.

Our approach to estimating model transferability, or predictive error, was inspired by techniques from the uncertainty quantification community. Uncertainty quantification has existed as long as probability and statistics and is the science of identifying, quantifying, and reducing uncertainty when predicting quantities of interest [31]. More recently, an interdisciplinary community has emerged to systematize the study of issues related to uncertainty, with particular relevance for estimating transferability in machine learning. In the broadest sense, there are two main types of uncertainty: aleatoric and epistemic uncertainty [31]. Aleatoric uncertainty, or statistical uncertainty, is inherent to a problem, and hence cannot be reduced and is generally represented in terms of probabilities [31]. For environmental sound level modeling, many sources of aleatoric uncertainty are associated with measuring the geospatial or acoustic data that make up the training and input data sets. Epistemic, or systematic, uncertainty originates from an incomplete knowledge or missing physics in a model and can be reduced through better modeling methods [31].

Following Kennedy et al. [32], uncertainty types may be further refined into six classes: parameter uncertainty, model inadequacy, residual variability, parametric variability, observation error, and code uncertainty. Of these six sources, this study is primarily interested in model inadequacy. Model inadequacy, or structural uncertainty, originates from uncertainty in the form of the model due to limited knowledge of the true underlying mechanisms that generate the data [32]. We anticipate that structural uncertainty, as measured by the standard deviation in ensemble model predictions in this work, is one of the primary sources of uncertainty for geospatial sounds level modeling.

Because we use a data-driven, machine learning approach, our model necessarily omits potentially relevant physics. In our case, this is not just a practical convenience; indeed, many physical principles relevant to ambient sound levels are unknown. Although machine learning methods are applicable to describing complex behaviors in which the underlying physical principles are unknown, success often hinges on the availability of large data sets on which to train the model. In the limited-data regime, a large portion of the uncertainty in model predictions is likely due to epistemic, rather than aleatoric, uncertainty. Furthermore, assessing the accuracy, precision, and transferability of the learned models is often not straightforward.

To make these ideas more concrete, consider the following problem formulation. Training data comes from a "true" (but unknown) generating function. Ideally, one would like to learn this generating function out of a candidate set, but in practice there are many different functions consistent with the available measurements. The set of functions consistent with the training data (i.e., the set of functions that fit the data equally well) form an equivalence class. Although each of these functions is consistent with the training data, they may disagree in their predictions under novel conditions. We lack confidence in our model's predictions because we do not know which model from the equivalence class is correct. Therefore, our goal is to characterize the range of predictions in this equivalence class of functions.

A benefit of using an ensemble model is that it provides some measure of uncertainty when the individual models must extrapolate outside their training regime. Although individual model predictions may look reasonable, they generally provide no measure of confidence to their predictions for input values that are statistically different from the training set. In contrast, the range of predictions within the ensemble provides an estimate of the expected uncertainty.

The standard deviation of the ensemble model's predictions is used here as a measure of uncertainty since the variation among ensemble members is a surrogate estimate of the accuracy. Because our ensemble was generated from the single optimal model in each class, this uncertainty estimate quantifies the variability due to the functional form of the model, i.e., the structural uncertainty. The confidence intervals provided by the standard deviation of ensemble predictions are similar to those generated by a Gaussian process regression model, which also uses an ensemble of models, or functions.

3.4 Results

We analyze ensemble model behavior for three different areas of analysis; namely, CONUS, Iowa, and the Asheville, North Carolina area. These studies focus primarily on understanding how geospatial layers affect the ensemble model. Results highlight advantages and limitations of an ensemble model in the limited-data regime and help identify possible ways of improving the model.

3.4.1 CONUS

Figure 3.1 shows ensemble model predictions, i.e., the median predicted levels from the six machine learning models, of the summer daytime L_{50} for CONUS. Figure 3.2 shows an estimate of the uncertainty, i.e., the standard deviation, associated with these predictions. Training sites are indicated by small circles, which are colored according to measured levels in Figure 3.1 and colored black in Figure 3.2. The mean and median uncertainty across CONUS are approximately 3.54 and 2.86 dBA, respectively. Only slightly more than 6.2% of sites have an uncertainty greater than 6 dBA, saturating the limits in Figure 3.2. However, the mean uncertainty of these sites is about 12.6 dBA.

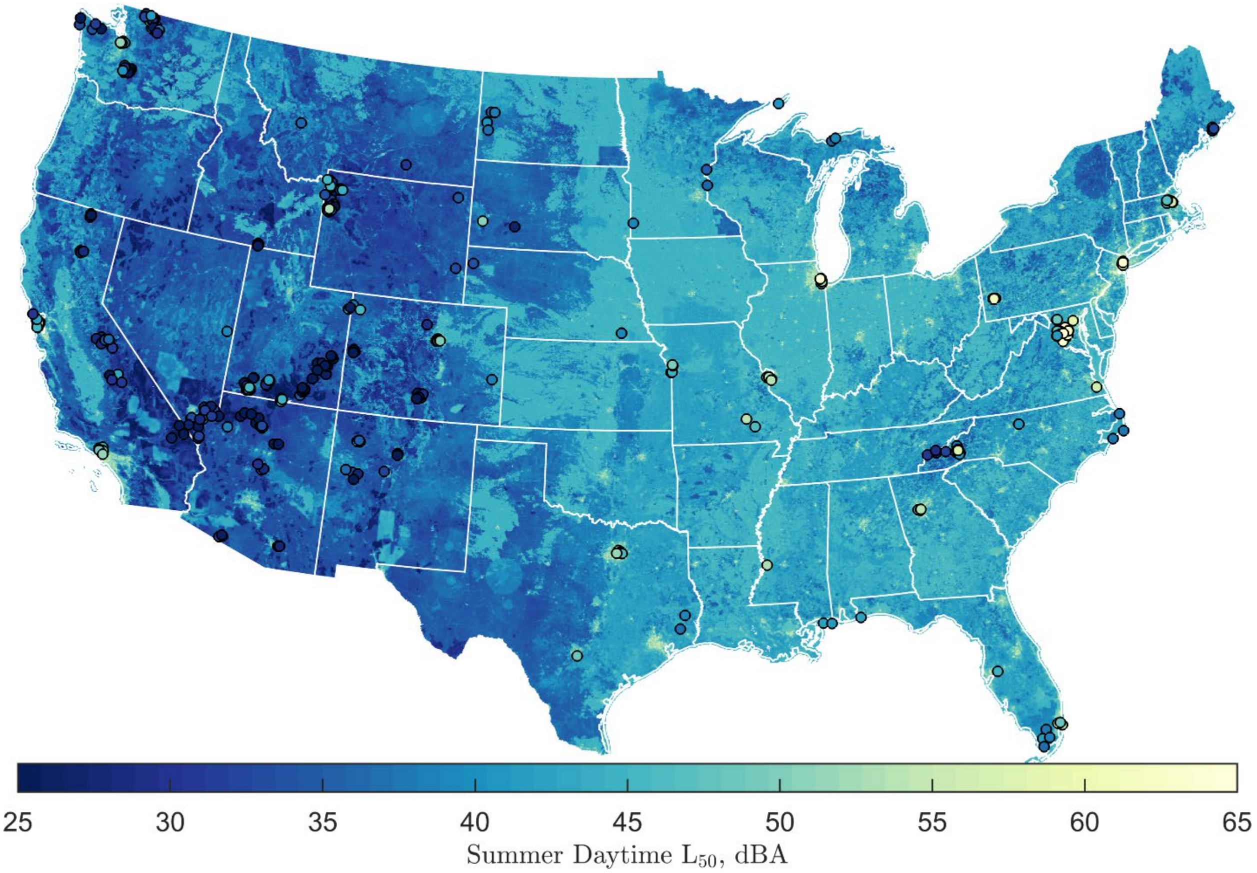

Figure 3.1 Ensemble model predictions for the A-weighted summer daytime L_{50} for CONUS.

It is important to recognize that the uncertainty estimates shown in Figure 3.2 are the standard deviation in the predicted dBA values, but they do not represent confidence intervals for model predictions because we are only quantifying one source of uncertainty, i.e., model inadequacy or structural uncertainty. In other words, relatively large uncertainty values indicate greater uncertainty in model predictions, but the actual level of uncertainty is still unclear. In part, this is due to the fact that the structural uncertainty is primarily an estimate of the epistemic uncertainty, and gives no indication of the aleatoric uncertainty. Additionally, although we anticipate the structural uncertainty comprises a large portion of the total epistemic uncertainty, there are other sources of epistemic uncertainty not considered here. For example, the geospatial database may be insufficient for predicting acoustic environments for some parts of CONUS. Hence, the standard deviation indicates a minimum level of confidence in model predictions, but does not provide rigorous confidence interval measures. These uncertainty estimates do strongly encourage caution when applying sound level model predictions for further studies, such as measuring correlations between high sound levels and ecological trends. More specifically, studies utilizing outdoor environmental sound level model predictions should be especially careful when looking at trends that rely on differences in sound levels of only a few dBA.

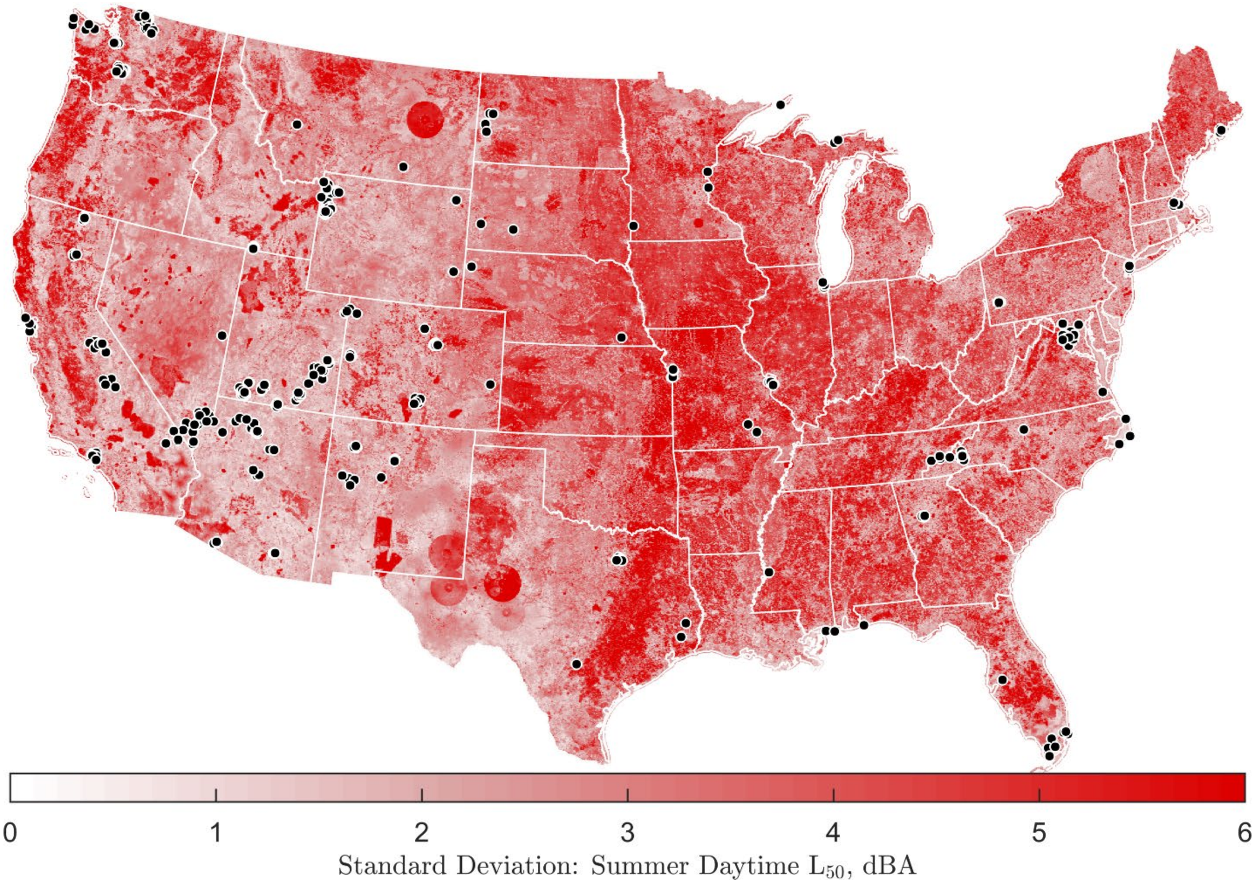

Figure 3.2 Estimated structural uncertainty, as measured by the standard deviation of ensemble model predictions, for the A-weighted summer daytime L_{50} for CONUS. Color bar limits were truncated at 6 dBA.

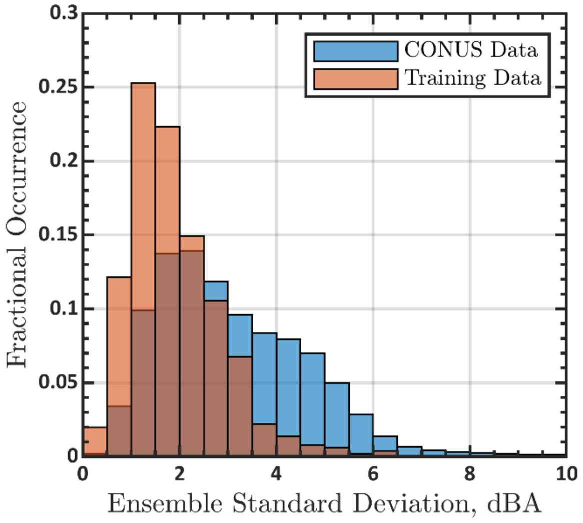

Figure 3.3 Histograms of the standard deviation of ensemble model predictions for CONUS data and the training data (when performing a LOO analysis) for the A-weighted summer daytime L_{50}.

Some areas of high structural uncertainty in Figure 3.2 can be associated with specific geospatial features. For example, the large circular regions in northwestern Texas and eastern Montana are due to the VIIRS layers, which measure the upward radiance at night, at the 69,120-m spatial resolution. Additionally, the odd rectangular shape in southern New Mexico and areas of high uncertainty in southeastern California, southern Nevada (northwest of Las Vegas), and southeastern Idaho (west of Rexburg) all correlate with several of the land use layers, including the institutional layers. Relatively few training sites are from areas of large institutional land use, which could account for the large uncertainties in these regions. Potential errors in the geospatial layers or acoustic data may also contribute to large uncertainties in model predictions.

Comparing the distribution of structural uncertainty for CONUS to that of the training sites when using LOO cross-validation can provide further insight into the uncertainty in model predictions. Figure 3.3 shows histograms of the ensemble standard deviation for the predictions at the training sites (when they are removed from the training data set) and CONUS data. Notice that the mean structural uncertainty is clearly lower for the training sites than for all of CONUS. This is because the training sites are not drawn from the same geospatial distribution as the CONUS data. Hence, sites that are geospatially different from those in the training data set are likely the greatest source of high structural uncertainty. Note that the ability to estimate the structural uncertainty for model predictions is a benefit to ensemble modeling; indeed, it is necessary to investigate the uncertainty in model predictions.

A qualitative analysis of Figure 3.1 indicates that sites on the eastern half of the map are generally louder than those on the western half. Because machine learning methods are data-driven and therefore search for and utilize patterns and correlations in the training data, it is possible that this is in part due to a biased training data set. More specifically, many of the training sites in national parks are located in western CONUS while many urban training sites are in eastern CONUS. Although eastern CONUS is more heavily populated on average, it is possible that predictions for much of eastern CONUS are higher than they otherwise would be due to correlations in the training data set between geospatial features and sound levels that are a result of working with limited data. Indeed, the map of standard deviations in ensemble model predictions (see Figure 3.2) indicates greater structural uncertainty for much of eastern CONUS.

3.4.2 Iowa

Although CONUS maps may be useful for looking at large-scale spatial patterns in model behavior, more local studies, particularly of the structural uncertainty, provide additional insight into understanding model performance. Figure 3.4 shows the structural uncertainty for the state of Iowa. Note that the color bar was not truncated here and the coloring up to 6 dBA is consistent with Figure 3.2. Additionally, note that no training sites are in Iowa and Iowa is geospatially dissimilar from the training data. In general, uncertainty in model predictions is

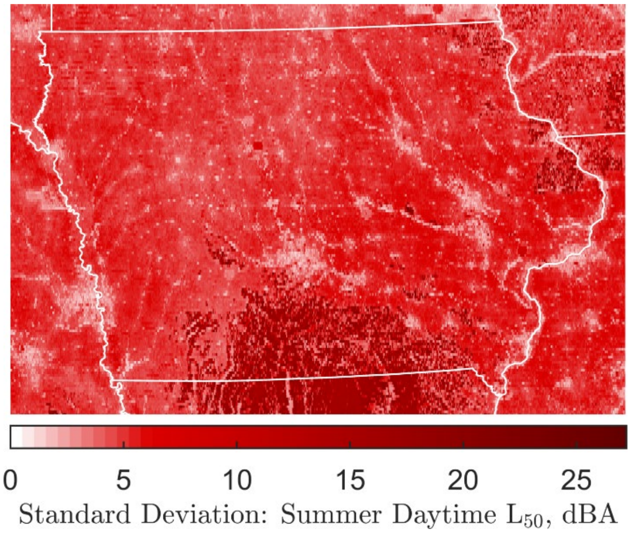

Figure 3.4 Estimated structural uncertainty, as measured by the standard deviation of ensemble model predictions, for the A-weighted summer daytime L_{50} for Iowa. The color bar was not truncated and the coloring up to 6 dBA is consistent with Figure 3.2.

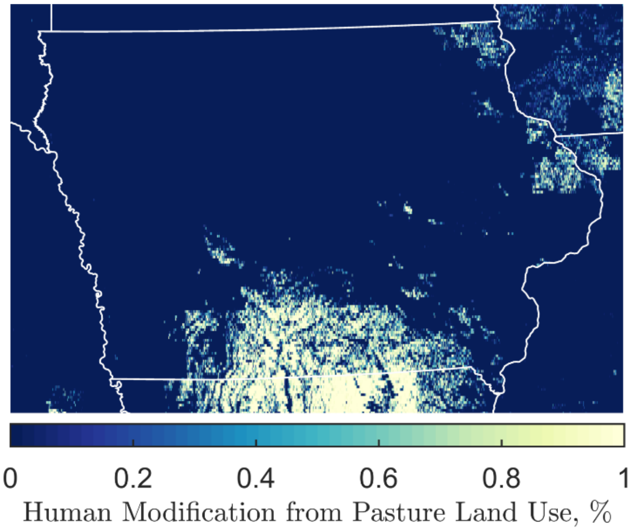

Figure 3.5 The degree of human modification from pasture land use (as a percentage) for Iowa. This geospatial feature is one of the inputs to the ensemble model.

relatively high (greater than 4 dBA) in Iowa, but tends to be low along rivers and in developed areas. The areas of highest uncertainty correlate most strongly with the pasture land use layer (with a 200 m area of analysis), which is shown in Figure 3.5. There is only one training site with any pasture land use component at the 200 m area of analysis, so it is unsurprising that the uncertainty is high for areas with a large amount of pasture land use. This suggests that areas with high pasture land use may be good candidates for future acoustic data collection. More generally, areas of high structural uncertainty are often good candidates for future data collection.

Although the pasture land use layer can account for the areas of largest uncertainty in Iowa, it does not explain why Iowa in general has large structural uncertainty compared to other areas in CONUS. The cropland land use geospatial layers may be the cause of this since these layers have relatively large values for much of Iowa and are poorly represented in the training data set. In particular, only seven training sites, or less than 2% of training data, have any cropland land use at the 200 m area of analysis, and only two of these sites have more than 20% cropland land use. Additionally, only fourteen sites have more than 10% cropland land use at the 5000 m area of analysis. Hence, areas of high cropland land use may also be good candidates for future acoustic data collection.

Although it is possible to identify the geospatial features correlated with large uncertainties, targeted acoustic data collection based on these geospatial features may not be the only method for reducing uncertainty. Future work will include physics-guided feature reduction to identify and remove irrelevant features that are under-sampled in the training set (and thus lead to high uncertainties), but that are also unlikely to improve predictive performance. For example, large uncertainties in Figure 3.2 are associated with the VIIRS layers for the 69,120 m area of analysis. These layers are unlikely to provide useful information to the model because they have unphysical (annular) shapes and are averaged over large spatial areas. So, we will likely remove these layers in future work. For the case of the VIIRS layers (at the 69,120 m area of analysis), it is unlikely they would correlate with sound levels even if sufficient training data were present. Outside of the limited-data regime, models would be less likely to identify such features as having an effect on environmental sound level predictions.

On the other hand, when uncertainty is associated with geospatial features that likely have a causal relationship to the acoustic environment (or are correlated in useful ways), we plan to use targeted acoustic data collection. More specifically, geospatial features that are poorly represented in the training data set but may have a physical or useful relationship with the acoustic environment, such as pasture and cropland land use, can direct future acoustic data collection. Because few training sites have cropland or pasture land use, we plan to take data at sites that have significant cropland or pasture land use. Ideally, the geospatial features in the training data set would be statistically similar to those for which predictions are made (e.g., CONUS). Uncertainty in ensemble model predictions can help identify areas where they are not.

3.4.3 Asheville, North Carolina

Unlike Iowa, for which we have no training sites, Asheville, North Carolina is a relatively well-sampled geographic area. Figure 3.6 shows the structural uncertainty for the Asheville area. Black circles indicate the location of training sites. Note that structural uncertainty is fairly low near Asheville, with a few exceptions. Many of the areas with higher uncertainty in Asheville, such as the Biltmore Estate in the middle of the map, are correlated with the timber land use layers. Timber land use is another geospatial layer that is poorly represented in the training data, so it is unsurprising that it correlates with areas of higher structural uncertainty. Additionally, the area of high uncertainty along the road (Interstate 26) in southern Asheville corresponds to the Asheville Regional Airport. Since the airport is relatively small, and the training data set does not contain many sites near airports, it is unsurprising that the model has higher uncertainty there.

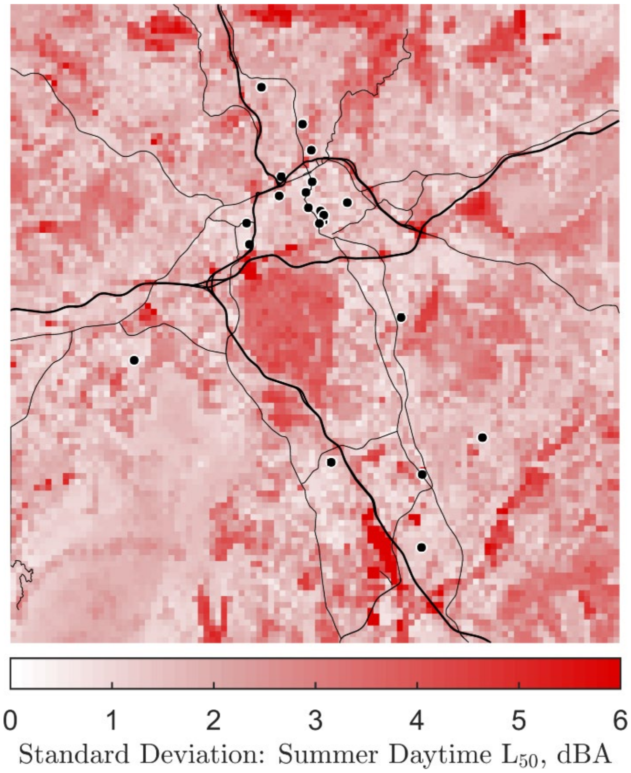

Figure 3.6 Estimated structural uncertainty, as measured by the standard deviation of ensemble model predictions, for the A-weighted summer daytime L_{50} for Asheville, North Carolina.

Despite these areas of higher uncertainty, ensemble model uncertainty is relatively low for much of Asheville. Recall that this does not necessarily indicate that model accuracy is high. However, the low uncertainty in the Asheville area suggests that machine learning methods may be successfully applied to the environmental sound level modeling problem when enough training data are present.

3.5 Conclusion and future work

Continental-scale environmental sound level modeling is challenging, but holds many potential applications in areas such as ecology and public health. However, environmental sound level modeling is also in its relatively early phases of development, so it is unclear how best to approach this problem. We used a computational pipeline to facilitate the development of models from six different machine learning algorithms: gradient boosted regression trees, neural networks, k-nearest neighbors, support vector machines, kernel ridge regression, and Gaussian process regression. One model from each of the six model classes was combined into an ensemble of machine learning models that made predictions of ambient sound levels and corresponding uncertainty estimates for sites within the contiguous United States.

From the uncertainty estimates, which we interpret as measures of the structural uncertainty (i.e., the uncertainty due exclusively to differences in functional forms), we found that the mean and median uncertainty across CONUS are approximately 3.54 and 2.86 dBA, respectively. Note that these uncertainties indicate a minimum level of confidence and low uncertainties do not guarantee high accuracy. For these reasons, the use of outdoor environmental sound level model predictions should always be in conjunction with prediction uncertainties. Additionally, our results indicate that conclusions drawn from continental-scale models that rely on differences of 3 dBA or less are unjustified. Applying sound level model predictions to ecological, public health, noise pollution, and other studies should be done carefully and take into account uncertainty estimates. Users of sound level model predictions should evaluate whether the geospatial inputs and acoustic training data are appropriate for their desired application. For example, an environmental sound level model trained only on National Park Service data is likely not appropriate for use in urban environments, even if structural uncertainty estimates are low.

Beyond estimating model accuracy, an ensemble model may help guide feature reduction strategies. Environmental sound level models that are produced using machine learning methods identify correlated, but not necessarily causal, features. Many of these correlations may only be present due to limited training data. Geospatial features that correlate with areas of high uncertainty may be good candidates for removal if they have a non-causal relationship with the acoustic environment and are unlikely to correlate in useful ways given sufficient training data. It is likely advantageous to remove such features to aid the model in identifying more causal relationships. However, it may be reasonable to keep a non-causal feature if there is no better feature available to represent or correlate with certain contributions to the acoustic environment. Note that removing geospatial features will likely decrease the estimated structural uncertainty in some areas while increasing it in others. In general, however, we anticipate that removing features will cause a decrease in the structural uncertainty because a smaller number of features will produce a smaller space of possible model structures (assuming there is not a significant change in model hyperparameters). Hence, we advise against comparing levels of measured uncertainty for feature data sets of different sizes.

In addition to aiding feature reduction strategies, uncertainty estimates can also guide future acoustic data collection efforts. Presumably many sites with large uncertainties have geospatial features that are underrepresented in the training set. Consequently, uncertainty measures can help identify ways in which training data are not statistically representative of areas for which predictions are made. Targeted acoustic data collection to improve the statistical similarity between the training data and data for predictions will likely aid the ensemble model in learning relationships between geospatial features and the acoustic environment. Hence, we anticipate that targeted acoustic data collection will not only reduce ensemble uncertainty at sites for which data are collected, but also for geospatially similar sites. Since acoustic data collection is an expensive and time-consuming task, it is extremely advantageous to use targeted, rather than random, data collection techniques.

This study motivates several future research directions for improving the uncertainty quantification of outdoor environmental sound level predictions. Our ensemble model quantified structural uncertainty, i.e., uncertainty due to the functional form of the model. However, more extensive ensembles could be generated to account for other sources of uncertainty or further evaluation of structural uncertainty. Recall that the geospatial and acoustic data are not free from error and may contribute significantly to the uncertainty in model predictions. In this regard, the uncertainty estimates reported here are conservative.

Environmental sound level modeling is an important acoustical question with potentially broad applications. However, the limitations of the available data demand sophisticated uncertainty quantification tools in order to assess model transferability and improve predictive performance for sites dissimilar from the training set. Uncertainty quantification methods will play an important role in guiding future feature reduction, data collection and management, and model selection.

Acknowledgements

This research was supported by a U.S. Army Small Business Innovation Research (SBIR) contract to Blue Ridge Research and Consulting, LLC. We thank Dr. James Stephenson, technical monitor for the SBIR, for his thoughtful comments and feedback on this manuscript.

Chapter 4

Environmental sound level modeling of summer hourly spectra

4.1 Introduction

In Chapters 2 and 3, supervised machine learning models were trained for the prediction of the summer daytime A-weighted L_{50} (i.e., the summer daytime median A-weighted sound pressure level). In this chapter, we present results of training models to predict summer hourly spectra. We use model inputs to specify the hour and frequency band for a given training instance. Large ensembles of gradient boosted regression trees and k-nearest neighbors models are constructed using bootstrap subsets of the training data and an analysis of leave-one-out predictions for three sites is presented.

4.1.1 Previous work

Previously, Mennitt et al. [23] trained individual models for the prediction of L_{10}, L_{50}, and L_{90} one-third octave band daytime levels for any day during the year. In other words, different models were trained for the prediction of each one-third octave band (e.g., separate models were trained for the prediction of L_{10} 200 Hz one-third octave band daytime levels and L_{10} 250 Hz one third-octave band daytime levels). We remind the reader that the L_{10}, L_{50}, and L_{90} are statistical time-exceeded levels corresponding to the sound pressure level exceeded 10 percent of the time, 50 percent of the time, and 90 percent of the time, respectively. Model inputs were used to specify the day of the year corresponding to measured/predicted levels. We are unaware of any other data-driven models for the purpose of predicting both temporal and spectral variation across geographic regions.

4.1.2 Motivation

In contrast to environmental sound level predictions from Chapters 2 and 3, summer hourly spectra predictions provide a greater degree of temporal and spectral resolution. Applications of models trained to predict summer hourly spectral levels are similar to those for the summer daytime A-weighted L_{50}. However, spectral and hourly information provides additional insight into which acoustic sources and/or propagation effects dominate within an acoustic environment. For example, large levels of high-frequency noise may often be caused by bird or insect activity, particularly in the evening and early morning.

Without spectral and hourly information, it is difficult to conclude much about the dominant contributions to overall sound pressure levels. An analysis of environmental sound models for the prediction of hourly spectra may therefore provide insight into which acoustic sources/mechanisms are well understood by models and which are not. Targeted acoustic data collection may then be used to gather data which characterize sources/mechanisms which are poorly understood by models.

4.1.3 Chapter outline

In Section 2 of this chapter, we describe the data set and modeling methods used to train environmental sound level models of summer hourly spectra. Section 3 describes summary statistical measures of model performance and presents an analysis of the leave-one-out predictions for three training sites. Leave-one-out predictions are realistic and we observe that models predict the appropriate spectral shape at some sites, despite differences in measured and predicted levels. Models struggle to identify locations at which bird/insect contribute to environmental sound levels. However, it is possible geospatial data do not contain sufficient information to determine locations of bird/insect activity. Despite these challenges, results indicate supervised machine learning methods may be successful if sufficient training data are obtained. Section 4 summarizes results and suggests future work.

4.2 Methods

4.2.1 Data sets

We use the set of scaled 51 geospatial features described in Appendix A. Additional model inputs specify the frequency band and time corresponding to a given instance. For example, f denotes the frequency bands and takes on values from 6.3 Hz to 20,000 Hz, which we transform using the logarithm (base 10). Because temporal variations are cyclic, they are treated differently. If h denotes the daily hour, it takes on integer values from 0 to 23. Ideally, however, hours 0 and 23 should be near one another. We therefore use $h_x = \sin\left(\frac{2\pi h}{24}\right)$ and $h_y = \cos\left(\frac{2\pi h}{24}\right)$ as two inputs to the model, rather than the hour h. We then augment the model to take (f, h_x, h_y) as inputs in addition to the geospatial layers described in previous chapters. All three of these inputs are scaled using a min-max scaler, which scales data to between zero and one and preserves the shape of the data distribution. Unlike models trained by Mennitt et al. [23] for daytime spectra, we do not train separate models for different one-third octave frequency bands. This allows models to potentially make use of correlations between different frequency bands.

Acoustic training data are composed of measured summer hourly L_{50} one-third octave bands (6.3 Hz-20 kHz) from 341 geographically unique sites. We manually cleaned acoustic data to remove data contaminated by wind noise and/or noise floor effects. Figure 4.1 shows an example of data that were remove due to wind noise contamination at lower frequencies and instrumentation noise floor at higher frequencies. Data from 55 sites were completely removed from the training data while data from many of the 341 remaining sites were partially removed. Therefore, different sites may contain data at different one-third octave frequency bands. Training data are composed of approximately 236,820 instances with each training instance corresponding to a specific location, hour, and frequency band.

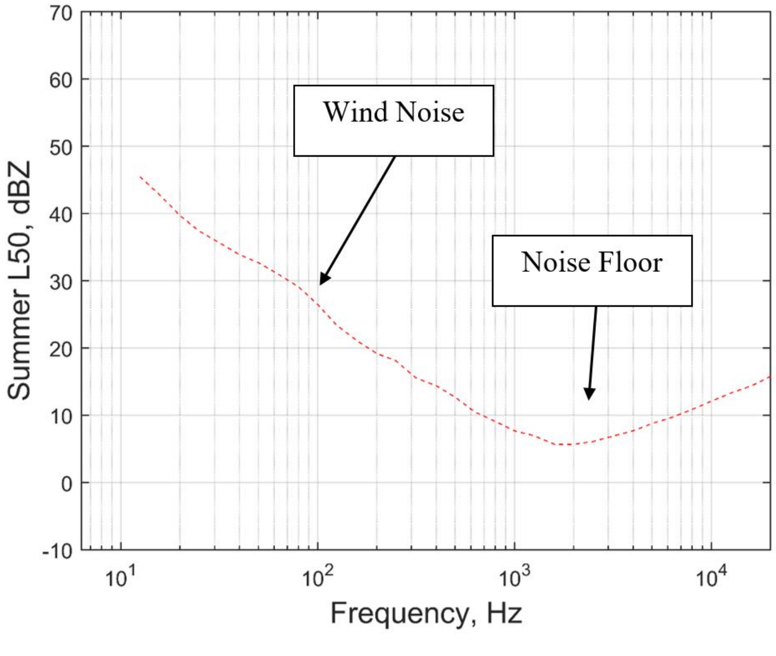

Figure 4.1 Example spectrum characterized by wind noise at low frequencies and instrumentation noise floor at higher frequencies. Data from this site were removed from the training data.

4.2.2 Ensemble models

One advantage of ensemble models is that they provide a range of predicted values from which uncertainty in model predictions can be estimated and from which the true value likely lies. In this chapter, we make use of ensemble models to predict summer hourly L_{50} spectra. Unlike the ensemble models described in Chapters 2 and 3, we use ensembles trained on 100 different bootstrap samples for two different model classes: Gradient Boosted Regression trees (GBR) and K-nearest Neighbors (KN). In contrast to ensembles composed of models from different model classes, which identify representational, or structural, uncertainty in model predictions, these bootstrap ensembles quantify uncertainty arising from computational and statistical uncertainties. We only use two of the six model classes used in the ensemble models from Chapters 2 and 3 due to computational limits and time constraints. GBR ensemble members were each trained for approximately six days on 24 cores using 64 GB of RAM, and KN ensemble members were each trained for approximately six hours on 24 cores using 32 GB of RAM. Training ensemble members from the other model classes using hyperparameter search techniques similar to those used for the ensemble of six models and the larger GBR and KN ensembles described here would take significantly longer (i.e., multiple months) or have higher memory requirements (i.e., greater than 128 GB RAM).

Despite only using two model classes, the range of predictions between the GBR and KN ensembles will still provide some indication of the structural uncertainty of model predictions. Note that model inputs determining frequency and hour were multiplied by a factor of 20 for the KN ensemble to help models identify dominant contributions for the prediction multiple acoustic metrics (i.e., multiple hours and one-third octave bands).

Hyperparameters for all models are tuned to minimize the Leave-One-Site-Out (LOSO) cross-validation Median Absolute Deviation (MAD) using the tree-structure Parzen estimator approach implemented in the Python library hyperopt [43, 44]. Note that LOSO predictions are calculated by leaving out all data for given a training site, training a model on the remaining data, and making predictions for all instances corresponding to the given training site. Ensemble LOSO predictions are made in a similar manner. Because ensemble members are trained on bootstrap subsets of the data however, predictions of some ensemble members will not change when a given training site is removed. More specifically, ensemble members whose training data did not originally contain the site being removed will not change.

4.3 Results

4.3.1 Summary error metrics

Fit and LOSO Root-Mean-Square Error (RMSE) and MAD for the GBR and KN ensembles are reported in Table 4.1. Although the KN ensemble has lower fit errors and slightly higher LOSO errors, the two ensemble models both perform reasonably on training data. To better understand what the models have learned, we look at LOO predictions at training sites.

Table 4.1 Fit and LOSO cross-validation RMSE and MAD for the GBR and KN ensemble models for the prediction of summer hourly L_{50} spectra.

Model	Fit RMSE/MAD, dBZ	LOSO RMSE/MAD, dBZ
GBR Ensemble Model	2.4/0.6	6.6/3.5
KN Ensemble Model	1.1/0.0	7.0/3.7

4.3.2 LOSO predictions: traffic noise

Figure 4.2 shows measured levels (dashed lines) and LOSO ensemble model predictions (solid lines) for the GBR (left) and KN (right) ensembles in 6-hour periods for a training site in Alexandria, Virginia near Washington D.C. Metadata indicate that traffic noise is the dominant acoustic source for this site. Levels are colored according to their corresponding hours. Going from bottom to top, the 6-hour periods are midnight to 6 a.m. (bottom), 6 a.m. to noon, noon to 6 p.m., and 6 p.m. to midnight (top). Note that ensemble predictions are median predicted values.

Although models do not predict peak levels of traffic noise, LOSO predictions are close to measured levels and spectral shapes are in fairly good agreement. Indeed, LOSO predictions still identify two peaks—one near 63 Hz and a slightly lower peak near 1 kHz—which have previously been observed as characteristic of traffic noise in cities [66]. LOSO predictions show that models predict higher, but more consistent levels during daytime hours and quieter levels at night. GBR and KN ensemble LOSO predictions are similar, suggesting structural uncertainty may be low for this site. The training data set contains similar data near roads in and around Washington, D.C., so it would not be surprising if structural uncertainty is low here.

It is difficult to visualize hourly spectra predictions and the spread of ensemble model predictions simultaneously. However, the spread of ensemble member predictions may provide insights into model learning. Figure 4.3 shows the standard deviation in fit (top) and LOSO (bottom) predictions as a function of hour (horizontal-axis) and frequency (vertical-axis) for the GBR (left) and KN (right) ensembles. Interestingly, uncertainty in both fit and LOSO predictions is largest for high frequencies. This may be due to differing instrumentation noise floors at different training sites. Although data were cleaned to minimize contamination due to wind noise and noise floor effects, these effects are still present in the training data to limited degrees.

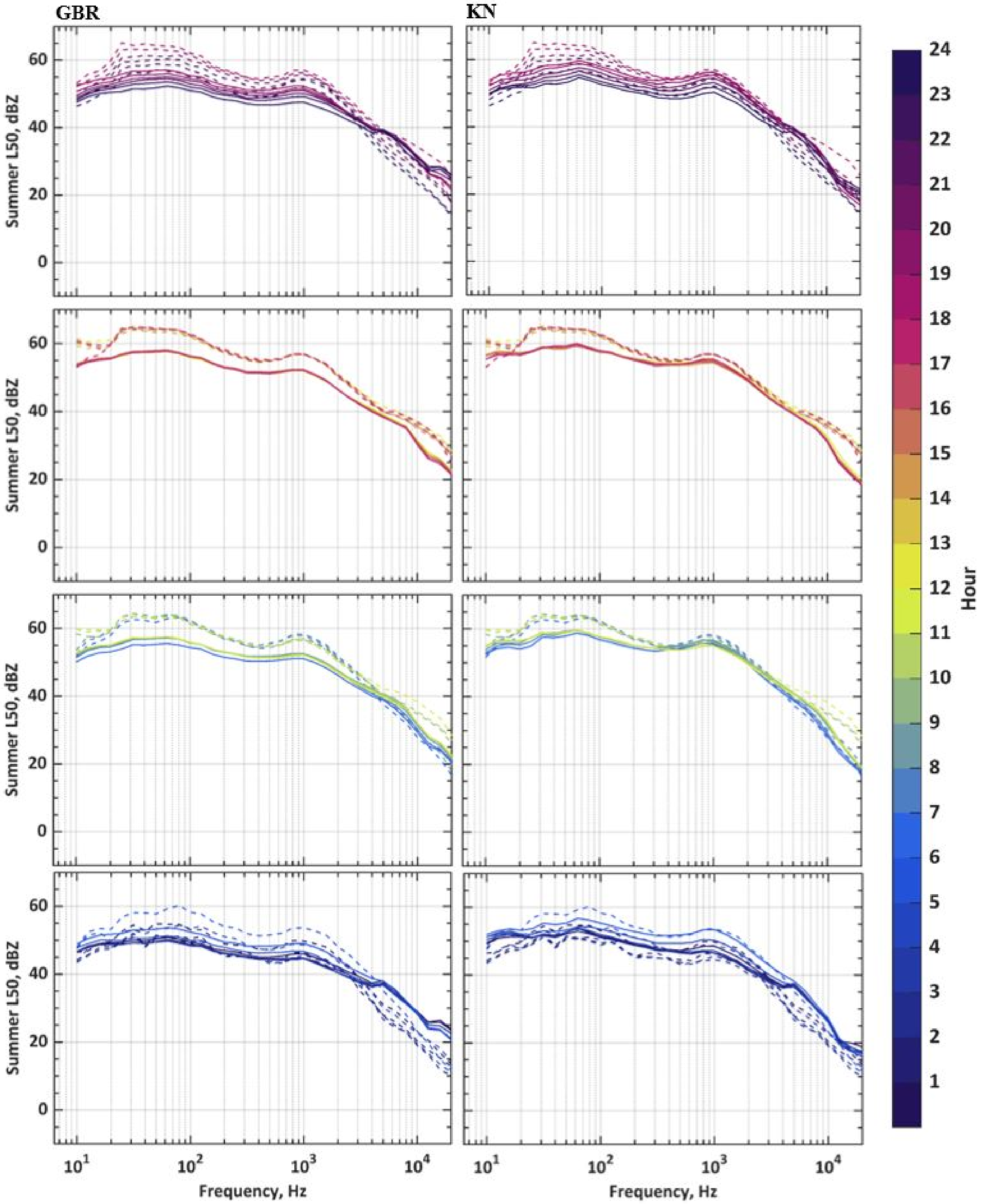

Figure 4.2 Hourly spectra measured levels (dashed lines) and LOSO predictions (solid lines) from an ensemble of GBR models (left) and KN models (right) for a site with significant traffic noise. Predictions are shown for 6-hour periods. From bottom to top these periods are: midnight to 6 a.m. (bottom), 6 a.m. to noon, noon to 6 p.m., and 6 p.m. to midnight (top).

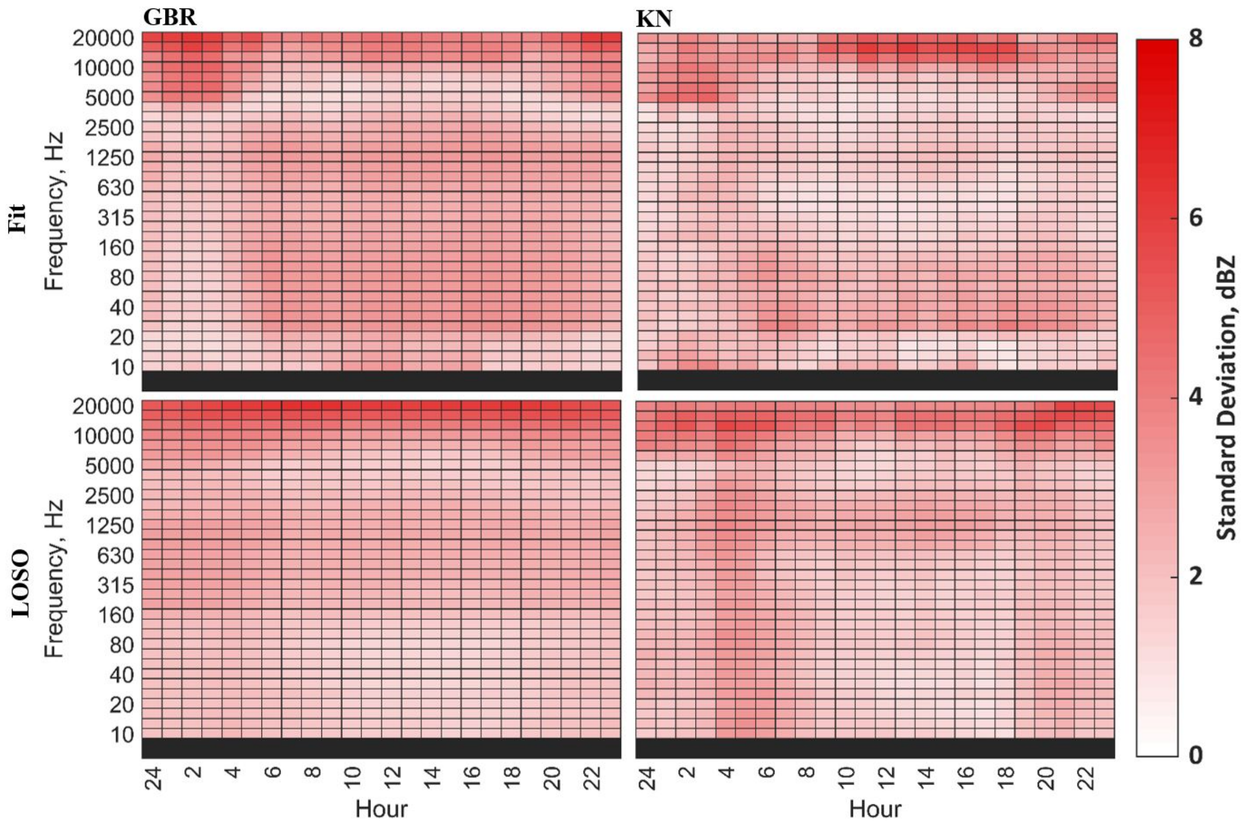

Figure 4.3 Standard deviation of ensemble model fit (top) and LOSO (bottom) predictions for the GBR (left) and KN (right) ensembles for a site with significant traffic noise. The horizontal and vertical axes indicate different hours and frequency bands, respectively.

4.3.3 LOSO predictions: insect noise

We show LOSO hourly L_{50} spectra predictions and the corresponding standard deviation in ensemble members for a different training site in Figures 4.4 and 4.5, respectively. Figures 4.4 and 4.4 have the same formatting as Figures 4.2 and 4.3. The training site used for validation in Figures 4.4 and 4.5 has significant insect noise as night, distant road noise, and occasional indirect overflights. It is located in Fletcher, near Asheville, North Carolina. Figure 4.4 shows that the KN ensemble tends to overpredict levels while the GBR ensemble does a fairly good job of predicting levels close to measured values. Note that the standard deviation of LOSO

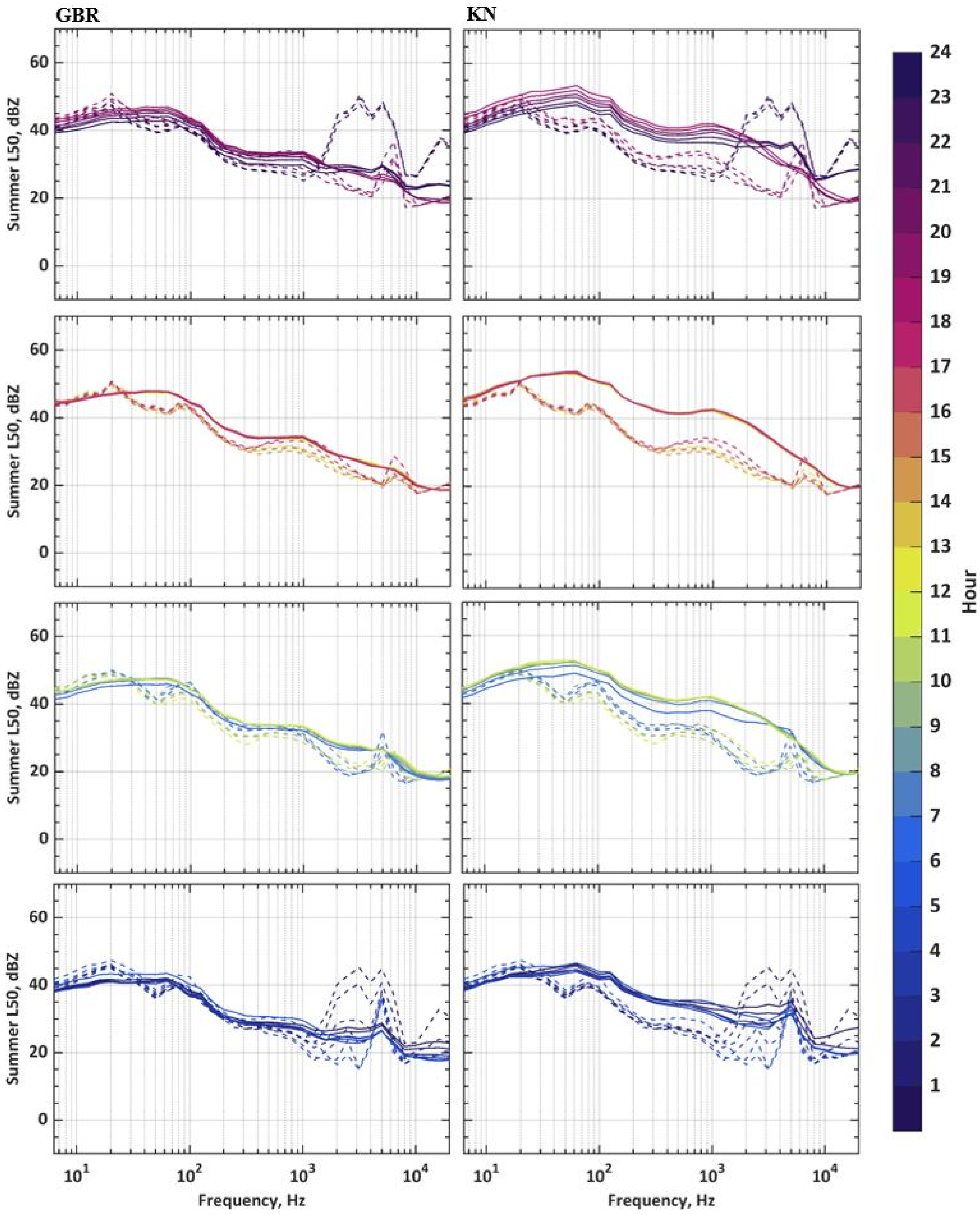

Figure 4.4 Hourly spectra measured levels (dashed lines) and LOSO predictions (solid lines) from an ensemble of GBR models (left) and KN models (right) for a training site with significant insect noise. Predictions are shown for 6-hour periods. From bottom to top these periods are: midnight to 6 a.m. (bottom), 6 a.m. to noon, noon to 6 p.m., and 6 p.m. to midnight (top).

predictions for the KN ensemble (see Figure 4.4) is in general larger than that of the GBR ensemble, particularly from 160 – 1250 Hz.

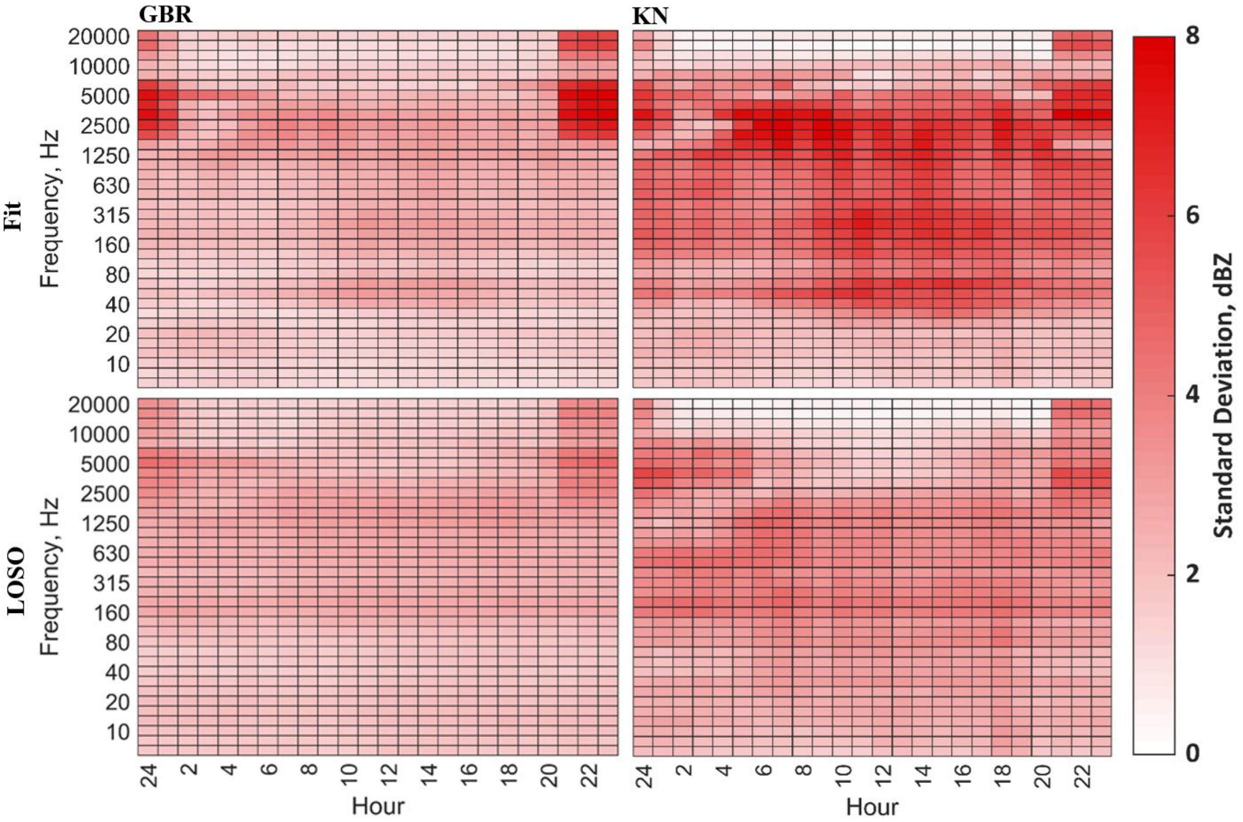

Figure 4.5 Standard deviation of ensemble model fit (top) and LOSO (bottom) predictions for the GBR (left) and KN (right) ensembles for a site with significant insect noise. The horizontal and vertical axes indicate different hours and frequency bands, respectively.

Figure 4.4 shows that insect noise at night covers a relatively wide range of one-third octave bands and reaches up to 40-50 dBZ. Both ensembles fail to predict levels of that magnitude. However, ensembles do predict a small spike near 5 kHz for some evening and nighttime hours. Interestingly, Figure 4.5 indicates that the uncertainty in fit predictions is larger at high frequencies than for LOSO predictions. This is likely due to similar sites in the training data that do not have insect noise. Therefore, when this somewhat atypical site is removed, uncertainty is lower because the model is unaware that similar sites may exhibit this amount of insect activity. The ensemble is overly confident because there are not enough training instances to capture the true variability in the data. This emphasizes the need for more data to improve model predictions and uncertainty estimates. Indeed, more training data may either help teach the model where and when increased insect noise is likely to occur or teach the model how confident it should be in predicting insect activity (or a lack thereof).

4.3.4 LOSO predictions: traffic and bird noise

Figures 4.6 and 4.7 show LOSO hourly L_{50} spectra predictions and corresponding standard deviations for the GBR and KN ensemble models for a site in Asheville, North Carolina with constant road noise, occasional direct overflights, and bird sounds. (Figures 4.6 and 4.7 are again formatted the same as Figures 4.2 and 4.3.) Figure 4.6 demonstrates that both models' LOSO predictions are close to measured levels. Even when predicted and measured levels disagree, much of the predicted spectral shape is accurate, particularly for traffic noise (i.e., at frequencies lower than about 1 kHz).

Models struggle more to predict bird noise (near 5 kHz). However, the models predict levels at 5 kHz quite well, and have higher errors and standard deviations (see Figure 4.7) at frequencies just below 5 kHz as well as above 10 kHz. This may be due to the contribution of insect noise at other sites such as the one described in Subsection 4.3.3. The high standard deviation shown in Figure 4.7 for some of the higher-frequency one-third octave bands is encouraging because it accurately identifies regions of relatively poor model accuracy.

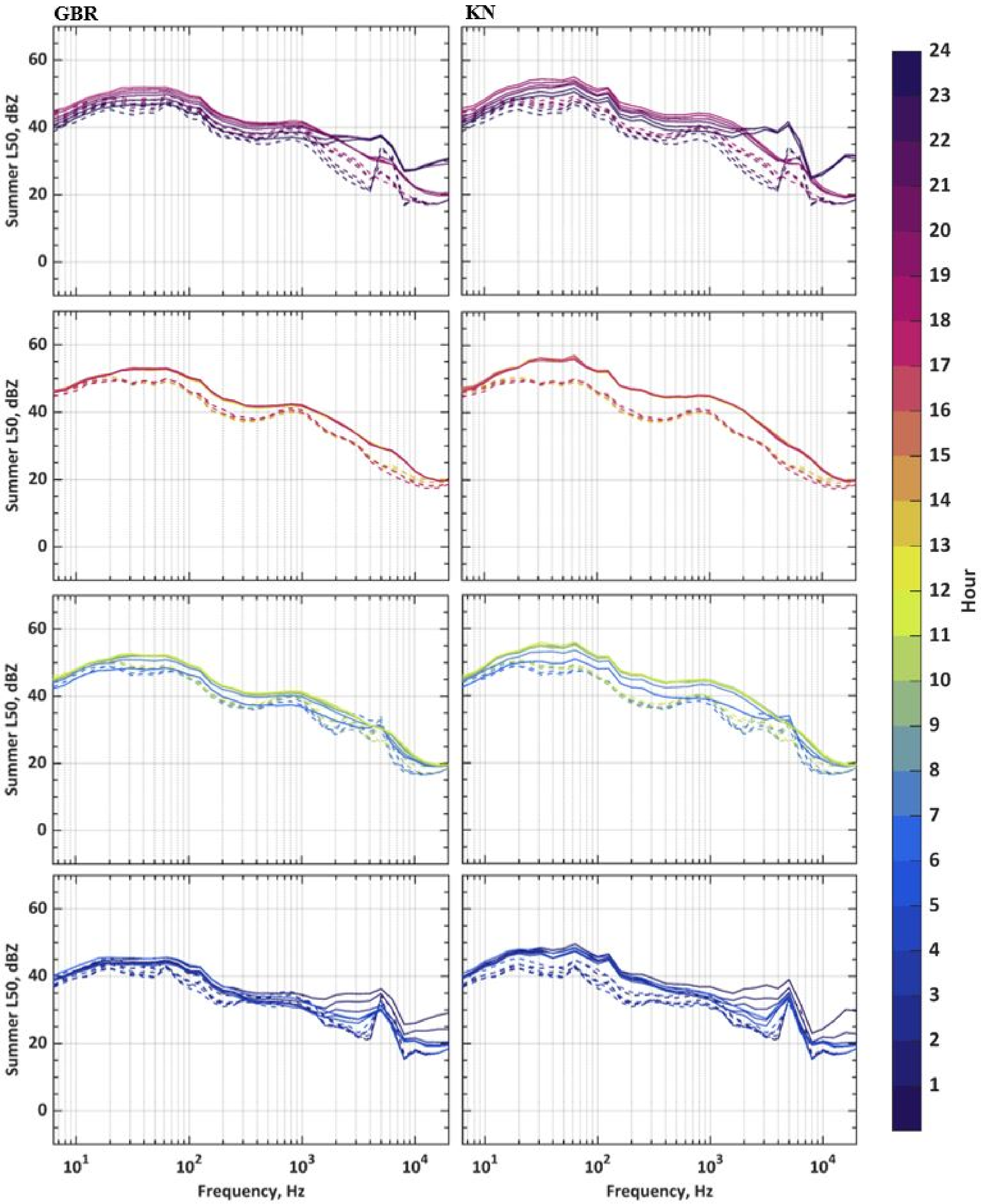

Figure 4.6 Hourly spectra measured levels (dashed lines) and LOSO predictions (solid lines) from an ensemble of GBR models (left) and KN models (right) for a training site with road and bird noise. Predictions are shown for 6-hour periods. From bottom to top these periods are: midnight to 6 a.m. (bottom), 6 a.m. to noon, noon to 6 p.m., and 6 p.m. to midnight (top).

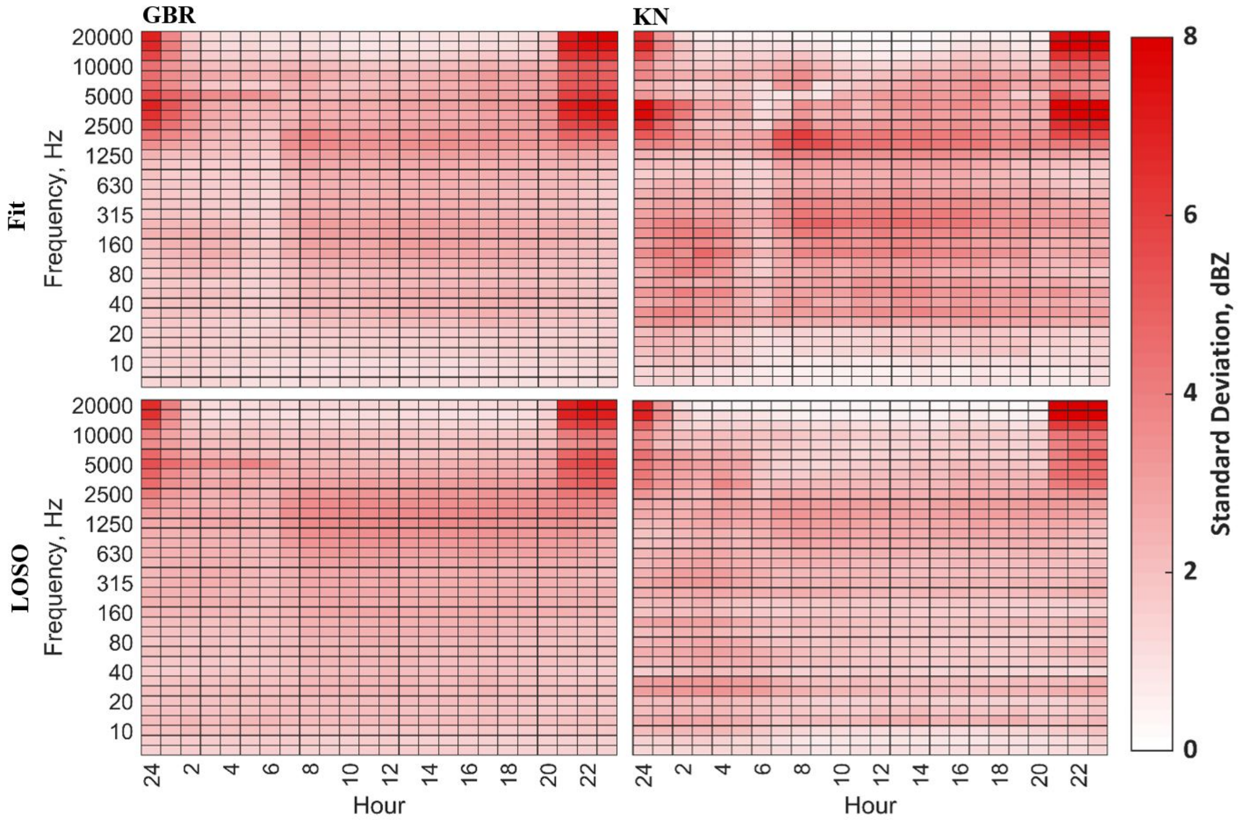

Figure 4.7 Standard deviation of ensemble model fit (top) and LOSO (bottom) predictions for the GBR (left) and KN (right) ensembles for a site with road and bird noise. The horizontal and vertical axes indicate different hours and frequency bands, respectively.

4.4. Conclusion

In this chapter, we discuss methods for training ensemble models for the prediction of summer hourly L_{50} spectra. In particular, we generate two ensembles of gradient boosted regression trees and k-nearest neighbors models trained on bootstrap subsets of the training data. We presented figures showing the leave-one-site-out hourly L_{50} spectra predictions and standard deviation in ensemble model fit and leave-one-site-out predictions for three training sites. One training site had significant road traffic noise, another had significant insect noise, and the third had road traffic and bird noise.

Leave-one-site-out predictions for all three training sites demonstrate that both ensembles have learned some relationships between geospatial data and summer hourly L_{50} spectra. Although spectral predictions may not always match predicted levels, the shape of the spectra is often in agreement with measured spectra; however, we observed that models struggle to identify when and to what extend insect and possibly bird noise is present. This may be due to inadequate training data or an inadequacy of the geospatial data to determine when insect/bird noise is likely present. The standard deviation of ensemble members is often useful in identifying predictions of higher uncertainty. However, we caution that this may not the case in extrapolation regions (i.e., geospatial regions not represented in the training data) because training data do not exist to either confirm or contradict model predictions in such regions.

Overall, the relative success of ensemble models for leave-one-site-out predictions of the summer hourly L_{50} spectra indicates supervised machine learning techniques may be successful for the prediction of acoustic metrics corresponding to different seasons, hours, one-third octave bands, exceedance levels, etc. if sufficient training data are available. The three sites shown in this chapter are likely similar to other sites in the training data. Indeed, the training data contain significant sites in both the Washington, D.C. and Asheville, North Carolina areas.

Therefore, future work includes applying the ensemble models described in this chapter to validation sites, which may be uncharacteristic of the training data. Additionally, constructing ensemble models from other model classes may help quantify structural uncertainty in predicted levels. Lastly, incorporating physics-based modeling for acoustic sources for which physical models exist (e.g., road traffic noise) may improve model accuracy near such sources.

Chapter 5

Feature selection for a geospatial model of environmental sound levels*

Katrina Pedersen,[1] Mark K. Transtrum,[1] Kent L. Gee,[1] Michael M. James,[2] and Alexandria R. Salton[2]

[1] Physics and Astronomy, Brigham Young University, Provo, UT, 84602
[2] Blue Ridge Research and Consulting, LLC, Asheville, NC, 28801

5.1 Abstract

Modeling environmental sound levels is challenging because they are the accumulation of all sounds in a geographic area and therefore depend upon a large number of geospatial features. In previous work, an ensemble of machine learning models was used to predict and estimate prediction uncertainty for environmental sound levels in the contiguous United States using a training set composed of 117 geospatial layers and acoustic data from 496 geographic sites [Pedersen et al., 2018]. To improve model accuracy, uncertainty estimates, and model interpretability, feature selection is performed on a set 120 geospatial features. An initial screening based on factors, such as data quality and correlations between features, reduces the set

*Chapter 5 is a manuscript to be submitted in the Journal of Acoustical Society of America.

to 51 features. Four different feature importance metrics are then utilized for further feature selection. Leave-one-out median absolute deviation cross-validation errors suggest that the number of geospatial features can be reduced to 15 without significant degradation of the model's predictive error. However, ensemble predictions demonstrate that results of feature selection are unstable to variations in details of the problem formulation, and therefore, should elicit some skepticism. Results suggest that more sophisticated dimensionality reduction techniques are necessary.

5.2 Introduction

5.2.1 Geospatial acoustics

Accurate modeling of environmental sound levels has many potential applications. In particular, accurate prediction of environmental sound levels may aid the National Park Service (NPS) in their charge to protect and restore natural acoustic environments within parks [45, 46]. Ecologists study the effects of environmental noise on animal behavior [50, 51, 53, 54, 56], especially for animals that respond to sound such as birds [17, 49-51], marine life [52-56], and anurans (i.e., frogs and toads) [57]. Additionally, public health studies have found that increased noise may be associated with changes in blood pressure, heart rate, and stress [6, 8] as well as mental health [67], cognitive function [68], and mental illnesses, such as depression and anxiety [2]. Accurate modeling of environmental sound levels may aid ecologists and public health workers in further identifying relationships between ecological and public health trends, respectively, and environmental noise. Accurate environmental sound modeling may also have applications in real estate, urban planning, and social justice.

Modeling of environmental sound levels is difficult due to the multitude of possible acoustic sources, propagation effects, etc. Remote sensing has increased the amount of geospatial data available for modeling, but these data are not free from error and may not always contain sufficient information to characterize an acoustic environment. Additionally, acoustic data are expensive to obtain and therefore relatively limited. Although some initial results are promising [1, 22, 23, 27, 65, 69], the problem is bottlenecked by limited data, making accurate environmental sound level modeling and model validation studies more challenging.

In previous work we used machine learning techniques to identify underrepresented areas for acoustic data collection [30, 65] and began to quantify uncertainty in model predictions in extrapolation regions [65]. In this paper, we explore feature selection methods in hopes of improving model accuracy, uncertainty estimates, and model interpretability.

5.2.2 Feature selection

Dimensionality reduction, or the process of reducing the number of features in a data set, has several benefits for machine learning such as minimizing the curse of dimensionality, improving model interpretability, decreasing computational and data requirements, and reducing uncertainty. The curse of dimensionality refers to challenges that occur when analyzing data in high-dimensional spaces, which do not occur in low-dimensional spaces [70]. Although data may be dense in low-dimensional spaces, data become more sparse as the number of dimensions (or features) increases. Machine learning identifies patterns and trends in data, so problems have higher data requirements in high-dimensional spaces. Since acoustic data are limited, it is likely that the sparsity of data is a challenge for environmental acoustic modeling and contributes to large errors and uncertainties reported in Chapters 2 and 3.

In the limited-data regime, there are additional benefits to dimensionality reduction. For example, models are more sensitive to noise when data are limited, so removing geospatial data that have large errors may improve model accuracy and uncertainty estimates. Another potential challenge of modeling limited data is that models may be prone to use correlated, rather than causal, geospatial features for predictions. Although this is desirable when the correlation extends to the extrapolation regime, with limited data machine learning can latch onto spurious correlations that limit transferability. Furthermore, when both a causal and correlated feature exist in the data set, it is beneficial to use the causal feature which will likely generalize better in extrapolation regions.

Dimensionality reduction can be divided into two types: feature extraction and feature selection [71]. Feature extraction methods, such as manifold learning, transform feature vectors into a lower dimensional space without losing information. In contrast, feature selection methods identify a subset of the original features. Feature extraction methods have the benefit that they can identify lower dimensional lossless representations; however, they lose physical interpretability. As an initial investigation into dimensionality reduction for environmental sound modeling, this work focuses on feature selection.

5.2.3 Previous work

Mennitt et al. [1, 22, 23, 69] have performed much of the pioneering work in continental-scale environmental sound level modeling, particularly for the Contiguous United States (CONUS). They used a training data set of geospatial and acoustic data to predict ambient outdoor sound levels in CONUS using random forest models. Mennitt et al. [1] performed feature selection by first removing features with high Pearson correlation coefficients. They then removed features one at a time, using the out-of-bag error due to permuting features to measure relative feature importance and removing the least important feature each iteration. The optimal number of features was determined by calculating the Leave-One-Out (LOO) Root-Mean-Square Error (RMSE) for all feature subsets and identifying the subset with the lowest LOO RMSE. After identifying the reduced feature set that minimized the LOO RMSE using default hyperparameters, five random forest hyperparameters were tuned. Mennitt et al. noted that although this process may not produce the best feature subset, it is computationally tractable.

Previous work done by both Mennitt et al. [1, 22, 23, 69] and the present authors [65] utilized LOO cross-validation to measure model performance. The LOO cross-validation error is computed by removing each site from the training data (one at a time), training a model on the remaining data, and then calculating the residual for the corresponding site. The LOO RMSE is the RMSE of all residuals. Although LOO cross-validation is more computationally expensive than other validation metrics, it is often more appropriate for small data sets, particularly when each training site may provide unique information to the model. Additionally, since data are limited, the computational costs of computing LOO cross-validation errors are not unreasonably large. However, using the LOO cross-validation to estimate model performance for extrapolation regions is not advisable.

The LOO cross-validation error is a traditional validation metric, and therefore assumes new input data are drawn from the same distribution as the training data. For the case of geospatial environmental acoustic modeling, this is unlikely to be true due to both limited acoustic data and a biased distribution of acoustic data. For example, many training sites (65%) are from national parks; however, national parks comprise only a small percent of total CONUS land area. Hence, LOO errors are not likely to be a good estimate of model uncertainty for much of CONUS. The problem of estimating model transferability, or the ability of a model to make accurate predictions for data statistically different from the training data, is an open area of research.

In previous work, we took steps to address the problem of model transferability for ambient acoustic environments by quantifying model structural uncertainty [65]. More precisely, we constructed an ensemble of six different machine learning models, each from a different model class: Gradient Boosted Regression trees (GBRs), Neural Networks (NNs), K-Nearest Neighbors (KNN), Support Vector Machines (SVMs), Kernel Ridge Regression (KRR), and Gaussian Process Regression (GPR). Model hyperparameters were tuned to minimize the LOO Median Absolute Deviation (MAD). Although LOO validation metrics are not good indicators of model transferability, they are computationally tractable and can indicate model accuracy on data drawn from a similar distribution as the training data. Hence, in the absence of more acoustic training data, it is reasonable to select model hyperparameters which minimize the LOO MAD. All six models had comparable LOO MAD errors.

The median of the ensemble model predictions was used to make environmental sound level predictions for CONUS. Additionally, the use of an ensemble model provided some measure of model transferability via the standard deviation of ensemble predictions, which was used as a surrogate for the structural uncertainty. There are many sources of uncertainty in the modeling process, including structural uncertainty, or model inadequacy, which arises due to lack of knowledge of the true generating function from which data are drawn [32]. Therefore, ensemble standard deviations do not correspond to rigorous confidence intervals; rather, they represent a minimum level of uncertainty.

In addition to aiding in quantifying model transferability, ensemble structural uncertainty estimates may help identify areas for model improvement. For example, areas of high structural uncertainty are likely good candidates for future acoustic data collection. Geospatial features that correlate with areas of large uncertainty are likely underrepresented in the training data and may also guide acoustic data collection efforts. However, geospatial features which correlate with areas of large uncertainty and which are unlikely to provide useful information about the acoustic environment are good candidates for removal.

5.2.4 Paper overview

In this paper, we extend our ensemble approach to environmental sound modeling and explore feature selection as a method to improve model accuracy, uncertainty estimates, and model interpretability.

Previously, 117 geospatial features were included in the modeling process (see Chapter 3). Here we start with a set of 120 geospatial features and manually remove 69 features due to large errors or uncertainties, significant correlations with other features, poor documentation, or the expected lack of a predictive relationship to the acoustic environment (e.g., latitude and longitude).

With the remaining 51 features, we then compare four feature selection methods. Each method identifies a reduced set of 15 features. These reduced sets give ensemble LOO MAD errors similar to those of all 51 features. Finally, we use an ensemble approach to calculate environmental sound level predictions in CONUS from the reduced-order models. Our results show that the predictions of the reduced-order models depend strongly upon details in the problem formulation, including the feature importance metric. This analysis indicates more sophisticated dimensionality reduction techniques are required.

5.3 Methods

5.3.1 Data sets

Data used in the modeling process were composed of both geospatial and acoustic data. The initial set of geospatial data contained 120 geospatial features described in Table A.1 of

Appendix A. Acoustic data for the summer daytime A-weighted L_{50} were obtained for 496 unique sites. These acoustic data and the geospatial data corresponding to the training sites are used to train the supervised machine learning models. For further explanation of the geospatial and acoustic data, we refer the reader to Appendix A.

5.3.2 Initial reduction to 51 features and feature scaling

Prior to utilizing the four feature importance metrics for feature selection, a quality review was performed for the existing 120 geospatial features. Feature processing, areas of analyses, sources of error, correlations with other features, and possible correlations to ambient sound levels were considered. Features were removed if they had large errors or uncertainties, significant correlations with other features, poor documentation, or expected lack of physical effect on the acoustic environment (e.g., latitude and longitude). A summary of the removed features is given below. The feature names are described in Table A.1 and the reduced set of 51 features is listed in Table A.2 of Appendix A. For convenience, we repeat Table A.2 here as Table 5.1.

Table 5.1 Subset of 51 geospatial layers used for environmental sound level modeling and feature selection in this paper.

Barren (200 m)	DistCoast	Herbaceous (5 km)	Slope
Barren (5 km)	DistMilitary	MilitarySum (40 km)	TdewAvgSummer
Cultivated (200 m)	DistRailroads	MixedForest (200 m)	TdewAvgWinter
Cultivated (5 km)	DistRoadsAll	MixedForest (5 km)	TMaxSummer
Deciduous (200 m)	DistRoadsMaj	PopDensity	TMaxWinter
Deciduous (5 km)	DistStreamO1	PPTSummer	TMinSummer
Developed (200 m)	DistStreamO3	PPTWinter	TMinWinter
Developed (5 km)	DistStreamO4	RddAll	VIIRSMean (270 m)
DistAirpHeli	Elevation	RddAll (5 km)	Water (200 m)
DistAirpHigh	Evergreen (200 m)	RddMajor	Water (5 km)
DistAirpLow	Evergreen (5 km)	RddMajor (5 km)	Wetlands (200 m)
DistAirpMod	FlightFreq (25 km)	Shrubland (200 m)	Wetlands (5 km)
DistAirpMoto	Herbaceous (200 m)	Shrubland (5 km)	

All land use features were removed due to possible errors in the layers, significant correlation with many of the land cover layers, and poor documentation. Many land use layers had sharp, unphysical discontinuities, which may be due to errors in the layers. For example, the Cropland layers had some unphysical looking discontinuities in south eastern North Dakota. All VIIRS layers, except the VIIRS mean upward radiance at night layer with a 270 m area of analysis, were removed due to large correlations with each other. The annual precipitation, minimum and maximum temperatures, and dew points were removed due to high correlations with the corresponding summer and winter layers. The RoadNoise and AviationNoise layers were removed because both are discontinuous for values below 35 dB. The forest land cover layer was omitted because the forest land cover layer is further divided into deciduous, mixed, and evergreen forest layers, which are included in the subset of 51 features. The DistWaterBody and PhyiscalAccess layers were removed due to concerns about errors in the layers. More specifically, some parts of rivers appear to be classified as bodies of water while other do not, and some values in the layer, particularly in California, Nevada, and Arizona, appear suspicious. The PhysicalAccess layer, on the other hand, had some extremely large unphysical values. The DistAirpSea layer was removed due to high correlations with the other DistAirp features. Lastly, Latitude and Longitude were removed since they should generally not have a physical effect on the acoustic environment.

In addition to reducing the feature set to 51 features, we rescaled geospatial features based on physical arguments. We scaled most features (i.e., all that do not depend on distance) based on their distribution within CONUS as opposed to their distributions in the training data. For these features, we use min-max scaling, which scales data to be between zero and one and preserves the shape of the data distribution. For some feature vector x, the scaled vector would be given by: $x_{scaled} = \frac{x-x_{min}}{x_{max}-x_{min}}$ where x_{min} and x_{max} are the minimum and maximum values of the feature x, respectively. For geospatial features that rely on distances however, such as DistAirpHigh and DistCoast, we use an arctangent function to scale data to be between zero and one: $x_{scaled} = \frac{2}{\pi}\arctan\frac{x}{x_0}$ where x_0 varies for different features and determines how quickly the function approaches one. Arctangent functions were selected for scaling of distance-dependent features because it is expected that after the distance exceeds some threshold, the feature's effect upon the ambient sound levels will not change. For example, after reaching 40 km from the nearest road, it is unlikely that the distance to the road will provide relevant information for environmental sound level predictions. Note that there is some ambiguity in the choice of appropriate distance thresholds (x_0) and further refinements to scaling methods may be considered in the future. However, these scaling methods are an improvement to those used in Chapter 3.

5.3.3 Feature importance metrics

All feature importance metrics have pros and cons—some of which are described here—and there is no single ‘best’ way to calculate feature importance. Previously, Mennitt et al. used the change in out-of-bag error due to permuting feature values in random forest models to estimate feature importance [23]. One disadvantage of permutation methods is that they often force models to focus on extrapolation regimes that may not correspond to allowed areas of the feature space [72]. Additionally, feature importance metrics, including permutation methods, tend to over-emphasize correlated features [73]. To investigate the stability of feature selection results for different feature importance metrics, we compare four different feature importance metrics: namely, Gini importance, Gini importance with a correlation penalty, neural network weights, and expert intuition.

The Gini importance metric, or mean decrease impurity, is a common feature importance measure for random forest or GBR models. The Gini importance for a given feature is calculated using the error reduction and number of instances split at each node corresponding to that feature. Although we do not provide an explanation of how to compute the Gini importance here, the interested reader is referred to [74, 75] for further details. The Gini importance is fast to calculate, but it is often biased in favor of features with higher cardinality or variability [76]. Additionally, the Gini importance has sometimes been shown to be biased towards correlated features [73]. For further information regarding the pros and cons of feature importance metrics for decision tree models, including the Gini importance and permutation method used in [23], the interested reader is referred to [77, 78].

To reduce bias due to correlations among features, we also calculated the Gini importance with a correlation penalty. After the calculation of the Gini importance for a given GBR model, we iterated through all features. For each feature $\mathbf{x}_i$ the most strongly correlated feature $\mathbf{x}_{corr}$ was identified. If the given feature $\mathbf{x}_i$ had a higher Gini importance than the correlated feature $\mathbf{x}_{corr}$, the Gini importance was unchanged. Otherwise, the Gini importance was decreased by a factor of $(1 - \mathrm{corr}_{max})$ where corr_{max} is the correlation corresponding to the maximally correlated feature $\mathbf{x}_{corr}$. If two features are almost identical, this metric will strongly penalize the feature with the lower Gini importance, giving it an importance score near zero, while leaving the Gini importance of the other feature unchanged. This feature importance metric has similar pros and cons to the original Gini importance metric, but penalizes correlated features.

The third feature importance metric is determined using the trained NN weights. There is no standard way to measure feature importance in a NN, but many methods have been suggested

[79]. We quantified the NN feature importance by first identifying all paths from an input feature to the output, and calculating the product of all weights along each path. Then, for each feature, the absolute value of all paths originating at that feature were summed together. Finally, these sums were normalized and the results were used as a feature importance measure. For the case of zero hidden layers, the feature importance was determined by the magnitude of the weights from the input features to the output.

The last feature importance metric was purely subjective. An expert familiar with environmental acoustic modeling used maps of the geospatial features, information about their processing methods, areas of analysis, correlations to other features, etc., to select which features would be most important for determining environmental sound levels in CONUS.

5.3.4 Feature selection process

After reducing the geospatial feature set from 120 to 51 features, we tuned model hyperparameters for the six supervised machine learning model classes (GBRs, NNs, KNN, SVMs, KRR, and GPR) to minimize the LOO MAD. Model hyperparameters are settings that a user selects for the learning process, such as the learning rate or activation function in a neural network. We used the tree-structured Parzen estimator approach implemented in hyperopt [43, 44], a Python library for automatic hyperparameter tuning, to determine appropriate hyperparameters. This approach tunes hyperparameters with minimal supervision so that we can periodically retune hyperparameters at different stages of feature selection. Hyperparameter search spaces were adjusted occasionally to account for varying feature subsets.

After tuning hyperparameters to minimize the LOO MAD and training all six members of the ensemble for the 51-feature model, we applied the four feature importance metrics described in the previous subsection to remove one feature at a time. For each metric, feature importance was calculated and the least important feature was removed to create four different feature subsets of size 50. All six model classes were retrained (using the hyperparameters identified from the 51-feature model) and the LOO MAD was calculated for all 24 models (6 models per subset of 50 features). At this point, there were four ensembles, each corresponding to a feature importance metric. For each ensemble, the corresponding feature importance metric was used to identify and then remove the least important feature again. This process was repeated every time a feature was removed.

Varying the number of geospatial features will change the optimal hyperparameters. Model hyperparameter tuning was performed at 51, 40, 30, 20, 15, 10, 5, 4, 3, 2, and 1 feature(s).

5.4 Results

5.4.1 Stability of feature selection results

Prior to examining the results of feature selection, it is beneficial to briefly explore the stability of results to changes in the problem formulation. Recall that the first two feature importance metrics are dependent upon the trained GBR model, the third metric is dependent upon the trained NN model, and the last metric is independent of all models and training data.

Since the Gini importance, Gini importance with correlation penalty, and importance calculated from the neural network weights all rely on trained models, varying hyperparameters affects feature importance estimates and the feature subsets identified by these metrics. Indeed, a comparison of the reduced feature sets generated using model hyperparameters tuned fewer times shows that results for these three data-driven feature importance metrics are sensitive to how often hyperparameters are tuned. For example, the top 15 ranked feature subsets vary both in features and rankings when hyperparameters are only tuned at 51 and 40 features rather than at 51, 40, 30, 20, and 15 features. For both the Gini importance and Gini importance with correlation penalty, 11 of the top 15 features are the same. For the importance calculated from neural network weights, only seven of the top 15 features are the same however. Therefore, tuning hyperparameters more often would result in different feature rankings and subsets.

There is also a certain amount of randomness in tuning hyperparameters and training models. Hence, there is some randomness in determining feature importance estimates for the three data-driven metrics. We found that both changing the random seed used to sample the hyperparameter spaces as well as making small changes to the hyperparameter search spaces resulted in different optimal hyperparameters and feature rankings.

Additionally, in the limited-data regime, the three quantitative methods are biased by the distribution of training data, and are therefore likely to select features that correlate with environmental sound levels in the training data, regardless of whether or not those correlations hold true for most of CONUS. More particularly, we found that when Longitude was added to the set of 51 geospatial features and hyperparameters were tuned for that set of 52 features, both model predictions and feature rankings changed. Note that longitude is strongly correlated (Pearson correlation coefficient of 0.59) with the training data due to sampling bias. The maximum, mean, and median absolute differences of ensemble model predictions in CONUS for the summer daytime A-weighted L_{50} for the sets of 51 and 52 features were 11.1, 0.9, and 0.7 dBA, respectively. Moreover, Longitude was ranked in the top 11 features for all three metrics when using hyperparameters tuned for the set of 52 features.

Despite, the sensitivity of the data-driven feature importance metrics to relatively small changes in the problem formulation, a comparison of feature selection results can help determine appropriate methods for dimensionality reduction for environmental sound level modeling. Since the last feature importance metric (i.e., expert intuition) does not rely upon trained models or the training data, it provides an interesting contrast to the other three data-driven methods of computing feature importance.

5.4.2 Changes in the LOO MAD error

Figure 5.1 shows the LOO MAD vs. the number of features for each of the four feature importance metrics and the six models that compose the ensemble model. All models were trained to predict the summer daytime A-weighted L_{50}. Figure 5.2 similarly shows the LOO

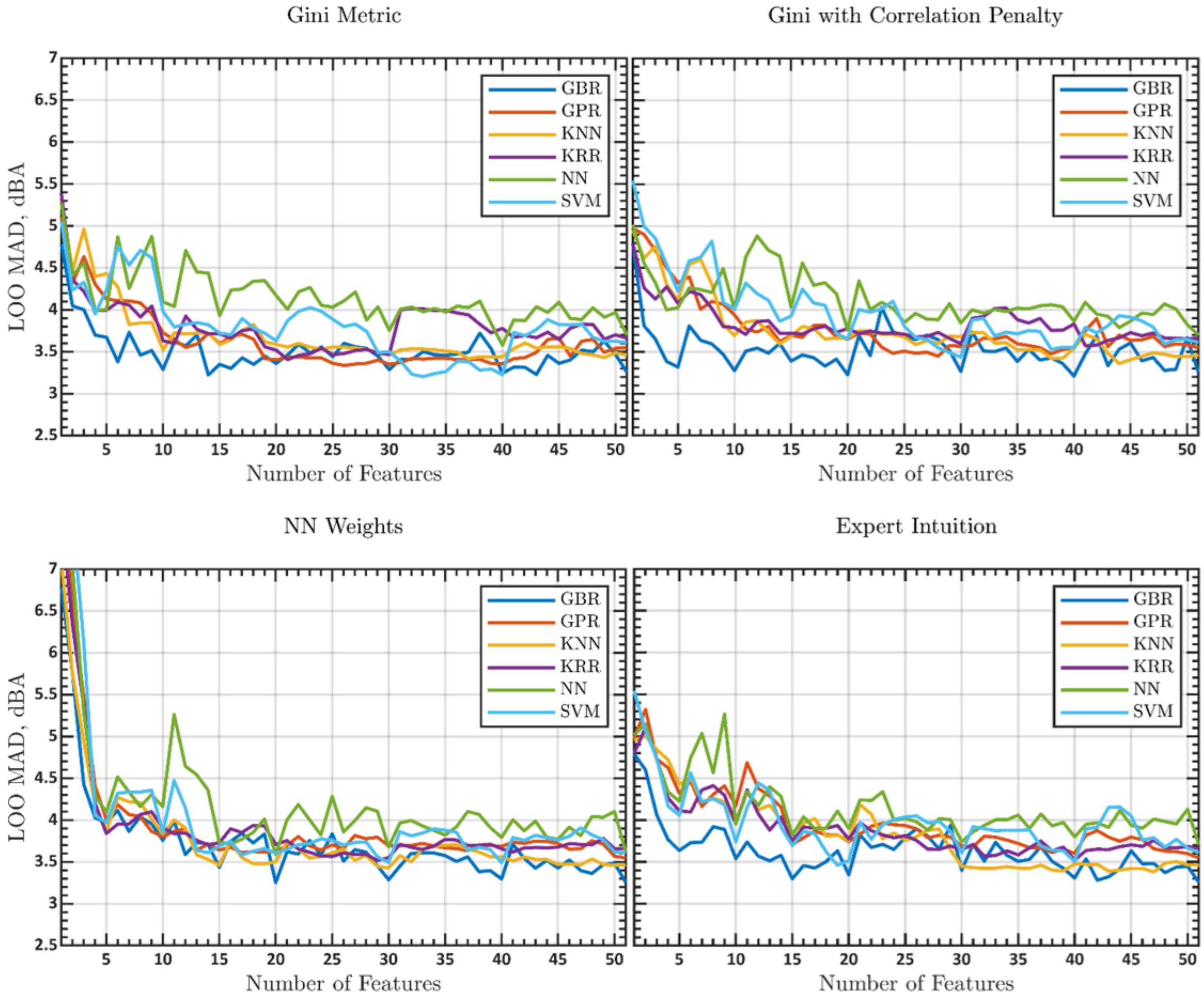

Figure 5.1 LOO MAD errors for the summer daytime A-weighted L_{50} as a function of the number of features. Model hyperparameters were tuned at 51, 40, 30, 20, 15, 10, 5, 4, 3, 2, and 1 feature(s).

MAD for the ensemble models, which are determined by the median prediction of all six members. Recall that hyperparameters were tuned at 51, 40, 30, 20, 15, 10, 5, 4, 3, 2, and 1 feature(s). All four feature importance metrics perform similarly, especially for larger feature sets. It is interesting that the number of features can be reduced significantly from the 51 initial features without much change to the LOO MAD. However, recall that the LOO MAD is a traditional validation metric. So, although models may perform well on the training data, LOO MAD makes no guarantees to how the models will generalize to data drawn from a different distribution (i.e., CONUS).

To further investigate the models generated during feature selection, we analyzed the reduced feature sets of 15 features and their corresponding CONUS ensemble predictions. We selected feature subsets of size 15 because all four ensembles had relatively low LOO MAD values there (likely in part due to hyperparameter tuning) and LOO MAD errors tended to start increasing as features continued to be removed.

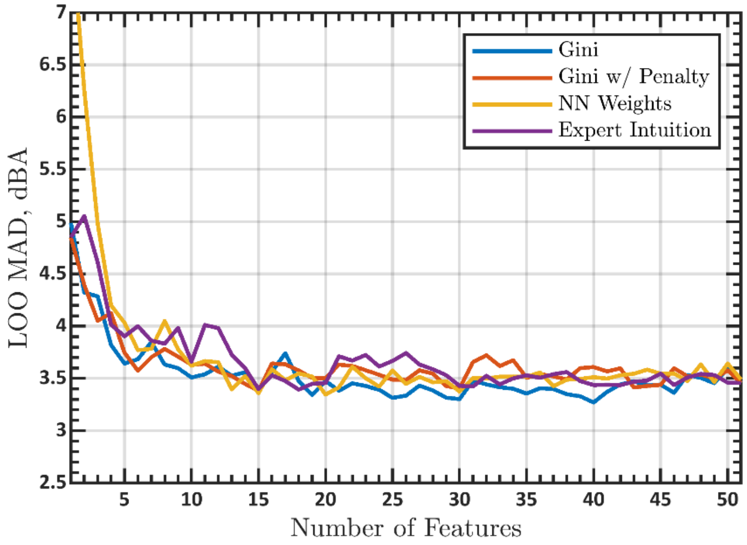

Figure 5.2 LOO MAD ensemble errors for the four metrics of determining feature importance as a function of the number of features. All models were trained to predict the summer daytime A-weighted L_{50} and ensemble predictions were determined by the median predicted level of the six ensemble members.

5.4.3 Comparison of top 15 features

Table 5.2 lists the top fifteen features identified by the four feature importance metrics in order of importance. (The features in the bottom row would have been removed next and the features on the top row were the last remaining features used to train the 1-feature models.) Interestingly, all four feature subsets include a feature that gives information about the distance to the nearest road, the distance to the nearest stream, and the amount of evergreen land cover. Three of the four subsets also include information about the mean upward radiance at night (VIIRS layer), road density, and the amount of shrubland and herbaceous land cover. It is unsurprising that all feature subsets include features that describe the distance to the nearest road, since road noise is likely a significant source of noise in urban areas. Likewise, features describing the distance to the nearest stream and the type of land cover are likely some of the most important features for the prediction of ambient sound levels in natural environments. Hence, the feature lists are not unreasonable.

Despite the similarities in the feature subsets, there are many significant differences. Each of the four subsets contains at least five unique feature layers, with the subset corresponding to the Gini importance with a correlation penalty containing the most (eight) unique layers. The subset corresponding to the Gini importance metric only contains one land cover feature while the three other subsets contain five to seven land cover features each. Additionally, the expert clearly favors land cover layers with a 200 m area of analysis while the Gini importance with a correlation penalty favors land cover layers with a 5000 m area of analysis. Given the many differences between the reduced feature sets, it is interesting that they all give comparable ensemble LOO MAD errors. This is likely due in part to limited training data and correlations among the geospatial features. Further information regarding model behavior can be gained by looking at ensemble predictions.

Table 5.2 Top fifteen features identified by various feature importance metrics (re-ranking after removing each lowest-ranked feature using the 15-, 10-, 5-, 4-, 3-, and 2-feature tuned models).

Gini Metric	Gini Metric with Correlation Penalty	Neural Network Weights	Expert Intuition
TdewAvgSummer	VIIRSMean (270 m)	TMinWinter	VIIRSMean (270 m)
VIIRSMean (270 m)	DistCoast	Water (5 km)	RddAll
Slope	DistRoadsMaj	Barren (5 km)	DistRoadsMaj
DistRoadsMaj	Shrubland (5 km)	RddAll (5 km)	DistStreamO3
DistStreamO3	PopDensity	Evergreen (5 km)	FlightFreq (25 km)
Evergreen (5 km)	Slope	DistRoadsMaj	PopDensity
DistMilitary	DistMilitary	Developed (200 m)	DistRailroads
PPTWinter	TMaxWinter	Barren (200 m)	Cultivated (200 m)
DistStreamO1	DistStreamO3	RddAll	Deciduous (200 m)
Elevation	Wetlands (5 km)	DistStreamO1	Wetlands (200 m)
RddAll (5 km)	Evergreen (5 km)	DistAirpMoto	Herbaceous (200 m)
DistAirpLow	DistRoadsAll	Shrubland (200 m)	Shrubland (200 m)
DistAirpHeli	Herbaceous (5 km)	FlightFreq (25 km)	Evergreen (200 m)
PPTSummer	TMaxSummer	DistStreamO4	Developed (200 m)
DistAirpMoto	Deciduous (5 km)	Herbaceous (200 m)	TdewAvgSummer

Figure 5.3 shows the ensemble predictions for the summer daytime A-weighted L_{50} for the 15-feature reduced feature sets corresponding to the features listed in Table 5.2. Training sites are indicated by small circles and colored according to measured levels. Even though the four 15-feature ensemble models give similar LOO MAD error measures, CONUS ensemble predictions vary significantly among the four ensembles. To emphasize this, the spread of ensemble model predictions for all four feature importance metrics and all sites in CONUS is plotted in a histogram in Figure 5.4. The spread of ensemble predictions has a mean, median, and maximum of 6.9, 6.5, and 29.8 dBA, respectively. Note that a difference of 6 dBA corresponds to a doubling of sounds pressure level, so these differences in the CONUS

predictions are not small. These results emphasize that the LOO MAD is not a reliable indicator of model accuracy in the extrapolation regime.

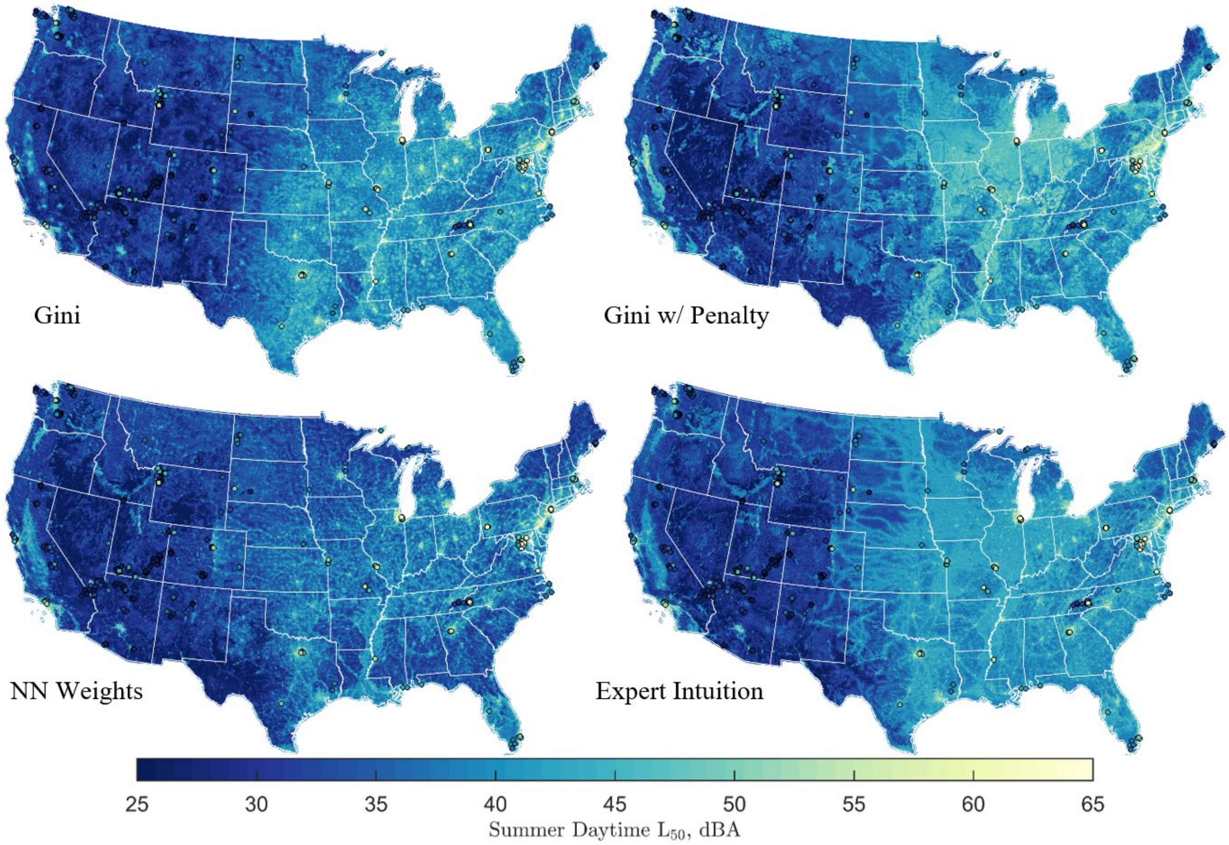

Figure 5.3 CONUS ensemble predictions of the summer daytime A-weighted L_{50} for models trained using the top 15 features identified from four different feature importance metrics. Training sites are indicated by small circles and are colored according to measured levels.

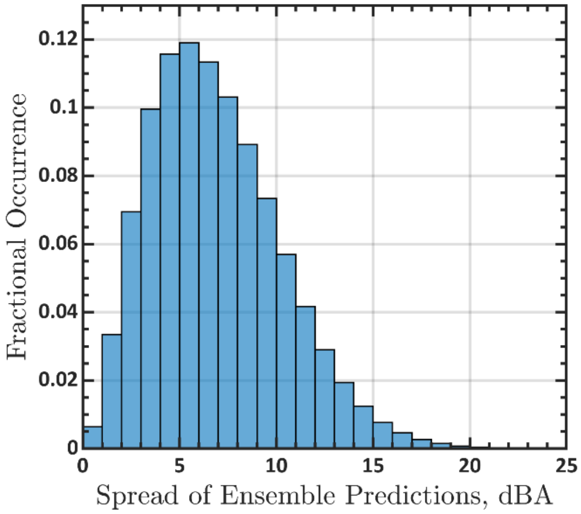

Figure 5.4 Histogram of the spread of CONUS ensemble predictions of the summer daytime A-weighted L_{50} for models trained using the top 15 features identified from four different feature importance metrics.

More generally, results suggest it is not beneficial to compare traditional validation errors, such as LOO MAD, for ensemble models generated from different feature subsets because they do not describe model accuracy in extrapolation regions. This points to the need for better dimensionality reduction methods, which do not rely on the model's predictive performance. Rather, dimensionality reduction techniques that attempt to characterize intrinsic dimensions of the data may be better suited to environmental acoustic modeling. Such feature extraction techniques are unsupervised and therefore do not rely on supervised machine learning models, hyperparameter optimization, validation error metrics, etc. Additionally, they can utilize information from all geospatial features while still reducing the dimensionality of feature space.

In Chapter 3, we quantified structural uncertainty of ensemble model predictions using the standard deviation of ensemble models. We do not include any maps of the estimated structural uncertainty in this paper because a comparison of these maps does not provide further insight into identifying appropriate feature selection methods or guiding acoustic data collection.

In particular, since acoustic training data are limited, training data are sparse in feature space. Removing features will change which regions of feature space are sparse, therefore also changing which points in CONUS are in the extrapolation regime. This in turn changes which points in CONUS are likely to have larger structural uncertainties and are therefore good candidates for acoustic data collection. Indeed, ensemble models trained on the four subsets of 15 features produce different estimates of structural uncertainty, indicating acoustic data should be collected in different locations for different models. In general, comparing structural uncertainty estimates for different feature subsets will not provide easily interpretable results since the feature subsets correspond to different feature spaces in which different regions are sparse.

Despite similar LOO MAD errors from the selected feature subsets of 15 geospatial features, there is notable disparity between the four feature subsets and their corresponding ensemble predictions in CONUS. This demonstrates that feature selection results are sensitive not only to the details of hyperparameter tuning (e.g., the size of the search space and frequency of hyperparameter tuning), but also to the choice of feature importance metric.

5.5 Conclusion and future work

Environmental sound level modeling is an important but challenging problem with various potential applications, including aiding the preservation of natural acoustic environments within national parks and informing ecological and public health studies. In this paper we explored the viability of dimensionality reduction via feature selection for environmental sound level modeling in attempts to improve model accuracy, uncertainty estimates, and model interpretability and decrease computational requirements.

A feature set of 120 geospatial features was reduced to a set of 51 geospatial features by removing features with large errors or uncertainties, significant correlations with other features, poor documentation, or lack of physical effect on environmental sound levels. Following the reduction to 51 features, we further reduced features using four feature importance metrics; namely, Gini importance, Gini importance with a correlation penalty, neural network weights, and expert intuition. Feature selection was performed iteratively by training an ensemble model, determining the least important feature, as measured by each of the four feature importance metrics, and removing that feature. Hyperparameters were tuned occasionally to minimize the leave-one-out median absolute deviation. All models were trained to predict the summer daytime A-weighted L_{50}.

Leave-one-out median absolute deviation measures indicated that the cardinality of feature space could be reduced to 15 using all four feature importance metrics before error started to increase noticeably. The four feature sets were significantly different (i.e., they did not generally contain the same geospatial features). Additionally, ensemble model predictions for the contiguous United States indicated large variability in extrapolation regions among the four models. More specifically, the spread between the four ensemble models of predicted summer daytime A-weighted L_{50} levels in the contiguous United States had a mean, median, and maximum of 6.9, 6.5, and 29.8 dBA, respectively.

These results further demonstrate that traditional validation metrics, such as the leave-one-out median absolute deviation are poor indicators of model transferability as discussed in Chapters 2 and 3. Additionally, results show that feature selection is strongly dependent upon the feature importance metric. An investigation of the stability of feature selection results also showed that reduced feature sets are sensitive to details of the problem formulation. In particular, results are sensitive to the frequency of hyperparameter tuning, the hyperparameter search space, and the random seed used to identify optimal hyperparameters. This should be cause for suspicion of feature selection for the problem of environmental sound level modeling. Indeed; since the results of feature selection are unstable to variations in details of the problem formulation, they should not be taken seriously. This motivates the need for more sophisticated dimensionality reduction techniques. In particular, feature extraction methods that describe the intrinsic dimensionality of the data and do not rely on a model may be better suited to environmental sound level modeling.

Acknowledgements

This research was supported by a U.S. Army Small Business Innovation Research (SBIR) contract to Blue Ridge Research and Consulting, LLC.

Chapter 6

K-means clustering of 51 geospatial layers for characterization of outdoor acoustic environments*

Katrina Pedersen,[a] Ryan R. Jensen,[b] Lucas K. Hall,[c] Mitchell C. Cutler,[a] Mark K. Transtrum,[a] Kent L. Gee[a]

[a]*Department of Physics and Astronomy, Brigham Young University, Provo, UT, 84602, USA*
[b]*Department of Geography, Brigham Young University, Provo, UT, 84602, USA*
[c]*Department of Biology, California State University Bakersfield, Bakersfield, CA, 93311, USA*

6.1 Abstract

Outdoor acoustic environments have a complex relationship with biotic and abiotic features. Acoustic environments influenced by anthropogenic activity are of particular interest because ambient noise may negatively affect humans and wildlife. Therefore, characterization of different outdoor acoustic environments, particularly those in which ambient noise dominates, may hold value in various scientific fields including public health and ecology. This paper

*Chapter 6 is a manuscript to be submitted with a different Introduction and Conclusion to *Applied Geography*. The Introduction and Conclusion here were written to better align with content in this dissertation.

investigates outdoor sound levels in the continental United States via a k-means clustering model that generated eight clusters using 51 geospatial layers, selected and scaled to describe outdoor acoustic environments. Cluster maps are shown and a subclustering analysis is presented in which each of the original eight clusters is further divided into two clusters. Clusters and subclusters describe different sound level clusters that may be linked to human or wildlife behavior or health outcomes.

6.2 Introduction

Overall outdoor sound levels in an area are influenced by anthropogenic and natural factors; likewise, outdoor sound levels affect both human and animal life. In particular, ambient noise, or unwanted outdoor sound due to anthropogenic activity, may negatively affect human and animal life. Conversely, the preservation of natural soundscapes enhances visitor experiences and helps mitigate anthropogenic effects upon animal life in protected areas (e.g., national parks) [13, 80].

Anthropogenic factors, including urban areas, transportation features/corridors (e.g., railways, airports, etc.), military bases, and energy development operations have all been linked to high levels of ambient noise [81-85]. Different anthropogenic features/activities may also impose additive and potentially interactive effects on overall sound levels. Despite the influence of anthropogenic activities on outdoor sound levels, their effects on largescale patterns of ambient noise are still in development.

Ambient noise can impact human health and well-being and studying noise in the context of public health is timely [86]. Ambient noise may be considered ‘the new secondhand smoke’ for millions of people, disrupting sleep, impacting hearing loss, and causing cardiovascular disease related to stress (e.g., high-blood pressure) [11, 87]. Further, higher noise levels can negatively impact human creativity [88], depression [3], and impair cognition in children [89]. Additionally, there is evidence that noise exposure may vary between different racial/ethnic and socioeconomic groups, possibly contributing to health inequalities [33]. In short, ambient noise is a biological stressor and public health hazard [87].

Ambient noise may also influence animal life and species interactions, especially for animals that utilize auditory signals. Many mammals, birds, amphibians, and insects use sound for an important life characteristic, such as antipredator defense or navigation [16], and some species are wholly reliant upon sound for reproduction (e.g., frogs, toads, birds, some ungulates). For species that rely upon sound, competing sounds (e.g., anthropogenic noise) can have varying levels of impact [90-97]. For example, ambient noise has been indicated as a causal factor for changes in avian behavior and community diversity [17, 18], marine life [19], and anurans (i.e., frogs and toads) [20, 21].

Organisms may have an ecological effect on shaping the acoustic landscape. Many invertebrate and vertebrate species use extensive acoustic displays that periodically influence outdoor sound levels. For example, the mating calls of cicadas occur in intense pulses that can persist over several weeks [98, 99]. While not as long in duration, the ‘avian dawn chorus’, created as multiple species of breeding birds vocalize at dawn (and sometimes at dusk), can result in marked sound levels [100, 101]. These acoustic displays are relatively short in duration, however, primarily occurring during reproductive events. In contrast, vegetation may contribute more to overall outdoor sound levels than animals; not by generating sound, but rather by attenuating or diffusing sound [102-104].

This paper examines outdoor sound levels throughout the continental United States (CONUS) using 51 biotic and abiotic geospatial layers and k-means clustering to map specific clusters of sound. It is based on the working hypothesis that outdoor sound levels are influenced by anthropogenic activity, landscape structure, landscape characteristics, land use, land cover, climatological, and geophysical variables. The paper does not link outdoor sound levels to human or ecosystem health; rather, it maps sound level clusters that may ultimately be linked to human health outcomes or ecological effects.

6.3 Methods

6.3.1 Geospatial layers

A set of 51 geospatial raster layers, each with a 270-meter spatial resolution, were obtained from the National Park Service Natural Sounds and Night Skies and Inventory and Monitoring Divisions database [24, 62]. These layers can be classified into five categories: topography, climate, land cover and land use, hydrology, and anthropogenic. A detailed description of each layer is given in Appendix A.

Prior to use in clustering, data were scaled to prevent biases in clustering due to variations in the range of values in different layers. In particular, layers that do not vary with distance (from some acoustic source) were scaled using min-max scaling, which scales data to be between zero and one while preserving the shape of the data distribution. For geospatial features that rely on distance (e.g., DistCoast, DistAirpHigh), an arc tangent function was used to scale data to be between zero and one. An arc tangent function was applied to distance-dependent features to emphasize changes in distance close to points of interest (e.g., the coast, airports).

6.3.2 K-means clustering

K-means clustering is an unsupervised machine learning algorithm which clusters data into *k* clusters. More specifically, the algorithm first randomly selects *k* samples from the data set to initialize the *k* cluster centroids. Each data point is then assigned to the cluster corresponding to the closest centroid as measured by the Euclidean distance. Centroid locations are then updated to correspond to the mean of all data points in the corresponding cluster. The process of assigning data points to the nearest cluster centroid and adjusting centroid locations is repeated until cluster centroids are stable. Hence, k-means clustering attempts to identify natural clusters in the data [105]. We used k-means clustering as implemented in the Python library scikit-learn [106].

One challenge of k-means clustering is determining the appropriate number of clusters *k* such that data are classified meaningfully and descriptively. Two common methods for determining the appropriate number of clusters are silhouette analysis and elbow analysis. Silhouette analysis is performed by calculating the average silhouette score for k-means clustering models trained using various values of *k* and selecting the model with the highest score. The silhouette score is a measure of how similar points within the same cluster are and how dissimilar points from different clusters are [107]. The silhouette score ranges from –1 to 1, where 1 represents perfectly clustered data and –1 represents poorly clustered data.

Similar to the silhouette analysis, elbow analysis requires training multiple k-means clustering models for different values of *k*. Elbow analysis uses the distortion (i.e., the sum of squares distance between the data and its nearest cluster) to identify an appropriate number of clusters. The distortion is a monotonic decreasing function and the optimal number of clusters is the point where adding another cluster to the model begins to only marginally reduce the distortion [108]. This happens at the "elbow" in a plot of the distortion. We use silhouette and elbow analyses to identify the appropriate number of clusters *k* for a set of 51 geospatial layers over CONUS.

6.3.3 Subclustering

To further examine the geospatial data and their clusters, we perform a clustering analysis on each of the initial *k* clusters. Silhouette analysis is used to identify an appropriate number of subclusters for each cluster. For all clusters, the optimal number of subclusters is determined to be two. Note that the silhouette score cannot be calculated for a single cluster, so silhouette analysis cannot indicate whether a single cluster (i.e., no subclustering) should be preferred.

Therefore, to determine if subclustering into two subclusters is beneficial for any individual cluster, the estimated probability densities of the distance (as measured by the Euclidean norm) between instances within each subcluster and the corresponding initial cluster centroid were plotted. These plots were overlaid with similar plots using the identified subcluster centroids (rather than the initial cluster centroid). This was done for each cluster to compare the results of subclustering into two subclusters and performing no subclustering. These distributions can be seen in Appendix B, Figures B.1-B.8. For the case in which adding a second centroid (i.e., subclustering into two clusters) significantly moved both distributions to the left, it is more likely that subclustering is beneficial for further describing the data. Marginal shifts indicate that subclustering made only minor improvements in accurately clustering the data at the cost of simplicity in the model.

In this paper, we present results of performing subclustering for all clusters into two subclusters, independent of the changes in the distributions of the distance to centroids shown in Figures B.1-B.8. Depending on the desired application of clustering results, subclustering may prove useful even when the results of subclustering do not immediately indicate improved clustering. Although we do not discuss which cases of subclustering appear most beneficial, the interested reader is referred to Appendix B for further subclustering results.

6.4 Results and discussion

6.4.1 Determining the number of clusters

The results of performing silhouette and elbow analyses on the 51 geospatial layers are shown on the left and right of Figure 6.1, respectively. Recall that a higher average silhouette score is indicative of better clustering, so the silhouette analysis identifies eight clusters as the optimal number. The results of the elbow analysis are more challenging to interpret because identifying the “elbow” in the plot (i.e., the location at which adding another cluster begins to only marginally reduce the distortion) is somewhat subjective. However, the “elbow” appears to be around seven or nine clusters. Given the subjective nature of determining the location of the “elbow,” we give more weight to the results of the silhouette analysis. Therefore, we use eight as the optimal number of clusters since both silhouette and elbow analyses indicate this is a reasonable choice.

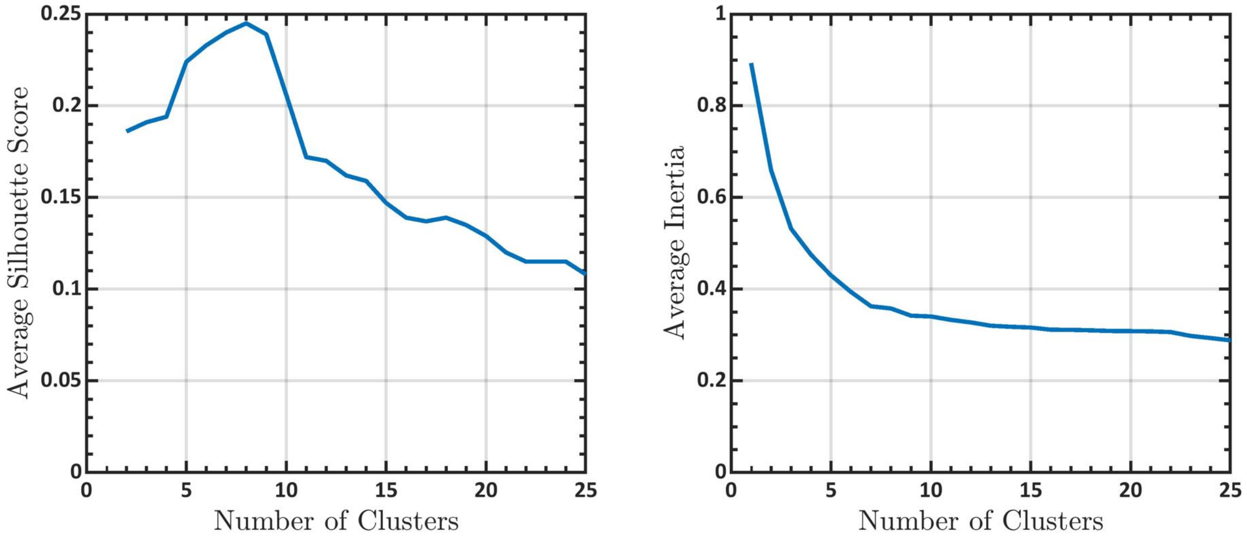

Figure 6.1 Silhouette (left) and elbow (right) analyses showing the average silhouette score and average inertia, respectively, as the number of clusters is varied.

6.4.2 Eight-cluster model

Letting *k* equal eight, k-means clustering was applied to the 51 geospatial layers for all of CONUS using a 270-meter spatial resolution. Each cluster is assigned a color and a map of the resulting clusters is shown in Figure 6.2.

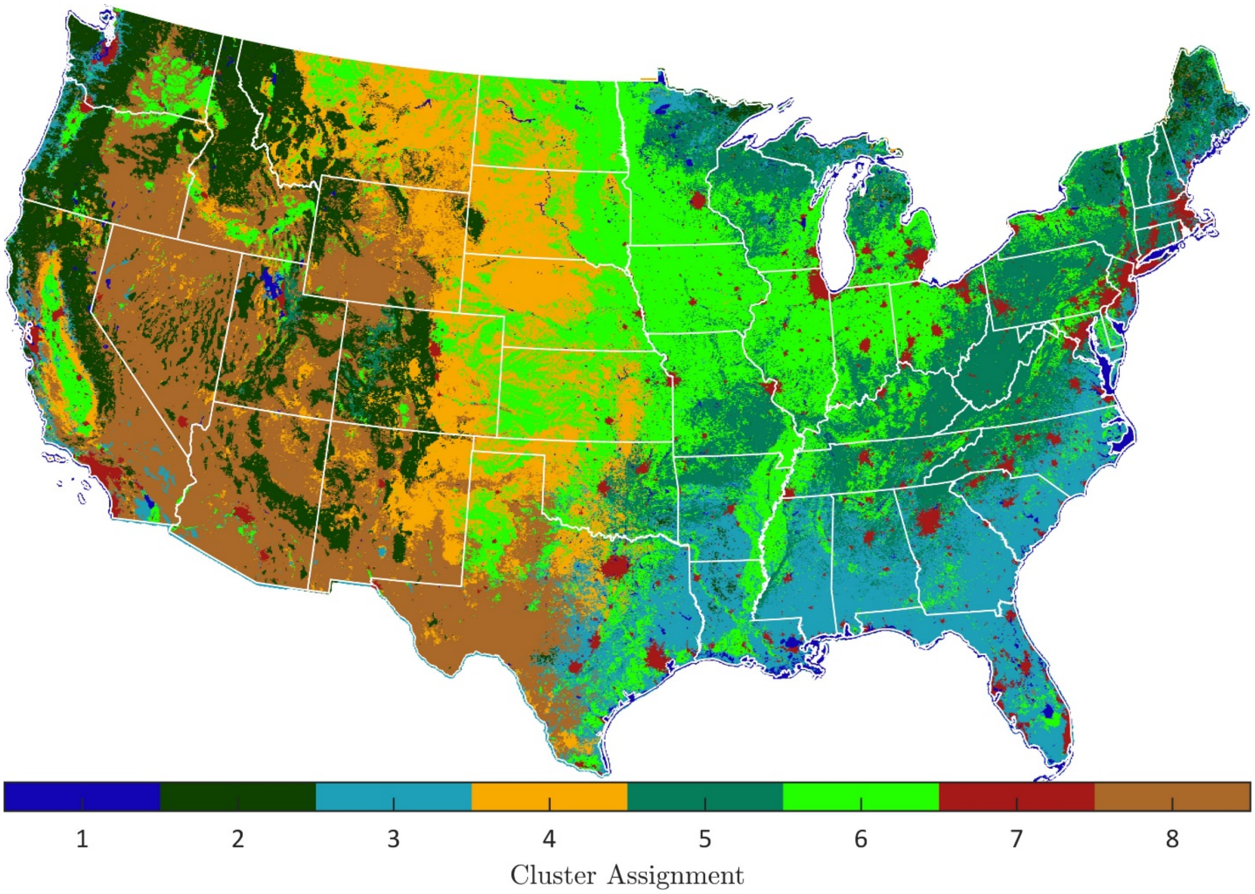

Figure 6.2 CONUS cluster assignments after clustering 51 geospatial features with a 270-meter spatial resolution.

Each of the eight clusters is impacted by several variables. Table 6.1 shows the top three distinct/unique correlated geospatial variables as ranked by the magnitude of the Pearson correlation coefficient. Correlation coefficients for each cluster were calculated using the scaled geospatial layers and Boolean values to denote whether a site resided within the given cluster.

Note that for geospatial variables corresponding to multiple layers due to differing areas of analysis, only the largest magnitude correlation among all layers is reported.

Some trends are immediately apparent in the eight-cluster map. Cluster one is influenced by water and is prevalent along the coasts and larger bodies of water – e.g., the Great Salt Lake. Cluster two represents evergreen forests and areas with higher degrees of slope. Cluster three is impacted by winter dew point temperatures and wetlands – thus representing relatively humid areas. Herbaceous vegetation and low amounts of winter precipitation are represented in cluster four throughout the northern and southern plains. Cluster five represents both deciduous and mixed forest environments while cluster six's most important variable is cultivated (crop) land. Cluster seven is heavily influenced by developed or urban areas. Finally, cluster eight represents shrubland and low summer precipitation and low summer dew point average temperature.

Table 6.1. Top three distinct/unique correlated geospatial variables ranked by magnitude for each cluster.

Cluster Number	Rank 1	Rank 2	Rank 3
1	Water (0.91)	DistCoast (-0.44)	DistRoadsAll (0.35)
2	Evergreen (0.78)	Slope (0.46)	TMinSummer (-0.46)
3	TdewAvgWinter (0.50)	Wetlands (0.49)	TMinWinter (0.42)
4	Herbaceous (0.88)	PPTWinter (-0.28)	DistAirpHigh (0.28)
5	Deciduous (0.84)	MixedForest (0.29)	PPTSummer (0.27)
6	Cultivated (0.89)	Shrubland (-0.32)	DistRailroads (-0.29)
7	Developed (0.76)	RddMajor (0.70)	RddAll (0.67)
8	Shrubland (0.90)	PPTSummer (-0.49)	TdewAvgSummer (-0.44)

Seven of the eight clusters' most impactful variables are land cover related and five of those seven are strongly correlated with vegetation. This may signify that land cover (especially vegetation) is the most important variable with regard to outdoor sound levels. The only cluster with a non-land cover variable ranked as the most important variable was cluster three (winter dew point average temperature ranked only slightly higher than wetlands land cover). However, other variables, such as dew point temperature and precipitation are also important in the cluster analysis (Table 6.1).

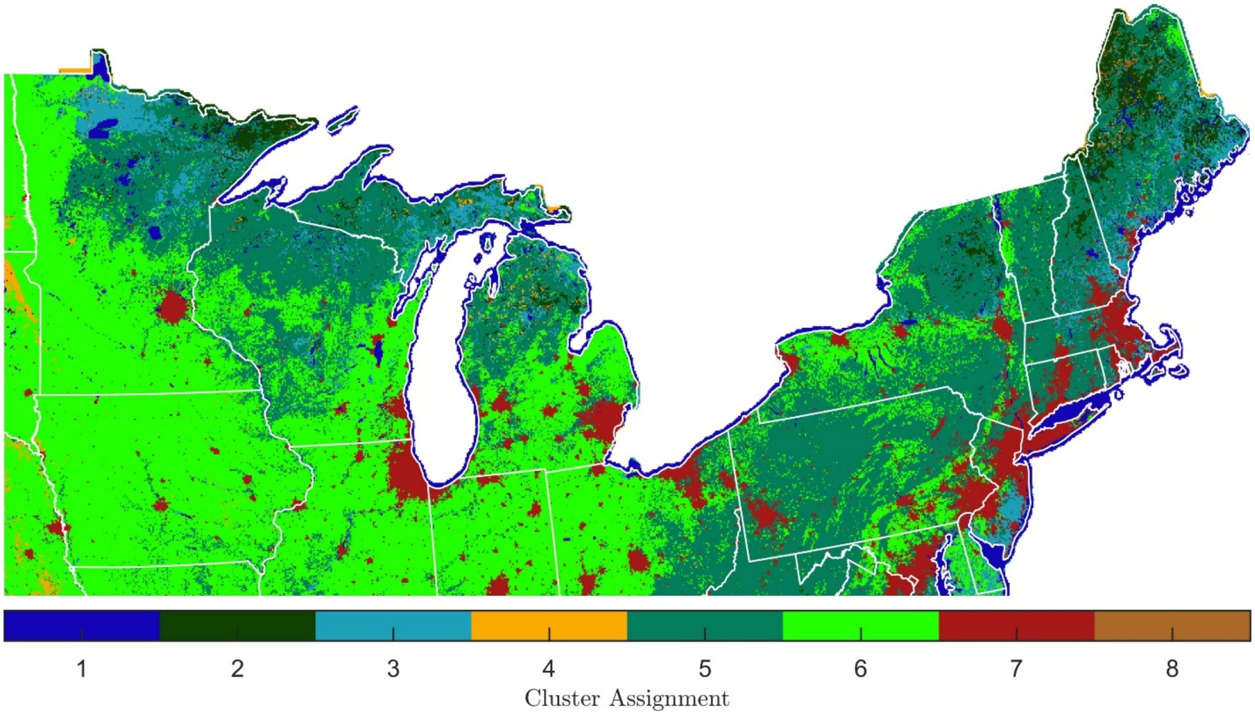

Figure 6.3 Cluster assignments in the Great Lakes region and North East as a result of clustering 51 geospatial features with a 270-meter spatial resolution.

Figure 6.3 shows a zoomed-in cluster map of the Great Lakes region and North East. The upper Midwest and northeastern United States are dominated by clusters five, six, and seven. Cluster five is heavily impacted by deciduous forest and cluster six is mostly influenced by cultivated (crop) land. The large urban/suburban areas of Chicago, Detroit, Minneapolis, Boston, New York City, etc. are in cluster seven, which is most strongly affected by developed land cover.

Figure 6.4 shows zoomed-in cluster maps of Utah (left) and the eastern/southeastern coastal plains (right). The west/southwest region of the United States is primarily represented by clusters eight and two, which are most strongly correlated with shrubland and evergreen land cover, respectively. Much of the eastern/southeastern coastal plains are represented by cluster three, which is positively related to average winter dew point temperatures and wetlands. Deciduous forests are evident in cluster five throughout much of the Piedmont and Appalachian Mountains. In both Utah and the eastern/southeastern coastal plains, urban/suburban areas are well-mapped in cluster seven. In particular, the developed northern Virginia-Washington DC-Baltimore-Philadelphia corridor is striking.

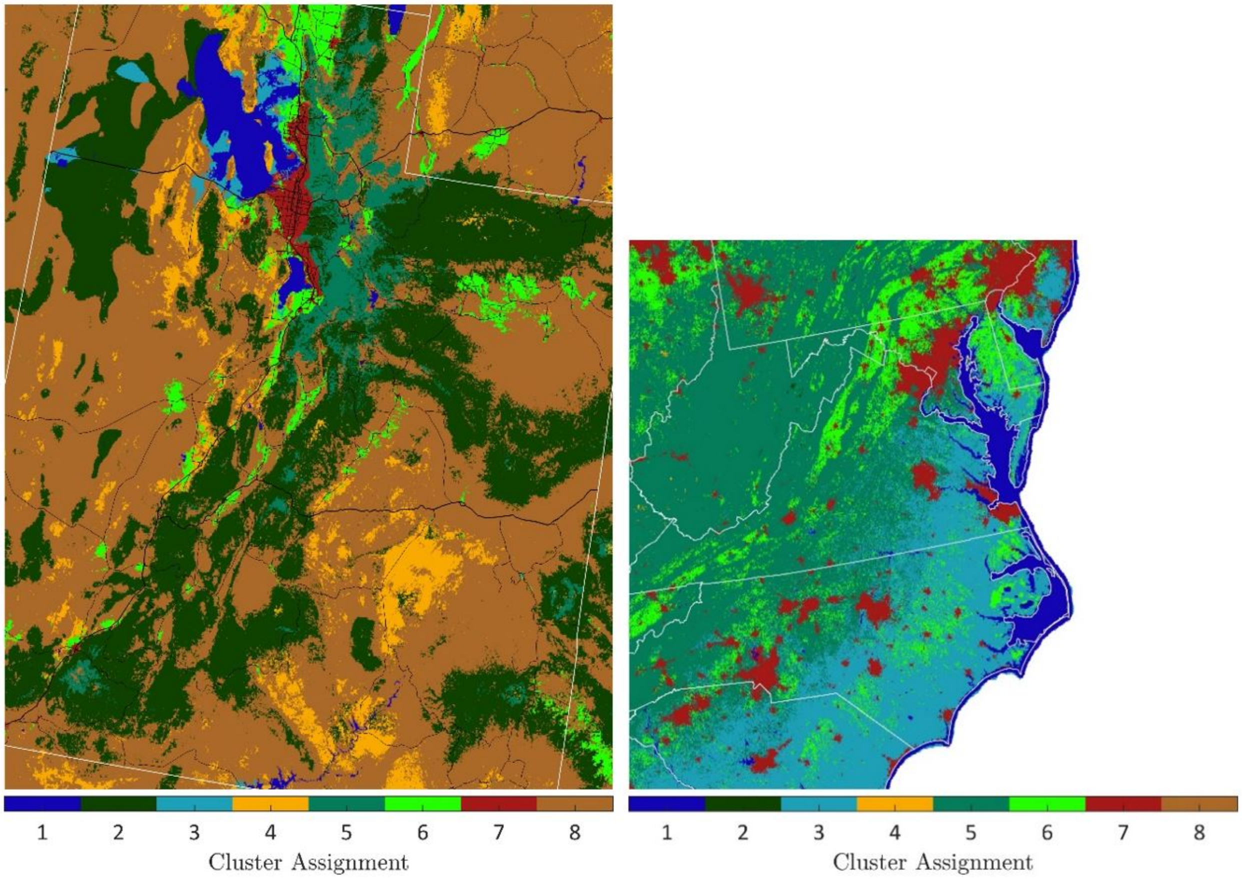

Figure 6.4 Cluster assignments in Utah (left) and along the eastern/southeastern coastal plains (right) as a result of clustering 51 geospatial features with a 270-meter spatial resolution.

6.4.3 Subclustering

Each of the initial eight clusters was further divided into two subclusters. We refer to the first and second subcluster for each cluster by the corresponding cluster number and a letter, “a” for the first subcluster and "b" for the second subcluster. For example, subclusters 8a and 8b are the subclusters corresponding to cluster eight (colored brown in maps above). For simplicity, we will refer to the model which subclusters each of the original eight clusters into two subclusters as the 16-subcluster model. A CONUS map of the subclusters is given in Figure 6.5 (see Figure B.9 for individual CONUS subcluster maps). The first color for each subcluster, corresponding to all "a" subclusters, is the same as the initial cluster color. The second color for each subcluster, corresponding to all "b" subclusters, is a lighter shade of each initial cluster color.

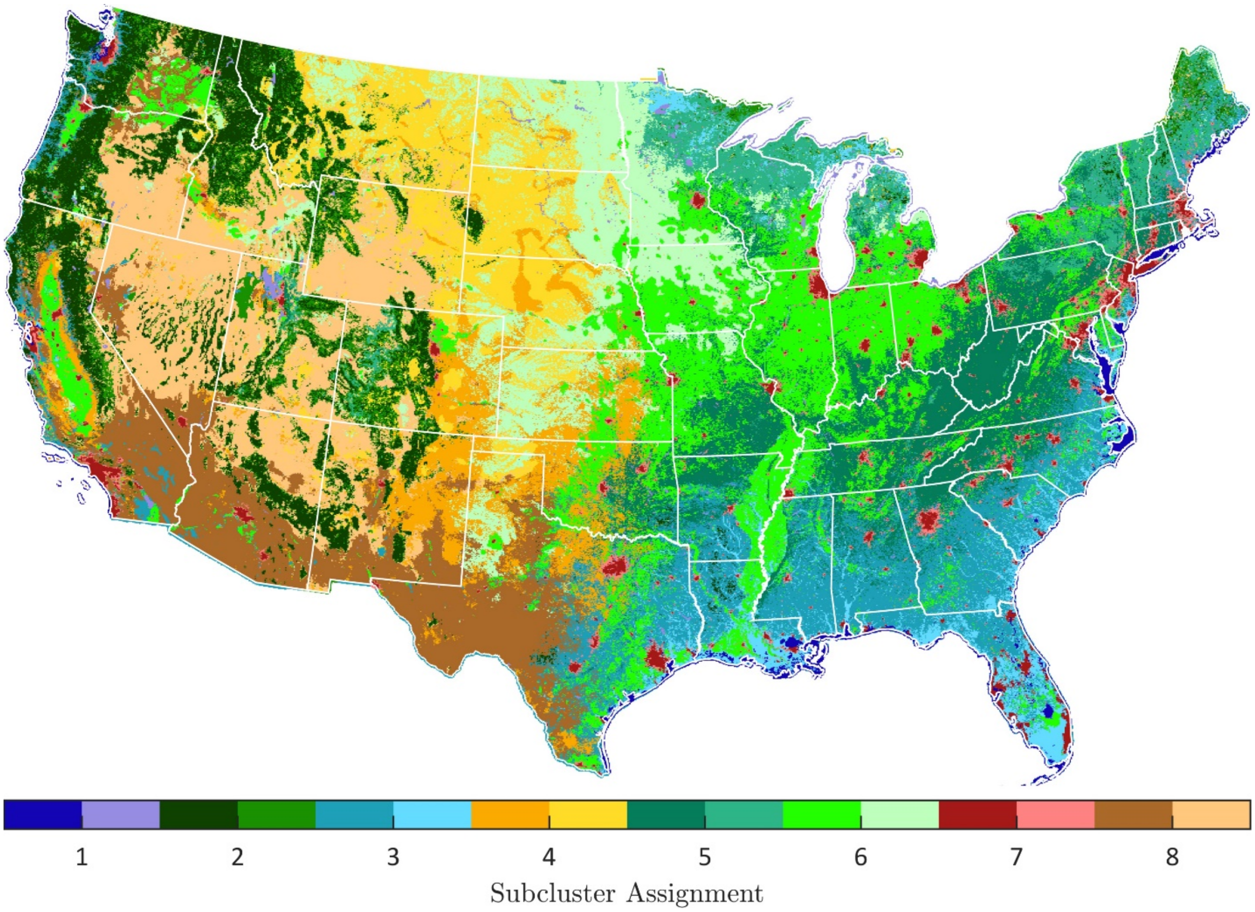

Figure 6.5 CONUS subcluster assignments after clustering each cluster from the 8-cluster model into two subclusters.

The 8-cluster and 16-subcluster models are overall similar in identifying influential environmental factors underlying the model predictions (see Tables 6.1 and 6.2). Note that rankings in Table 6.2 were calculated in a similar manner to those in Table 6.1. Some of the largest observable differences between the 8-cluster and 16-subcluster models are the further distinction of clusters three, seven, and eight. In the 16-subcluster model, predominant wetland areas within cluster three are more clearly separated from the rest of the coastal plains. Subclustering also helps differentiate between two densities of urban activity (subclusters 7a and 7b). In the 8-cluster model, the deserts of the western United States are represented well by cluster eight. In the 16-subcluster model however, cluster eight subclusters distinguish between the cold (subcluster 8b) and hot (subcluster 8a) deserts of the western United States.

Table 6.2 Top three distinct/unique correlated geospatial variables ranked by magnitude for each subcluster.

Subcluster Number	Rank 1	Rank 2	Rank 3
1a	Water (0.70)	DistCoast (-0.58)	DistRoadsAll (0.34)
1b	Water (0.60)	DistRoadsAll (0.14)	TMaxWinter (-0.08)
2a	Evergreen (0.78)	Slope (0.38)	TMinSummer (-0.34)
2b	Elevation (0.31)	Evergreen (0.30)	TMinSummer (-0.29)
3a	TdewAvgWinter (0.43)	TMaxWinter (0.37)	TMinWinter (0.37)
3b	Wetlands (0.79)	PPTSummer (0.27)	TdewAvgWinter (0.23)
4a	Herbaceous (0.57)	DistMilitary (-0.18)	TMaxSummer (0.15)
4b	Herbaceous (0.63)	DistAirpHigh (0.31)	TMinWinter (-0.27)
5a	Deciduous (0.78)	PPTSummer (0.25)	FlightFreq_25km (0.25)
5b	MixedForest (0.39)	Deciduous (0.29)	TMaxWinter (-0.27)
6a	Cultivated (0.65)	DistAirpHeli (-0.28)	Elevation (-0.27)
6b	Cultivated (0.53)	TMinWinter (-0.31)	TMaxWinter (-0.27)
7a	Developed (0.44)	RddAll (0.31)	RddMajor (0.30)
7b	RddMajor (0.74)	Developed (0.74)	RddAll (0.67)
8a	Shrubland (0.59)	TMaxSummer (0.41)	TMaxWinter (0.36)
8b	Shrubland (0.62)	TdewAvgSummer (-0.48)	Elevation (0.45)

6.4.4 Applications

The ability to map and determine characteristics of sound has many potential applications. Geographic models of environmental sound levels may utilize clustering results to identify areas that may have similar acoustic characteristics across large geographic scales. For data-driven environmental sound models which rely on acoustic measurements, clustering may also help identify optimal locations for further acoustic data collection.

Characterization of acoustic environments has applications in the study of human and animal health and behavior. Although there are many sources of natural sound, ambient noise often results in much higher sound pressure levels and can have harmful effects on human health [86] and wildlife [96]. However, the relationship between different acoustic environments and their effects upon humans and wildlife is still under investigation; indeed, it has only recently been found that noise may affect wildlife [109].

This research could help inform public health studies by aiding in determining correlations between human health and acoustic environments. Additionally, this research could help wildlife policy and management – including identifying suitable wildlife corridors and examining how ambient noise impacts both aquatic and terrestrial animal habitats as many species experience anthropogenically driven habitat changes [109]. Further, the methods used and maps generated in this study may enable large-scale studies on the effect of noise on racial and ethnic diversity, socioeconomic disparity, animal communication, distribution, and other topics.

6.4.5 Limitations

The 51 geospatial layers and scaling of layers were selected with the goal of distinguishing different acoustic environments. However, it is likely that there are acoustic sources, propagation effects, etc., which are not well-represented in the geospatial data. Additionally, it is possible that some acoustic effects are not well-represented in clusters due to trends in a larger number of geospatial layers, which dominate clustering assignments. We also note that the spatial resolution (270 meters) makes cluster maps ineffective at investigating local (i.e., over small spatial regions; e.g., <1km) acoustic clusters.

6.5 Conclusion

Largescale patterns of vegetation are influential in the 8-cluster CONUS model; five out of the possible eight clusters in the CONUS model are strongly correlated with vegetation. These results present two interesting thoughts: (1) vegetation plays a large role in outdoor sound levels even across large expanses of terrain (e.g., CONUS) and (2) outdoor sound levels are influenced by the type of vegetation present. This indicates that, to the extent acoustic environments vary with vegetation type, they are in principle resolvable by geospatial data. Each of the five clusters that have a strong correlation with vegetation are best represented by a different type of vegetation community. These unique vegetation communities also likely correspond to unique animal communities, which may in part, contribute to sound levels (e.g., wetlands). Taken together, our findings highlight the potential importance of the acoustic properties of unique biotic communities (plants and animals) and should be considered in future largescale models of outdoor sound.

As apparent with clusters six and seven, anthropogenic features (e.g., cultivation and urban land development) play a role in landscape level outdoor sound clusters. However, the way in which anthropogenic features influence outdoor sound levels varies across spatial scales. For example, urban and suburban areas (cluster seven) demonstrate a concentrated effect on outdoor sound clusters, whereas cultivated areas (cluster six) have a much broader effect on outdoor sound clusters that occurs over large expanses (e.g., much of the Midwest). Interestingly, we also notice that cultivated lands are moderately correlated with the distance to the nearest airport or heliport – indicating that in the absence of ‘natural’ vegetation other factors are more important. Humans have an influential role on ecosystems and it is not surprising that anthropogenic features may impose additive effects on outdoor sound levels across landscapes.

The outdoor sound level clusters presented in this paper may inform studies of ambient noise and its effect upon human and animal life and behavior. Indeed, we anticipate that different outdoor acoustic environments may correlate with trends in public health and ecology. As human populations grow, ambient noise is likely to increase in some manner, whether that be in sound level or spatial extent or both. Therefore, it is important to study both the relationship between ambient noise and biotic life and the underlying mechanisms that determine outdoor acoustic environments.

Acknowledgments

This work was partially funded by the College of Physical and Mathematical Sciences at Brigham Young University and a U.S. Army Small Business Innovation Research (SBIR) contract to Blue Ridge Research and Consulting, LLC.

Chapter 7

Manifold learning of geospatial data via diffusion maps

7.1 Introduction

As discussed in Chapter 5 of this dissertation, feature selection results for environmental sound level modeling are sensitive to details of the problem formulation (e.g., the feature importance metric and hyperparameter search space). However, dimensionality reduction minimizes the curse of dimensionality and decreases computational and data requirements. Unstable feature selection results suggest more sophisticated dimensionality reduction techniques are needed. In particular, this chapter investigates the application of manifold learning techniques via diffusion maps [110] for dimensionality reduction of the geospatial data.

7.1.1 Manifold learning

In mathematics, a manifold is a topological space that is locally isomorphic to Euclidean space. Formally, this means that there exists a bijective function that maps points in the space onto a set of real numbers known as coordinates. A good mental model of a manifold is that of a curved surface embedded in a higher-dimensional space. For example, each point on the surface of a sphere can be uniquely identified by a pair of real numbers, which means that the surface of the sphere forms a two-dimensional manifold embedded in the larger three-dimensional space.

In manifold learning, it is assumed that data are drawn from a manifold (and possibly corrupted by noise). The goal is to identify coordinates to describe points on the manifold from which the data were sampled. This process can identify potentially complex, nonlinear relationships among features. Importantly, manifold learning performs dimensionality reduction on a data set while preserving the intrinsic geometric relationships between neighboring points. Manifold learning is similar to principal component analysis in many ways. The former, however, is able to explicitly learn *nonlinear* spaces, while principal component analysis finds the projection onto the optimal *linear* spaces. Due to high correlations between geospatial features in the geospatial database, it is natural to assume that the geospatial data are constrained to lie on a low-dimensional manifold embedded in the full-featured space.

7.1.2 Additional advantages of manifold learning

In addition to minimizing the curse of dimensionality and reducing computational and data requirements, performing dimensionality reduction via manifold learning may improve optimal experimental design procedures and allow for transferability of models trained only on data from the Contiguous United States (CONUS), or CONUS-based models, to regions Outside of the Contiguous United States (OCONUS).

Clustering methods have been used previously to identify optimal locations for further acoustic data collection [30]. Somewhat like clustering, the manifold coordinates identify spatially separated locations with similar geospatial features. However, the manifold learning approach has several advantages over previous clustering analyses. In particular, clustering requires the selection of an appropriate number of clusters, which is often subjective. Additionally, clustering identifies a discrete number of clusters and pigeon-holes each location into exactly one cluster while manifold coordinates allow for the possibility of smooth transitions between points without forcing points to be categorized as a single type.

Manifold coordinates may also enable OCONUS predictions of environmental sound levels using CONUS-based sound level models. One of the primary challenges of producing OCONUS predictions is that different geospatial data are available in different regions, preventing direct use of CONUS-based models for many OCONUS regions. Unlike previous feature selection which identified a minimal set of *bare* features, the manifold coordinates are *nonlinear combinations* of geospatial features. Machine learning models may be trained on these manifold coordinates to produce environmental sound level models. Although OCONUS regions may not have the same available geospatial data as in CONUS, we need only identify the underlying manifold coordinates of the OCONUS data to then apply CONUS-based models.

More specifically, we can utilize the shared bare geospatial layers of CONUS and OCONUS regions to train machine learning models to map subsets of bare geospatial layers into manifold coordinates. The final model for OCONUS regions will therefore consist of a two-stage process: first, a subset of geospatial features is mapped into manifold coordinates, and then the manifold coordinates are used to predict ambient sound. In this way, models of environmental sound levels may be extended to OCONUS regions in which only a limited subset of geospatial features are available. In practice, any subset of features that map injectively into the manifold coordinates can be used as a reduced feature set. As we consider strategies for

OCONUS geospatial database development, a strategy based on manifold learning will provide maximum flexibility in reducing high-correlated features with low effective dimensionality.

7.1.3 Challenges

There are several potential challenges to performing manifold learning on the CONUS geospatial database. For example, manifold learning is often computationally intensive; therefore, it is not possible to perform manifold learning on all of CONUS. Hence, manifold learning will only be performed on a subset of CONUS data. Care should be taken to downsample in a way that does not introduce bias.

Geospatial data also presents interesting challenges. In particular, if we consider geospatial data drawn from a very small geographic region, it should always appear two-dimensional because it can be mapped to latitude and longitude. However, on larger scales, latitude and longitude will not be manifold coordinates because points that are geographically separated may have very similar geospatial features. As a rough mental model, the data should be approximately a two-dimensional surface that intersects itself in potentially complicated ways.

In mathematical language, geospatial data are drawn from an *immersion* rather than a *manifold*. (An immersion is similar to a manifold except that it could possibly self-intersect. A figure-eight is an example of a one-dimensional immersion, whereas a circle is a one-dimensional manifold.) We believe self-intersections to be important for two related reasons. First, self-intersections are the mathematical reason for needing more features in a predictive model. Next, for the case of geospatial data, self-intersections occur when geographically separated sites have identical geospatial features. For example, sites in distant cities may have the same geospatial features, and by extension (one hopes), similar acoustic environments.

Indeed, at a more fundamental level, this is the reason we expect continental-scale models to be possible.

When applied to data drawn from immersions, manifold learning will learn the effective dimensionality of the minimal manifold into which the immersion can be embedded. We therefore expect the data to exhibit interesting dependence on the scale of the geographic region from which it was drawn. Understanding this scale-dependence will be important both as we continue to collect data across CONUS and as we consider OCONUS predictions on even larger scales.

7.1.4 Chapter outline

Section 2 of this chapter describes the selection of CONUS geospatial data for diffusion maps. This includes the selection of geospatial features and downsampled sites in CONUS. In Section 3, diffusion maps are applied to CONUS geospatial data and we perform some initial analyses on the diffusion coordinates. Section 4 gives concluding remarks and suggests possible future work.

7.2 Determining data for diffusion maps

We decided to use the set of 51 geospatial features described in Appendix A, as this set of features was selected by a human expert to provide maximal information while minimizing error. Features were removed from the initial set of 120 due to large errors or uncertainties, redundancies or large correlations with layers in the set of 51, poor documentation and metadata, or expected lack of physical effect on the soundscape. Geospatial features were scaled based on all CONUS data (rather than training data) for each layer. Scaling methods are described in more detail in Table 7.1.

Table 7.1 Features and scaling methods selected for the application of diffusion maps to CONUS geospatial data.

Feature	Scaling
Elevation	Min-max
Slope	Min-max
PPTSummer	Min-max
PPTWinter	Min-max
TMaxSummer	Min-max
TMaxWinter	Min-max
TMinSummer	Min-max
TMinWinter	Min-max
TdewAvgSummer	Min-max
TdewAvgWinter	Min-max
Barren (200 m, 5 km)	Min-max
Cultivated (200 m, 5 km)	Min-max
Deciduous (200 m, 5 km)	Min-max
Developed (200 m, 5 km)	Min-max
Evergreen (200 m, 5 km)	Min-max
Herbaceous (200 m, 5 km)	Min-max
MixedForest (200 m, 5 km)	Min-max
Shrubland (200 m, 5 km)	Min-max
Water (200 m, 5 km)	Min-max
Wetlands (200 m, 5 km)	Min-max
DistCoast	$\frac{2}{\pi}\arctan(\frac{x}{8})$
DistStream (O1, O3, O4)	$\frac{2}{\pi}\arctan(\frac{x}{2})$
DistAirpHeli	$\frac{2}{\pi}\arctan(\frac{x}{2})$
DistAirpHigh	$\frac{2}{\pi}\arctan(\frac{x}{4})$
DistAirpLow	$\frac{2}{\pi}\arctan(\frac{x}{2})$
DistAirpMod	$\frac{2}{\pi}\arctan(\frac{x}{2})$
DistAirpMoto	$\frac{2}{\pi}\arctan(\frac{x}{2})$
DistMilitary	$\frac{2}{\pi}\arctan(\frac{x}{4})$
DistRailroads	$\frac{2}{\pi}\arctan(\frac{x}{2})$
DistRoadsAll	$\frac{2}{\pi}\arctan(\frac{x}{2})$
DistRoadsMaj	$\frac{2}{\pi}\arctan(\frac{x}{2})$
FlightFreq	Min-max
MilitarySum	Min-max

RddAll (Point, 5 km)	Min-max
RddMajor (Point, 5 km)	Min-max
PopDensity	Min-max
VIIRSMean (270 m)	Min-max

Ideally, we would apply diffusion maps to geospatial data with a 270-m spatial resolution (i.e., the highest spatial resolution available). However, computational constraints prevent us from applying diffusion maps to all geospatial data in CONUS. Preliminary benchmarks suggested we could calculate diffusion maps for approximately 75,000 sites (or ~0.07% of CONUS). Therefore, we needed to determine an appropriate sample of CONUS data on which to apply diffusion maps. In particular, we wanted to identify a sample of CONUS data that would produce an embedding (or diffusion coordinates) similar to that of the true embedding (i.e., the embedding produced using all 270-m CONUS data).

We performed numerical experiments on downsampled regions in Utah and found that the results of diffusion mapping were generally robust to downsampling when data were sampled randomly and equally from each subcluster. (Subclusters are described in Chapter 6 of this dissertation.) In other words, simple random sampling of an equal number of sites from each subcluster resulted in consistent diffusion maps for large and small sampled percentages (i.e., down to 0.07%) of sites. These results suggest that applying diffusion maps with this sampling method to 0.07% of CONUS data should give a good approximation to the true embedding.

7.3 Diffusion maps of CONUS data

We randomly sampled the same number of sites (approximately 4,530) from each of the 16 subclusters described in Chapter 6 and applied diffusion maps to the corresponding geospatial data (51 scaled geospatial features). The total number of sites sampled was 72,538 (close to 0.066% of CONUS). The Nyström extension for out-of-sample points was used to estimate diffusion coordinates for all sites not initially used to generate the mapping [111]. The Nyström extension uses the similarity between out-of-sample input data points and training data points to estimate the diffusion coordinates for the out-of-sample points.

A CONUS map of the first diffusion coordinate is shown in Figure 7.1. Visually, the first diffusion coordinate is similar to the predicted summer daytime A-weighted L_{50}. The correlation coefficient between the first diffusion coordinate and summer daytime A-weighted L_{50} predicted by the ensemble model using the set of 51 scaled geospatial features is 0.876. This correlation coefficient is greater than the correlation between the first diffusion coordinate and any individual geospatial feature in the set of 51 scaled features. This result suggests that the first diffusion coordinate describes trends in the geospatial data that correlate with sound levels. Additionally, the high correlation between the first coordinate and sound level predictions indicates that only a small number of diffusion coordinates may be needed for environmental sound level modeling.

We calculated the correlation coefficients between the 51 scaled geospatial features and each of the first 20 diffusion coordinates. Table 7.2 shows the 15 geospatial features most strongly correlated with the first diffusion coordinate. All of the top 15 correlated features have relatively strong magnitudes of correlation (i.e., greater than 0.5). This result indicates that the first diffusion coordinate describes dominant trends in the data, as expected. The mix of natural and anthropogenic features in Table 7.2 demonstrates that the first diffusion coordinate identifies both natural and anthropogenic trends. In general, the other diffusion coordinates have fewer strongly correlated geospatial features because they describe regions of smaller variance in the underlying manifold.

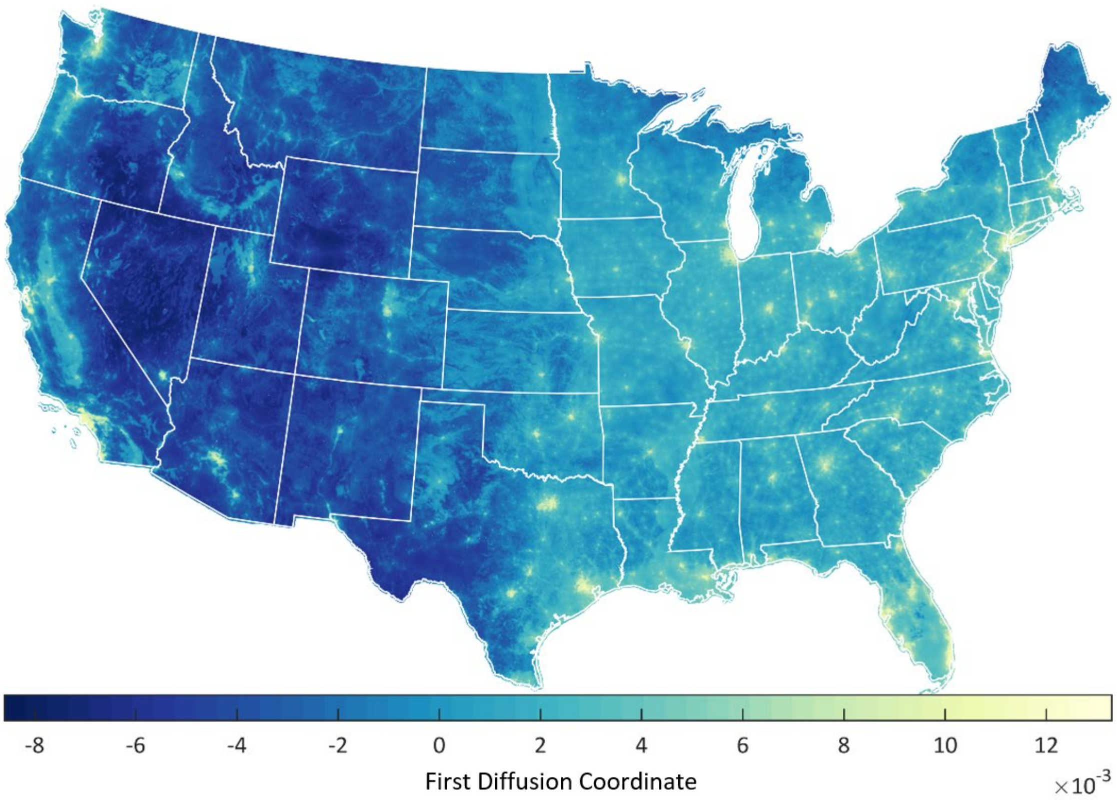

Figure 7.1 CONUS map of the first diffusion coordinate.

Table 7.2 Table of the top 15 correlated geospatial features with the first diffusion coordinate.

Geospatial Feature	Correlation
TdewAvgSummer	0.761
Elevation	-0.749
Shrubland_5000m	-0.713
PPTSummer	0.707
DistAirpHeli	-0.686
Shrubland_200m	-0.672
DistAirpMoto	-0.635
DistRailroads	-0.624
DistAirpMod	-0.612
Developed_5000m	0.571
RddAll_5000m	0.540
FlightFreq_25km	0.535
DistAirpHigh	-0.524
DistAirpLow	-0.519
RddMajor_5000m	0.517

We also compared the distributions of training data values and CONUS values for the first 20 diffusion coordinates. Figure 7.2 shows the distributions of training data values and CONUS values for the first diffusion coordinate. Although the distributions are different, they are more similar than many of the distributions between individual geospatial features in CONUS and the training data. The third and fourth diffusion coordinates show similar agreement between CONUS and training data distributions to the first coordinate. However, the histograms for the other diffusion coordinates show even closer agreement between the distributions of CONUS and training data. These results suggest that diffusion coordinates describe the relationship between training points and prediction sites in CONUS better than the geospatial features.

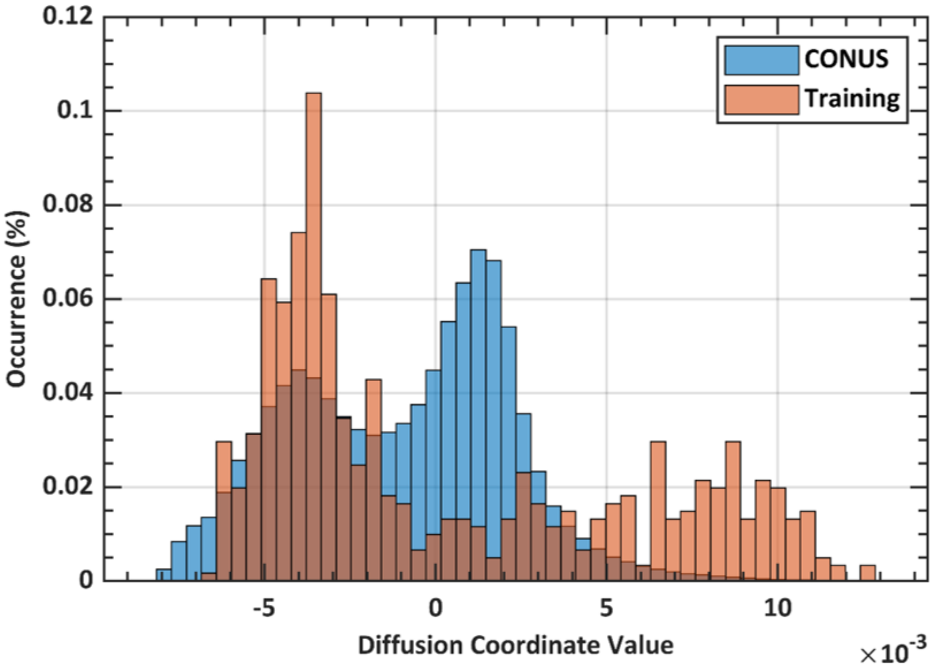

Figure 7.2 Histogram comparing the distributions of values for the first diffusion coordinate in CONUS and the training data.

7.4 Concluding remarks and future work

There are many questions regarding the best use of diffusion coordinates for environmental sound level modeling. For example, it is unclear how many diffusion coordinates should be used; although, we anticipate the optimal number of diffusion coordinates is likely small (i.e., close to 15) because of the results of Chapter 5. Additionally, it is possible that the use of diffusion coordinates (and the number of diffusion coordinates used) may change trends in model predictions as well as estimates of model uncertainty. Because diffusion coordinates attempt to describe the underlying manifold from which data are drawn, a study of changes in model behavior and uncertainty estimates may yield physically meaningful results, which can guide further improvements in the modeling approach and data collection methods.

We expect that in addition to minimizing the curse of dimensionality and reducing computational requirements, CONUS diffusion coordinates may help identify optimal locations for data collection. Because diffusion coordinates minimize redundancies among features, diffusion coordinates may be more effective at identifying geospatial regions which are poorly represented in the training data. Diffusion coordinates may also be used to measure the similarity between geographically distinct sites as well as indicate densely populated areas of the model manifold.

The ability to apply CONUS-based models to OCONUS regions through the use of diffusion coordinates is exciting and may have potentially broad implications. Indeed, the application of CONUS-based models to OCONUS regions not only has implications for environmental sound modeling, but also for the study of model transferability in machine/deep learning. For environmental sound modeling, validation studies of CONUS-based model transferability to OCONUS regions may help distinguish between acoustic trends which either persist across or are unique to certain continents or regions. The success (or failure) of CONUS-based models in OCONUS regions will also have implications for the success of model transferability with limited training data.

Chapter 8

Conclusion

Modeling outdoor acoustic environments on continental scales is a challenging problem with broad potential applications in areas such as public health, social justice, and ecology. In this dissertation, we implemented supervised machine learning models trained on geospatial layers and acoustic data to predict different acoustic metrics across the contiguous United States. Many challenges of continental-scale environmental sound modeling stem from limited availability of acoustic measurements on which to train supervised machine learning algorithms. Within this data-limited regime, models are forced to make predictions in extrapolation regions (i.e., areas of feature space poorly represented by the training data) for which traditional validation metrics do not apply. Therefore, leave-one-out cross-validation predicts overly optimistic expected errors.

Indeed, we observe that leave-one-out cross-validation is a poor indicator of model accuracy across continental-scales. This motivates better uncertainty quantification of model predictions. We use ensemble models to improve uncertainty estimates; however, we do not quantify all sources of uncertainty and therefore, present only minimum estimates of uncertainty. More specifically, we use an ensemble of models from different machine learning classes to estimate the structural uncertainty in model predictions. Additionally, we utilize bootstrap sampling to better estimate uncertainty due to computational and statistical uncertainties. We emphasize that these minimal uncertainty estimates do not account for all sources of uncertainty (e.g., we do not quantify uncertainty due to errors in geospatial layers or measured acoustic data) and strongly caution the use of environmental sound level model predictions without a thorough consideration of uncertainty estimates.

However, evaluation of model predictions for the summer daytime A-weighted L_{50} and summer hourly spectra indicate that models have learned some relationships between geospatial data and the acoustic environment. We observe that leave-one-out predictions for the summer hourly spectra are realistic and often have the same spectral shape as measured levels, even if overall levels do not agree. Although models struggle to identify locations for which bird and/or insect contributions should be present, geospatial data may not provide sufficient detail and resolution for accurate modeling of bird/insect activity. The relative success of models for leave-one-out predictions of different hours and frequency bands indicates that supervised machine learning methods may be successful if sufficient training data are obtained. Much of the training data are obtained from similar environments, which may explain the relative success of models for leave-one-out cross-validation.

In this dissertation, we also investigated methods of dimensionality reduction for the geospatial data. We found that feature selection results were unstable to changes in the problem formulation, suggesting more sophisticated methods of dimensionality reduction were required. In particular, we used diffusion maps to identify a reduced set of coordinates by which the geospatial data can be described. These reduced coordinates provide a means of making predictions for regions outside of the contiguous United States as well as identifying undersampled geospatial environments.

Future work may include training environmental sound level models on diffusion coordinates and applying and validating models in regions outside the contiguous United States. Additionally, identifying optimal methods for acoustic data collection as well as incorporating physics-based modeling approaches may improve model accuracy and uncertainty estimates. Indeed, validating models on measured and physics-based model outputs will likely identify geospatial regions in which models struggle, and which are therefore good candidates for acoustic data collection. It is possible that such methods will also identify potential geospatial information that may benefit models. Indeed, it is unlikely the geospatial data are sufficient to accurately characterize all acoustic environments. Therefore, additional geospatial layers may be needed for accurate modeling within such environments.

In conclusion, continental-scale environmental sound level modeling is a challenging problem with broad potential applications. Many challenges are a result of limited acoustic data on which to train machine learning models, which force models to extrapolate when making predictions on large geographic scales. We have laid the foundation for uncertainty quantification in continental-scale environmental sound level models using two types of ensembles. We have also shown that feature selection techniques are not appropriate for sound level models, and have suggested an alternative dimensionality reduction method.

The results of the validation studies and uncertainty estimates reported in this dissertation are deeply concerning and bring into question the validity of studies dependent upon the accuracy of the NPS published sound level map [24]. Despite concerns for the accuracy of continental-scale sound level predictions, this dissertation also demonstrates that models may perform well when training data are well-sampled for a geographic region or geospatial environment. Therefore, we expect that continental-scale environmental sound level models may achieve relatively high accuracy in most geospatial environments given sufficient representative training data.

Appendix A

Geospatial and acoustic database

A.1 Geospatial data

Data from 120 geospatial layers for the contiguous United States were considered for ensemble models (see Table A.1). We note that many of the layers considered are the same as those considered or used by Mennitt et al. [1] Additionally, Table A.1 is organized in the same manner as Table 1 of [1] to facilitate comparisons between geospatial layers considered for Mennitt et al.'s model and the ensemble model. The geospatial layers all have a 270-m spatial resolution and can be categorized into one of six categories: topography, climate, land cover, hydrology, anthropogenic, and position.

All geospatial data were obtained from the National Park Service Natural Sounds and Night Skies and Inventory and Monitoring Divisions database [62, 63] with the exception of the AviationNoise and RoadNoise layers which were obtained from the U.S. Department of Transportation's Bureau of Transportation Statistics [112].

Table A.1 Geospatial layers for the contiguous United States, their area of analysis, description, and units. This table follows the same organization of Table 1 from [1].

Variable	Area of Analysis	Description	Units
Topography			
Elevation	Point	Digital elevation, height above sea level	m
Slope	Point	Rate of change of elevation	Degrees
Climate			
PPTSummer	Point	10-year average summer precipitation	mm
PPTWinter	Point	10-year average winter precipitation	mm
PPTAnnual	Point	10-year average yearly precipitation	mm
TMaxSummer	Point	10-year average summer maximum temperature	°C
TMaxWinter	Point	10-year average winter maximum temperature	°C
TMaxAnnual	Point	10-year average yearly maximum temperature	°C
TMinSummer	Point	10-year average summer minimum temperature	°C
TMinWinter	Point	10-year average winter minimum temperature	°C
TMinAnnual	Point	10-year average yearly minimum temperature	°C
TdewAvgSummer	Point	10-year average summer minimum dew point	°C
TdewAvgWinter	Point	10-year average winter maximum dew point	°C
TdewAvgAnnual	Point	10-year average yearly minimum dew point	°C
Land Cover			
Barren	200 m, 5 km	Proportion of barren land cover	%
Cultivated	200 m, 5 km	Proportion of cultivated land cover	%
Deciduous	200 m, 5 km	Proportion of deciduous forest land cover	%
Developed	200 m, 5 km	Proportion of developed land cover	%
Evergreen	200 m, 5 km	Proportion of evergreen forest land cover	%
Forest	200 m, 5 km	Proportion of forest land cover	%
Herbaceous	200 m, 5 km	Proportion of herbaceous land cover	%
MixedForest	200 m, 5 km	Proportion of mixed forest land cover	%
Shrubland	200 m, 5 km	Proportion of shrubland land cover	%
Water	200 m, 5 km	Proportion of water (only) land cover	%
Wetlands	200 m, 5 km	Proportion of wetlands land cover	%
Hydrology			
DistCoast	Point	Distance to nearest coastline	m
DistStreamO	Point	Distance to nearest stream with Strahler order greater than 1, 3, or 4	m
DistWaterBody	Point	Distance to nearest body of water	m
Anthropogenic			
AviationNoise	Point	Aviation model noise	dB

Built	200 m, 5 km	Degree of human modification from built land use	Ratio
Commercial	200 m, 5 km	Degree of human modification from commercial land use	Ratio
Cropland	200 m, 5 km	Degree of human modification from cropland land use	Ratio
DistAirpHeli	Point	Distance to nearest heliport	m
DistAirpHigh	Point	Distance to nearest high-volume airport	m
DistAirpLow	Point	Distance to nearest low-volume airport	m
DistAirpMod	Point	Distance to nearest moderate-volume airport	m
DistAirpMoto	Point	Distance to nearest motorized airport	m
DistAirpSea	Point	Distance to nearest seaplane airport	m
DistMilitary	Point	Distance to nearest military flight path	m
DistRailroads	Point	Distance to nearest rail line	m
DistRoadsAll	Point	Distance to nearest road (all roads)	m
DistRoadsMaj	Point	Distance to nearest road (major roads)	m
Extractive	200 m, 5 km	Degree of human modification from extractive land use	Ratio
ExurbanHigh	200 m, 5 km	Degree of human modification from high exurban land use	Ratio
ExurbanLow	200 m, 5 km	Degree of human modification from low exurban land use	Ratio
FlightFreq	25 km	Total weekly flight observations	Count
Grazing	200 m, 5 km	Degree of human modification from grazing land use	Ratio
Industrial	200 m, 5 km	Degree of human modification from industrial land use	Ratio
Institutional	200 m, 5 km	Degree of human modification from institutional land use	Ratio
MilitarySum	40 km	Sum of designated military flight paths	Count
Mining	200 m, 5 km	Degree of human modification from mining land use	Ratio
Park	200 m, 5 km	Degree of human modification from park land use	Ratio
Pasture	200 m, 5 km	Degree of human modification from pasture land use	Ratio
PhysicalAccess	Point	Travel time given transportation infrastructure and off-trail permeability	Ratio
PopDensity	Point	2015 estimated population density data	persons/km^2
RddAll	Point, 5 km	Road density, sum of road lengths (all roads) divided by area of interest	km/km^2
RddMajor	Point, 5 km	Road density, sum of road lengths (major roads only) divided by area of interest	km/km^2
RecCon	200 m, 5 km	Degree of human modification from recreation-conservation land use	Ratio

RoadNoise	Point	Department of Transportation road model noise	dB
Suburban	200 m, 5 km	Degree of human modification from suburban land use	Ratio
Timber	200 m, 5 km	Degree of human modification from timber land use	Ratio
Transportation	200 m, 5 km	Degree of human modification from transportation land use	Ratio
UrbanHigh	200 m, 5 km	Degree of human modification from high urban land use	Ratio
UrbanLow	200 m, 5 km	Degree of human modification from low urban land use	Ratio
VIIRS	270 m, 1080 m, 4320 m, 17280 m, 69120 m	Maximum, mean, and minimum upward radiance at night	nW/cm^2/sr
WaterHum	200 m, 5 km	Degree of human modification from water land use	Ratio
WaterNat	200 m, 5 km	Degree of human modification from natural water land use	Ratio
Wet	200 m, 5 km	Degree of human modification from wet land use	Ratio
Position			
Latitude	Point	Latitude value of raster cell in decimal degrees	Degrees
Longitude	Point	Longitude value of raster cell in decimal degrees	Degrees

Following a quality review of these 120 layers, a reduced set of 51 geospatial layers were selected for environmental sound level modeling (see Table A.2). Features were removed due to large errors or uncertainties, significant correlations with other features, poor documentation, or the expected lack of a predictive relationship to the acoustic environment (e.g., latitude and longitude).

Table A.2 Subset of 51 geospatial layers used for environmental sound level modeling.

Barren (200 m)	DistCoast	Herbaceous (5 km)	Slope
Barren (5 km)	DistMilitary	MilitarySum (40 km)	TdewAvgSummer
Cultivated (200 m)	DistRailroads	MixedForest (200 m)	TdewAvgWinter
Cultivated (5 km)	DistRoadsAll	MixedForest (5 km)	TMaxSummer
Deciduous (200 m)	DistRoadsMaj	PopDensity	TMaxWinter
Deciduous (5 km)	DistStreamO1	PPTSummer	TMinSummer
Developed (200 m)	DistStreamO3	PPTWinter	TMinWinter
Developed (5 km)	DistStreamO4	RddAll	VIIRSMean (270 m)
DistAirpHeli	Elevation	RddAll (5 km)	Water (200 m)
DistAirpHigh	Evergreen (200 m)	RddMajor	Water (5 km)
DistAirpLow	Evergreen (5 km)	RddMajor (5 km)	Wetlands (200 m)
DistAirpMod	FlightFreq (25 km)	Shrubland (200 m)	Wetlands (5 km)
DistAirpMoto	Herbaceous (200 m)	Shrubland (5 km)	

A.2 Acoustic data

The acoustic data contain acoustic measurements of the summer daytime A-weighted L_{50} from 492 geographically unique sites. Data are compiled from multiple sources including Blue Ridge Research and Consulting, LLC's internal acoustic data, the NPS acoustic database [62, 63], and a 1974 Environmental Protection Agency study [64]. Note that only summer daytime and hourly acoustic data were utilized in this dissertation (i.e., data from other seasons/times were not used).

Appendix B

Additional figures for subclustering analysis

B.1 Silhouette scores and distances to centroids

This section provides figures of silhouette analyses for each cluster from the 8-cluster model described in Chapter 6 (Figures B.1-B.8; left). Additionally, the estimated probability densities of the distance (as measured by the Euclidean norm) between instances within each subcluster and the correspond initial cluster centroid are plotted. These estimated probability densities are overlaid with similar plots using the identified subcluster centroids (rather than the initial cluster centroid; Figures B.1-B.8; right). These figures may help determine if subclustering into two subclusters is beneficial for a given application and cluster. In particular, when adding a second centroid (i.e., subclustering into two clusters) significantly moves both distance distributions to the left, it is more likely that subclustering is beneficial for further describing the data.

Lastly, we plot individual maps of subcluster assignments in the contiguous United States for each of the original eight clusters (Figure B.9).

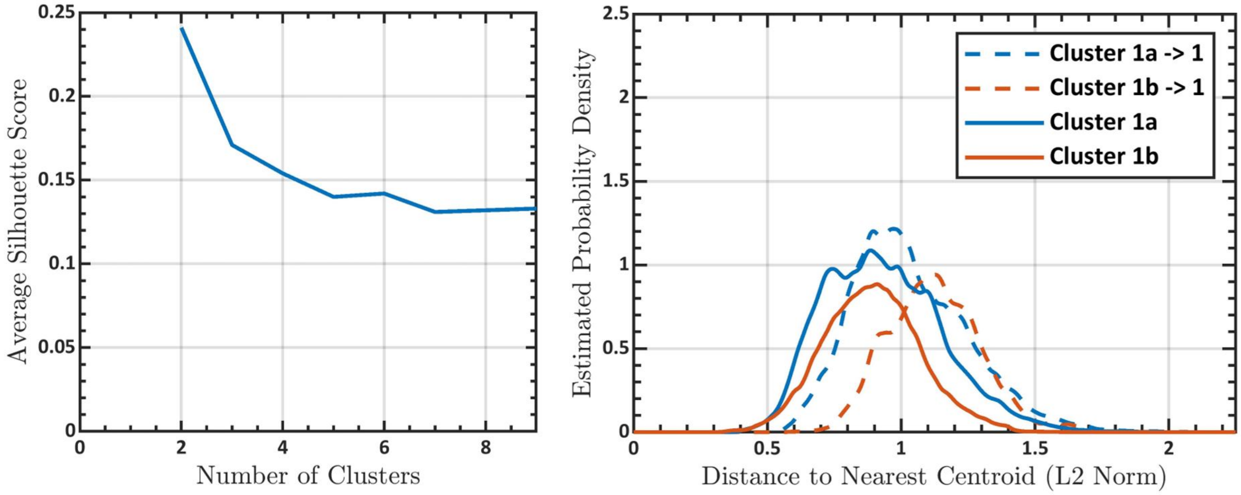

Figure B.1 Left: average silhouette scores as Cluster 1 is subclustered. Right: estimated probability density of the distance from data within subclusters 1a and 1b to Cluster 1's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

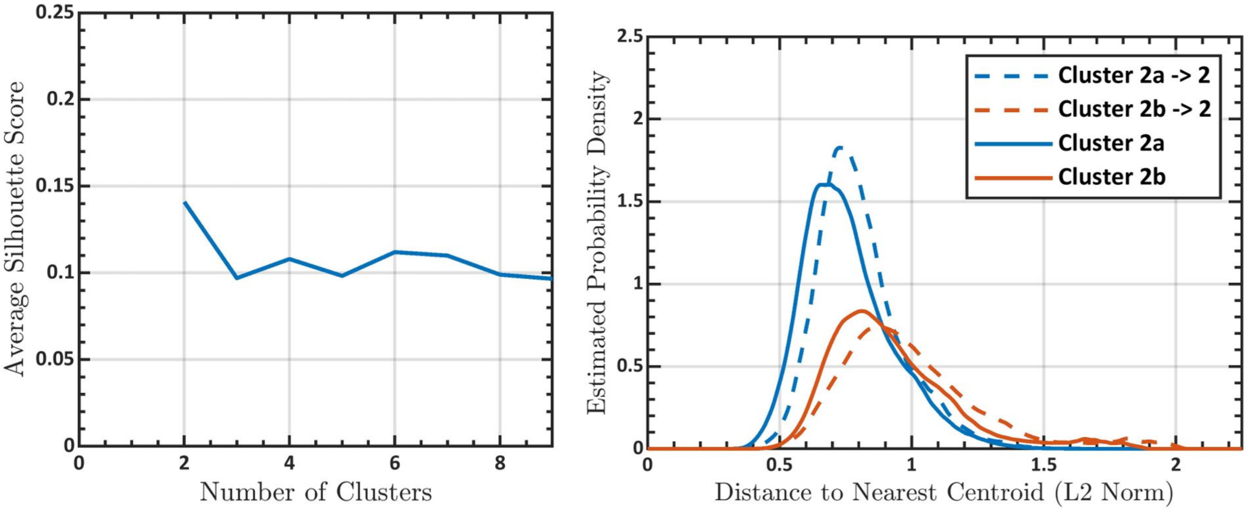

Figure B.2 Left: average silhouette scores as Cluster 2 is subclustered. Right: estimated probability density of the distance from data within subclusters 2a and 2b to Cluster 2's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

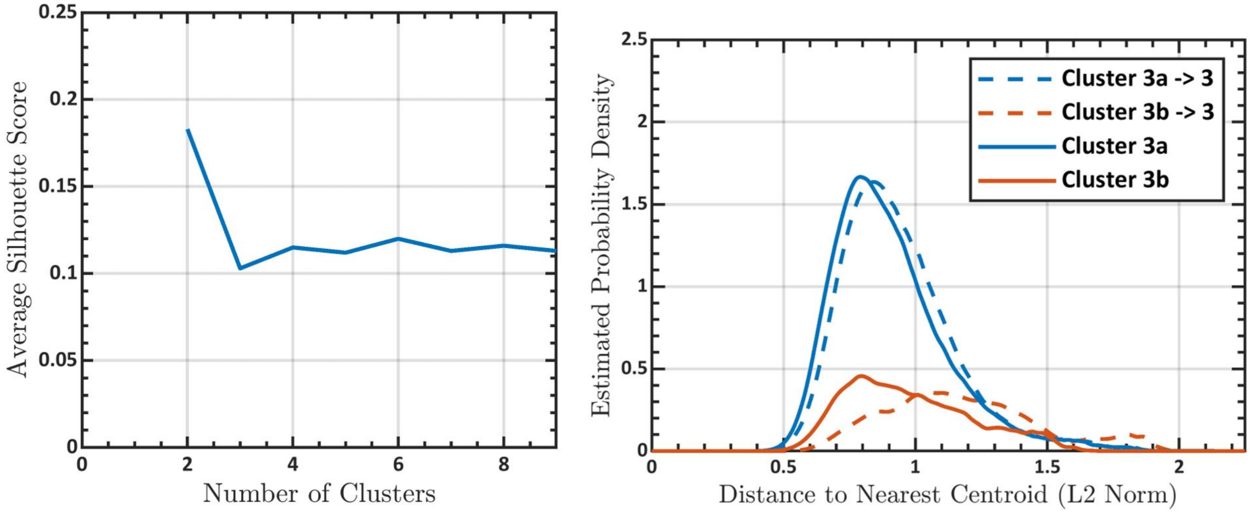

Figure B.3 Left: average silhouette scores as Cluster 3 is subclustered. Right: estimated probability density of the distance from data within subclusters 3a and 3b to Cluster 3's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

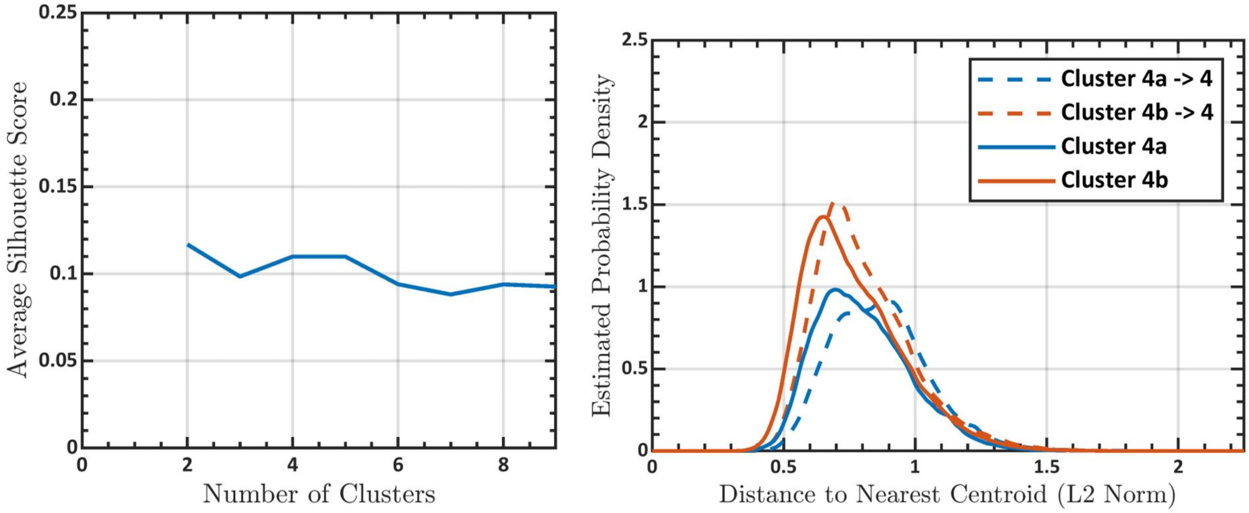

Figure B.4 Left: average silhouette scores as Cluster 4 is subclustered. Right: estimated probability density of the distance from data within subclusters 4a and 4b to Cluster 4's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

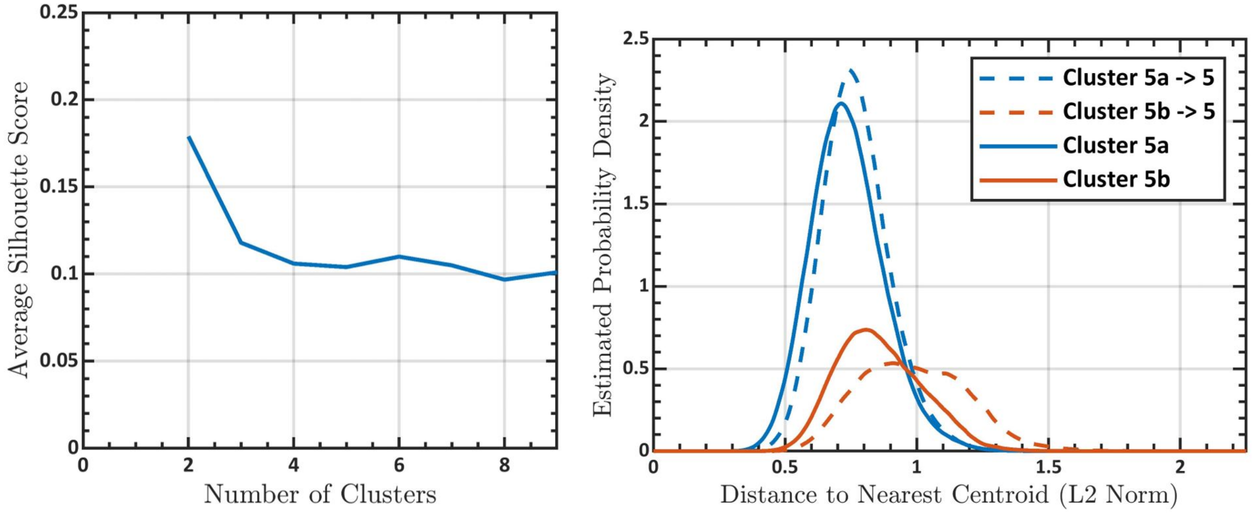

Figure B.5 Left: average silhouette scores as Cluster 5 is subclustered. Right: estimated probability density of the distance from data within subclusters 5a and 5b to Cluster 5's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

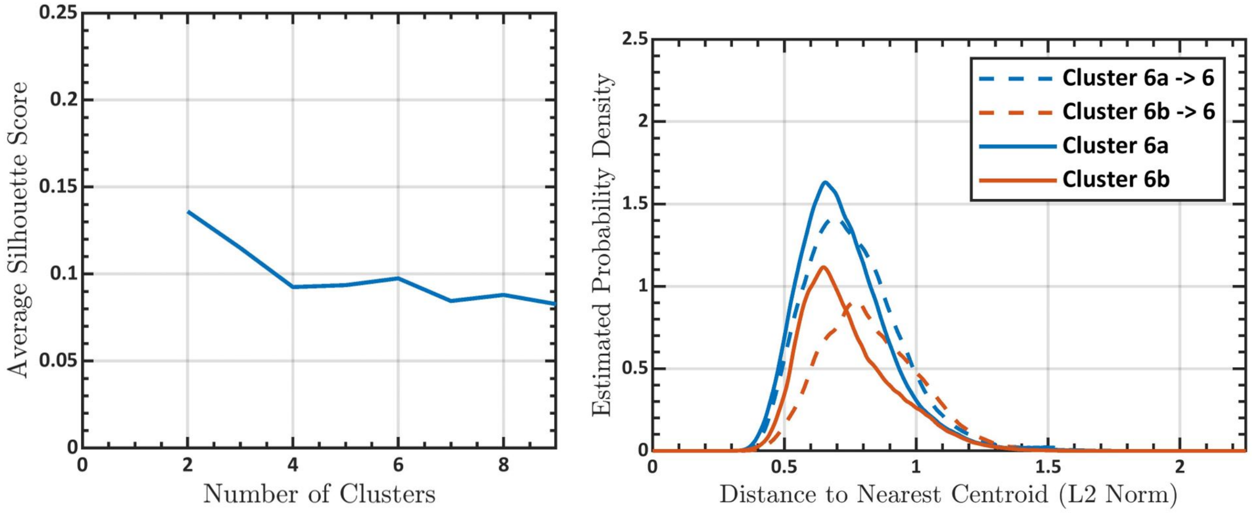

Figure B.6 Left: average silhouette scores as Cluster 6 is subclustered. Right: estimated probability density of the distance from data within subclusters 6a and 6b to Cluster 6's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

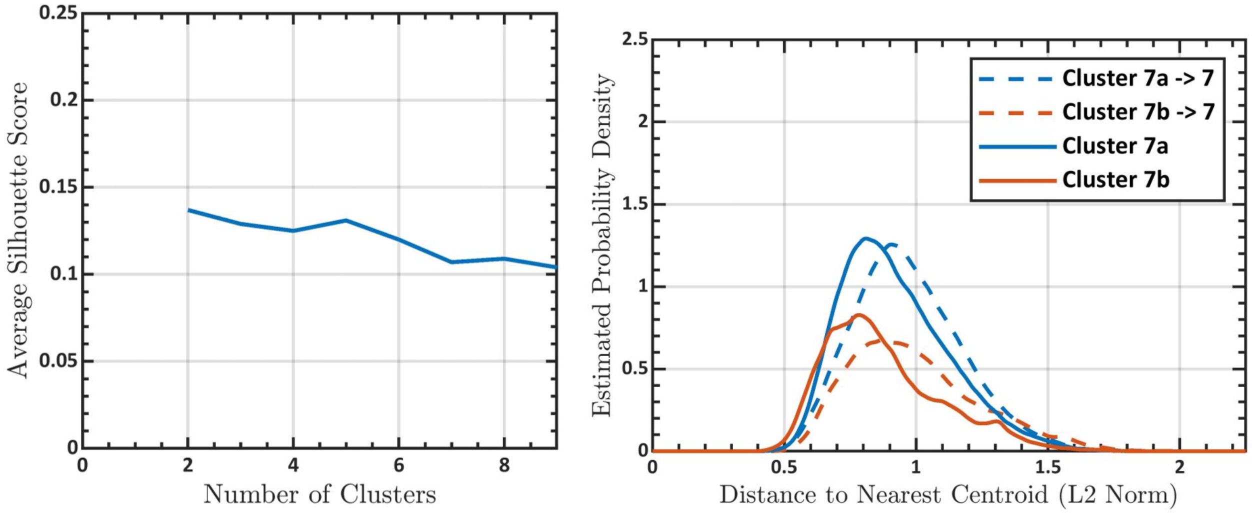

Figure B.7 Left: average silhouette scores as Cluster 7 is subclustered. Right: estimated probability density of the distance from data within subclusters 7a and 7b to Cluster 7's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

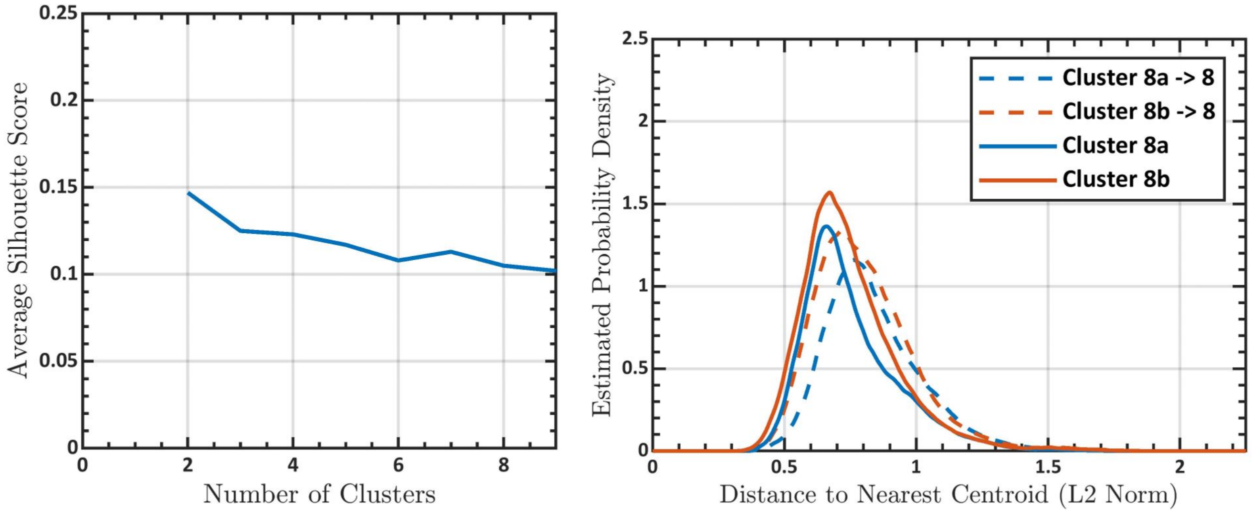

Figure B.8 Left: average silhouette scores as Cluster 8 is subclustered. Right: estimated probability density of the distance from data within subclusters 8a and 8b to Cluster 8's centroid (dashed lines) and their corresponding subcluster centroids (solid lines).

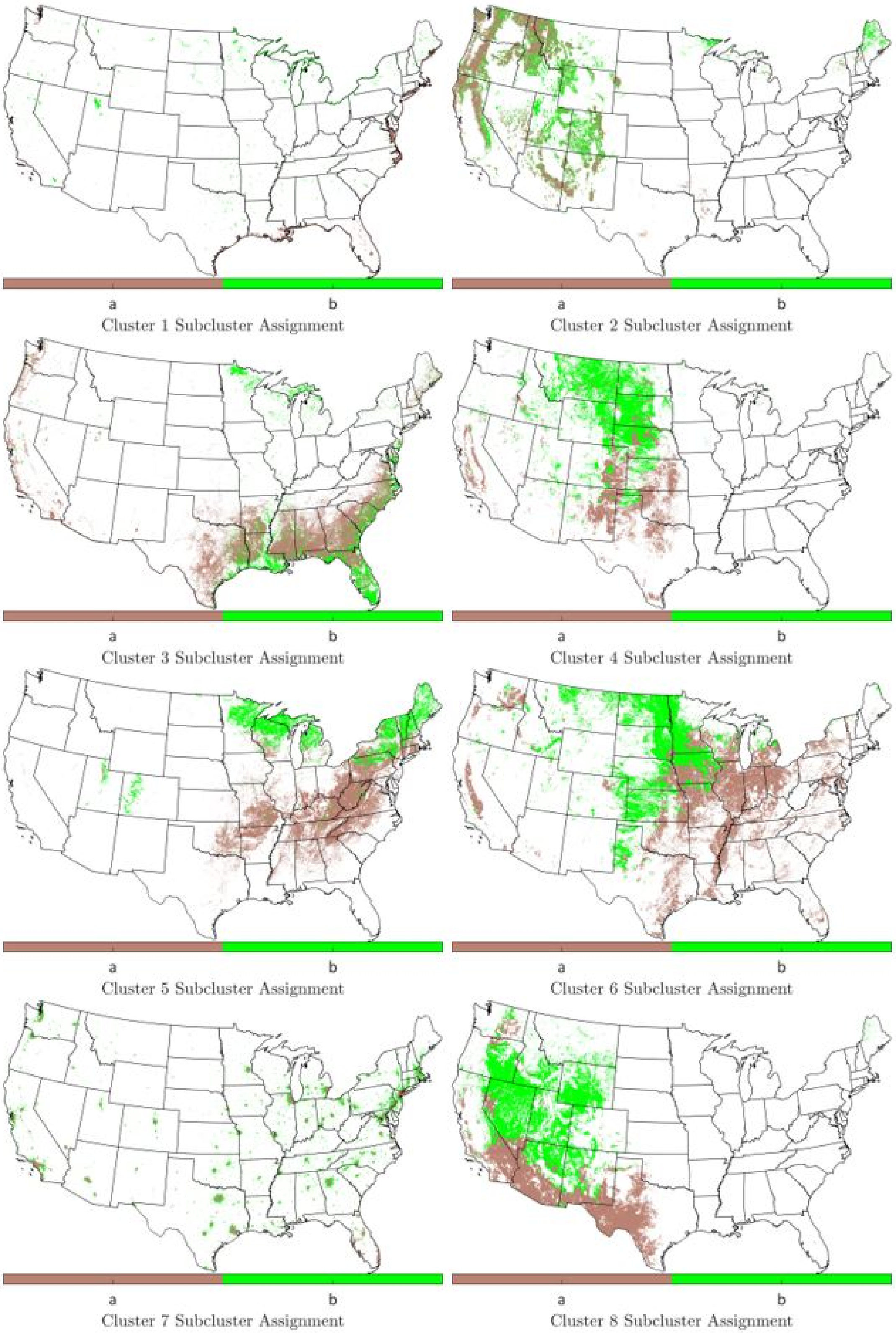
a
b
Cluster 1 Subcluster Assignment
a
b
Cluster 2 Subcluster Assignment
a
b
Cluster 3 Subcluster Assignment
a
b
Cluster 4 Subcluster Assignment
a
b
Cluster 5 Subcluster Assignment
a
b
Cluster 6 Subcluster Assignment
a
b
Cluster 7 Subcluster Assignment
a
b
Cluster 8 Subcluster Assignment

Figure B.9 CONUS maps of two subclusters for each of the original eight clusters.

Bibliography

[1] D. Mennitt and K. Fristrup, “Influential factors and spatiotemporal patterns of environmental sound levels in the contiguous United States,” *Noise Control Engr. J.*, vol. 64, no. 3, pp. 342–353, 2016.

[2] M. E. Beutel, C. Junger, E. M. Klein, P. Wild, K. Lackner, M. Blettner, H. Binder, M. Michal, J. Wiltink, E. Brahler, and T. Munzel, “Noise annoyance is associated with depression and anxiety in the general population - the contribution of aircraft noise,” *PLoS One*, vol. 11, no. 5, 2016.

[3] E. Orban, K. McDonald, R. Sutcliffe, B. Hoffmann, K. B. Fuks, N. Dragano, A. Viehmann, R. Erbel, K. H. Jöckel, N. Pundt *et al.*, “Residential road traffic noise and high depressive symptoms after five years of follow-up: results from the Heinz Nixdorf recall study,” *Environ. Health Perspect.*, vol. 124, no. 5, pp. 578–585, 2016.

[4] T. Münzel, M. R. Miller, M. Sørensen, J. Lelieveld, A. Daiber, and S. Rajagopalan, “Reduction of environmental pollutants for prevention of cardiovascular disease: itâ€™s time to act,” *Eur. Heart J.*, vol. 41, no. 41, p. 3989, 2020.

[5] E. Van Kempen, M. Casas, G. Pershagen, and M. Foraster, “WHO environmental noise guidelines for the European region: a systematic review on environmental noise and cardiovascular and metabolic effects: a summary,” *Int. J. Environ. Res. Public Health*, vol. 15, no. 2, p. 379, 2018.

[6] T. Bodin, M. Albin, J. Ardö, E. Stroh, P. Östergren, and J. Björk, “Road traffic noise and hypertension: Results from a cross-sectional public health survey in southern Sweden,” *Environ. Health*, vol. 8, no. 1, p. 38, 2009.

[7] L. Jarup, W. Babisch, D. Houthuijs, M. Dudley, G. Pershager, K. Katsouyanni, E. Cadum, P. Savigny, I. Seiffert, W. Swart, and O. Breugelmans, “Hypertension and exposure to noise near airports: the HYENA study,” *Environ. Health Perspect.*, vol. 116, no. 3, p. 329, 2008.

[8] T. Münzel, F. P. Schmidt, S. Steven, J. Herzog, A. Daiber, and M. Sørensen, “Environmental noise and the cardiovascular system,” *J. Am. Coll. Cardiol.*, vol. 71, no. 6, pp. 688–697, 2018.

[9] Z. J. Andersen, J. T. Jørgensen, L. Elsborg, S. N. Lophaven, C. Backalarz, J. E. Laursen, T. H. Pedersen, M. K. Simonsen, E. V. Bräuner, and E. Lynge, “Long-term exposure to road traffic noise and incidence of breast cancer: a cohort study,” *Breast Cancer Res.*, vol. 20, no. 1, pp. 1–13, 2018.

[10] T. Münzel, T. Gori, W. Babisch, and M. Basner, “Cardiovascular effects of environmental noise exposure,” *Eur. Heart J.*, vol. 35, no. 13, pp. 829–836, 2014.

[11] D. Fink, “Ambient Noise Is “The New Secondhand Smoke”,” *J. Acoust. Soc. Am.*, vol. 146, no. 4, p. 2835, 2019.

[12] "EEA Report No 22/2019 Environmental noise in Europe - 2020," https://-www.eea.europa.eu/publications/environmental-noise-in-europe.

[13] C. D. Francis, P. Newman, B. D. Taff, C. White, C. A. Monz, M. Levenhagen, A. R. Petrelli, L. C. Abbott, J. Newton, S. Burson, C. B. Cooper, K. M. Fristrup, C. J. McClure, D. Mennitt, M. Giamellaro, and J. R. Barber, "Acoustic environments matter: Synergistic benefits to humans and ecological communities," *Environ. Manage.*, vol. 203, pp. 245–254, 2017.

[14] P. Morano, F. Tajani, F. Di L., and M. Darò, "Economic Evaluation of the Indoor Environmental Quality of Buildings: The Noise Pollution Effects on Housing Prices in the City of Bari (Italy)," *Buildings*, vol. 11, no. 5, p. 213, 2021.

[15] E. Murphy and E. King, *Environmental noise pollution: Noise mapping, public health, and policy*. Newnes, 2014.

[16] B. M. Siemers, P. Stilz, and H. U. Schnitzler, "The acoustic advantage of hunting at low heights above water: behavioural experiments on the European â€˜trawlingâ€™bats Myotis capaccinii, M. dasycneme and M. daubentonii," *J. Exp. Biol.*, vol. 204, no. 22, pp. 3843–3854, 2001.

[17] C. D. Francis, C. P. Ortega, and A. Cruz, "Noise pollution changes avian communities and species interactions," *Curr. Biol.*, vol. 19, no. 16, pp. 1415–1419, 2009.

[18] D. S. Proppe, C. B. Sturdy, and C. C. St. Clair, "Anthropogenic noise decreases urban songbird diversity and may contribute to homogenization," *Global Change Biol.*, vol. 19, no. 4, pp. 1075–1084, 2013.

[19] N. Rako-Gospic and M. Picciulin, "Únderwater noise: Sources and effects on marine life," in *World Seas: An Environmental Evaluation*. Elsevier, 2019, pp. 367–389.

[20] K. M. Parris, M. Velik-Lord, and J. M. A. North, “Frogs call at a higher pitch in traffic noise,” *Ecol. Soc.*, vol. 14, no. 1, 2009.

[21] J. W. Sun and P. M. Narins, “Anthropogenic sounds differentially affect amphibian call rate,” *Biol. Conserv.*, vol. 121, no. 3, pp. 419–427, 2005.

[22] D. J. Mennitt, K. Fristrup, K. Sherrill, and L. Nelson, “Mapping sound pressure levels on continental scales using a geospatial sound model,” *InterNoise13*, 2013.

[23] D. Mennitt, K. Sherrill, and K. Fristrup, “A geospatial model of ambient sound pressure levels in the contiguous United States,” *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2746–2764, May 2014.

[24] National Park Service, “Geospatial sound modeling,” https://irma.nps.gov/Datastore/-Reference/Profile/2217356, accessed January, 2020.

[25] Y. Liu, S. Goudreau, T. Oiamo, D. Rainham, M. Hatzopoulou, H. Chen, H. Davies, M. Tremblay, J. Johnson, A. Bockstael, and T. Leroux, “Comparison of land use regression and random forests models on estimating noise levels in five Canadian cities,” *Environ. Pollut.*, vol. 256, p. 113367, 2020.

[26] I. Aguilera, M. Foraster, X. Basagaña, E. Corradi, A. Deltell, X. Morelli, H. C. Phuleria, M. S. Ragettli, M. Rivera, A. Thomasson, and R. Slama, “Application of land use regression modelling to assess the spatial distribution of road traffic noise in three European cities,” *J. Exposure Sci. Environ. Epidemiol.*, vol. 25, no. 1, pp. 97–105, 2015.

[27] D. Xie, Y. Liu, and J. Chen, “Mapping urban environmental noise: A land use regression method,” *Environ. Sci. Technol.*, vol. 45, no. 17, pp. 7358–7364, 2011.

[28] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws,

Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

[30] B. A. Butler, K. Pedersen, C. Maekawa, K. L. Gee, M. K. Transtrum, M. M. James, and A. R. Salton, "K-means clustering of inputs to a geospatial model for optimizing acoustic data collection," in *Proceedings of Meetings on Acoustics 176ASA*, vol. 35, no. 1. Acoustical Society of America, 2018, p. 055008.

[31] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*. Philadelphia, PA, USA: SIAM Comput. Sci. & Eng. Series, 2014.

[32] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *J. R. Statist. Soc.*, vol. 63, pp. 425–464, 2001.

[33] 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," *Environ. Health Perspect.*, vol. 125, no. 7, p. 077017, 2017.

[34] B. S. Johnson, K. M. Malecki, P. E. Peppard, and K. M. Beyer, "Exposure to neighborhood green space and sleep: evidence from the Survey of the Health of Wisconsin," *Sleep Health*, vol. 4, no. 5, pp. 413–419, 2018.

[35] K. E. Rudolph, A. Shev, D. Paksarian, K. R. Merikangas, D. J. Mennitt, P. James, and J. A. Casey, "Environmental noise and sleep and mental health outcomes in a nationally

representative sample of urban US adolescents," *Environ. Epidemiol. (Philadelphia, Pa.)*, vol. 3, no. 4, 2019.

[36] R. T. Buxton, M. F. McKenna, D. J. Mennitt, K. Fristrup, K. Crooks, L. Angeloni, and G. Wittemyer, "Noise pollution is pervasive in US protected areas," *Science*, vol. 356, no. 6337, pp. 531–533, 2017.

[37] R. T. Buxton, B. M. Seymoure, J. White, L. M. Angeloni, K. R. Crooks, K. Fristrup, M. F. McKenna, and G. Wittemyer, "The relationship between anthropogenic light and noise in US national parks," *Landsc. Ecol.*, vol. 35, pp. 1371–1384, 2020.

[38] W. L. Rice, P. Newman, Z. D. Miller, and B. D. Taff, "Protected areas and noise abatement: A spatial approach," *Landsc. Urban Plan.*, vol. 194, p. 103701, 2020.

[39] B. T. Klingbeil, F. A. L. Sorte, C. A. Lepczyk, D. Fink, and C. H. Flather, "Geographical associations with anthropogenic noise pollution for North American breeding birds," *Global Ecol. Biogeogr.*, vol. 29, no. 1, pp. 148–158, 2020.

[40] M. Senzaki, J. R. Barber, J. N. Phillips, N. H. Carter, C. B. Cooper, M. A. Ditmer, K. M. Fristrup, C. J. McClure, D. J. Mennitt, L. P. Tyrrell, J. Vukomanovic, A. A. Wilson, and C. D. Fancis, "Sensory pollutants alter bird phenology and fitness across a continent," *Nature*, vol. 587, no. 7835, pp. 605–609, 2020.

[41] W. J. Doebler, "Estimated ambient sonic boom metric levels and X-59 signal-to-noise ratios across the USA," in *Proceedings of Meetings on Acoustics 179ASA*, vol. 42, no. 1. Acoustical Society of America, 2020, p. 040003.

[42] N. Japkowicz and M. Shah, *Evaluating learning algorithms: A classification perspective*. Cambridge University Press, 2011.

[43] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*, 2013, pp. 115–123.

[44] J. Bergstra, "Hyperopt: Distributed asynchronous hyperparameter optimization in python," http://jaberg.github.com/hyperopt, 2013.

[45] National Park Service, "NPS Director's Order #47: Soundscape preservation and noise management," 2000.

[46] National Academy of Engineering, "Protecting national park soundscapes," 2013.

[47] C. I. Merchan, L. Diaz-Balteiro, and M. Soliño, "Noise pollution in national parks: Soundscape and economic valuation," *Landsc. Urban Plan.*, vol. 123, pp. 1–9, 2014.

[48] D. Weinzimmer, P. Newman, D. Taff, J. Benfield, E. Lynch, and P. Bell, "Human responses to simulated motorized noise in national parks," *Leis. Sci.*, vol. 36, pp. 251–267, 2014.

[49] H. Slabbekorn and W. Halfwerk, "Behavioural ecology: Noise annoys at community level," *Curr. Biol.*, vol. 19, no. 16, pp. R693–R695, 2009.

[50] B. C. Pijanowski, L. T. 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, no. 3, pp. 203–216, 2011.

[51] E. P. Derryberry, R. M. Danner, J. E. Danner, G. E. Derryberry, J. N. Phillips, S. E. Lipshutz, K. Gentry, and D. A. Luther, "Patterns of song across natural and anthropogenic soundscapes suggest that white-crowned sparrows minimize acoustic masking and maximize signal content," *PLoS One*, vol. 11, no. 4, e0154456, 2016.

[52] G. Buscaino, M. Ceraulo, N. Pieretti, V. Corrias, A. Farina, F. Filiciotto, V. Maccarrone, R. Grammauta, F. Caruso, A. Giuseppe, and S. Mazzola, “Temporal patterns in the soundscape of the shallow waters of a Mediterranean marine protected area,” *Sci. Rep.*, vol. 6, no. 34230, 2016.

[53] P. A. Hastings and A. Sirovic, “Soundscapes offer unique opportunities for studies of fish communities,” in *PNAS USA*, vol. 112, no. 19, 2015, pp. 5866–5867.

[54] L. Ruppé, G. Clément, A. Herrel, L. Ballesta, T. Décamps, L. Kéver, and E. Parmentier, “Environmental constraints drive the partitioning of the soundscape in fishes,” in *PNAS USA*, vol. 112, no. 19, 2015, pp. 6092–6097.

[55] F. Bertucci, E. Parmentier, G. Lecellier, A. D. Hawkins, and D. Lecchini, “Acoustic indices provide information on the status of coral reefs: An example from Moorea Island in the South Pacific,” *Sci. Rep.*, vol. 6, no. 33326, 2016.

[56] S. M. Haver, H. Klinck, S. L. Nieukirk, H. Matsumoto, R. P. Dziak, and J. L. MiksisOlds, “The not-so-silent world: Measuring Arctic, Equitorial, and Antarctic soundscapes in the Atlantic Ocean,” *Deep Sea Res. Part I: Oceanogr. Res. Pap.*, 2017.

[57] S. Goutte, A. Dubois, and F. Legendre, “The importance of ambient sound level to characterise anuran habitat,” *PLoS One*, vol. 8, no. 10, e78020, 2013.

[58] K. Kaliski, E. Duncan, and J. Cowan, “Community and regional noise mapping in the United States,” *Sound Vib.*, vol. 41, no. 9, p. 12, 2007.

[59] D. owicki and S. Piotrowska, “Monetary valuation of road noise. Residential property prices as an indicator of the acoustic climate quality,” *Ecol. Indic.*, vol. 52, pp. 472–479, 2015.

[60] Boeing, "Airports with noise and emissions restrictions," https://www.boeing.com/-commercial/noise/list.page.

[61] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. on Knowl. and Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[62] 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, 2015.

[63] National Park Service, "Data store," https://irma.nps.gov/DataStore/.

[64] W. J. Galloway, K. M. Eldred, and M. A. Simpson, "Population distribution of the United States as a function of outdoor noise level, Volume 2," *U.S. EPA (Washington, D.C.)*, 1974.

[65] K. Pedersen, M. K. Transtrum, K. L. Gee, B. A. Butler, M. M. James, and A. R. Salton, "Machine learning-based ensemble model predictions of outdoor ambient sound levels," in *Proceedings of Meetings on Acoustics 176ASA*, vol. 35, no. 1. Acoustical Society of America, 2018, p. 022002.

[66] S. P. Shukla, S. K. Yadav, B. Lohani, S. Biswas, S. N. Behra, N. Singh, and N. K. Singh, "Characterization of traffic noise for a typical Indian road crossing," *Curr. Sci. India*, pp. 1193–1201, 2012.

[67] S. A. Stansfeld, B. Berglund, C. Clark, I. Lopez-Barrio, P. Fischer, E. Öhrström, M. . M. Haines, J. Head, S. Hygge, I. van Kamp, and B. F. Berry, "Aircraft and road traffic noise and children's cognition and health: a cross-national study," *Lancet*, vol. 365, pp. 1942–1949, 2005.

[68] S. P. Banbury, W. J. Macken, S. Tremblay, and D. M. Jones, “Auditory distraction and short-term memory: Phenomena and practical implications,” *Hum. Factors*, vol. 43, no. 1, pp. 12–29, 2001.

[69] D. J. Mennitt and K. M. Fristrup, “Influential vactors and spatiotemporal patterns of environmental sound levels,” *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 250, no. 5, pp. 2029–2040, 2015.

[70] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton University Press, 2015, vol. 2045.

[71] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, “A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction,” *JASTT*, vol. 1, no. 2, pp. 56–70, 2020.

[72] G. Hooker and L. Mentch, “Please stop permuting features: An explanation and alternatives,” *arXiv preprint arXiv*, p. 1905.03151, 2019.

[73] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, “Conditional variable importance for random forests,” *BMC Bioinf.*, vol. 9, no. 1, p. 307, 2008.

[74] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, October 2001.

[75] ——, “Manual on setting up, using, and understanding random forests v3. 1,” *Statistics Department University of California Berkeley, CA, USA*, vol. 1, p. 58, 2002.

[76] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution,” *BMC Bioinf.*, vol. 8, no. 1, p. 25, 2007.

[77] G. Louppe, *Understanding random forests*. Cornell University Library, 2014.

[78] ——, “Understanding random forests: From theory to practice,” *arXiv preprint arXiv:1407.7502*, 2014.

[79] M. Gevrey, I. Dimopoulos, and S. Lek, “Review and comparison of methods to study the contribution of variables in artificial neural network models,” *Ecol. Modell.*, vol. 160, pp. 249–264, 2003.

[80] S. Olson, P. Reid *et al.*, *Protecting national park soundscapes*. National Academies Press, 2013.

[81] L. M. Kuehne and J. D. Olden, “Military flights threaten the wilderness soundscapes of the Olympic Peninsula, Washington,” *Northwest Sci.*, vol. 94, no. 2, pp. 188–202, 2020.

[82] K. Kalawapudi, T. Singh, J. Dey, R. Vijay, and R. Kumar, “Noise pollution in Mumbai Metropolitan Region (MMR): An emerging environmental threat,” *Environ. Monit. Assess.*, vol. 192, no. 2, pp. 1–20, 2020.

[83] P. S. Sharma, S. G. Raju, M. Murali, and C. S. R. K. Prasad, “Assessment of noise level due to vehicular traffic at Warangal city, India,” *Int. J. Environ. Pollut.*, vol. 30, no. 1, pp. 137–153, 2007.

[84] R. T. Buxton, M. F. McKenna, D. Mennitt, E. Brown, K. Fristrup, K. R. Crooks, L. M. Angeloni, and G. Wittemyer, “Anthropogenic noise in US national parks–sources and spatial extent,” *Front. Ecol. Environ.*, vol. 17, no. 10, pp. 559–564, 2019.

[85] N. F. Jones, L. Pejchar, and J. M. Kiesecker, “The energy footprint: how oil, natural gas, and wind energy affect land for biodiversity and the flow of ecosystem services,” *BioScience*, vol. 65, no. 3, pp. 290–301, 2015.

[86] A. V. Moudon, “Real noise from the urban environment: how ambient community noise affects health and what can be done about it,” *Am. J. Prev. Med.*, vol. 37, no. 2, pp. 167–171, 2009.

[87] J. Hays, M. McCawley, and S. B. Shonkoff, “Public health implications of environmental noise associated with unconventional oil and gas development,” *Sci. Total Environ.*, vol. 580, pp. 448–456, 2017.

[88] R. Mehta, R. Zhu, and A. Cheema, “Is noise always bad? Exploring the effects of ambient noise on creative cognition,” *J, Consum. Res.*, vol. 39, no. 4, pp. 784–799, 2012.

[89] P. Lercher, G. W. Evans, M. Meis, and W. W. Kofler, “Ambient neighbourhood noise and children’s mental health,” *Occup. Environ. Med.*, vol. 59, no. 6, pp. 380–386, 2002.

[90] D. K. Delaney, T. G. Grubb, P. Beier, L. L. Pater, and M. H. Reiser, “Effects of helicopter noise on Mexican spotted owls,” *The Journal of Wildlife Management*, pp. 60–76, 1999.

[91] M. L. Leonard and A. G. Horn, “Ambient noise and the design of begging signals,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 272, no. 1563, pp. 651–656, 2005.

[92] J. L. Quinn, M. J. Whittingham, S. J. Butler, and W. Cresswell, “Noise, predation risk compensation and vigilance in the chaffinch Fringilla coelebs,” *J. Avian Biol.*, vol. 37, no. 6, pp. 601–608, 2006.

[93] L. A. Rabin, R. G. Coss, and D. H. Owings, “The effects of wind turbines on antipredator behavior in California ground squirrels (Spermophilus beecheyi),” *Biol. Conserv.*, vol. 131, no. 3, pp. 410–420, 2006.

[94] J. P. Swaddle and L. C. Page, “High levels of environmental noise erode pair preferences in zebra finches: implications for noise pollution,” *Anim. Behav.*, vol. 74, no. 3, pp. 363–368, 2007.

[95] M. I. Herrera-Montes and T. M. Aide, “Impacts of traffic noise on anuran and bird communities,” *Urban Ecosyst.*, vol. 14, no. 3, pp. 415–427, 2011.

[96] C. R. Kight and J. P. Swaddle, “How and why environmental noise impacts animals: an integrative, mechanistic review,” *Ecol. Lett.*, vol. 14, no. 10, pp. 1052–1061, 2011.

[97] H. E. Ware, C. J. W. McClure, J. D. Carlisle, and J. R. Barber, “A phantom road experiment reveals traffic noise is an invisible source of habitat degradation,” *PNAS*, vol. 112, no. 39, pp. 12105–12109, 2015.

[98] J. Sueur, “Cicada acoustic communication: potential sound partitioning in a multispecies community from Mexico (Hemiptera: Cicadomorpha: Cicadidae),” *Biol. J. Linn. Soc.*, vol. 75, no. 3, pp. 379–394, 2002.

[99] J. A. Simmons, E. G. Wever, and J. M. Pylka, “Periodical cicada: sound production and hearing,” *Science*, vol. 171, no. 3967, pp. 212–213, 1971.

[100] A. Leopold and A. E. Eynon, “Avian daybreak and evening song in relation to time and light intensity,” *The Condor*, vol. 63, no. 4, pp. 269–293, 1961.

[101] K. S. Berg, R. T. Brumfield, and V. Apanius, “Phylogenetic and ecological determinants of the neotropical dawn chorus,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 273, no. 1589, pp. 999–1005, 2006.

[102] J. Kragh, “Pilot study on railway noise attenuation by belts of trees,” *J. Sound Vib.*, vol. 66, no. 3, pp. 407–415, 1979.

[103] D. Aylor, “Noise reduction by vegetation and ground,” *J. Acoust. Soc. Am.*, vol. 51, no. 1B, pp. 197–205, 1972.

[104] Z. Zhang, Q. Li, H. Nan, X. Yang *et al.*, “Study on noise attenuation of green belts in plain area.” *Forest Research, Beijing*, vol. 30, no. 2, pp. 329–334, 2017.

[105] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[106] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Duborg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *JMLR*, vol. 12, pp. 2825–2830, 2011.

[107] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[108] P. Bholowalia and A. Kumar, “EBK-means: A clustering technique based on elbow method and k-means in WSN,” *International Journal of Computer Applications*, vol. 105, no. 9, 2014.

[109] H. P. Kunc and R. Schmidt, “The effects of anthropogenic noise on animals: a meta-analysis,” *Biol. Lett.*, vol. 15, no. 11, p. 20190649, 2019.

[110] R. R. Coifman and S. Lafon, “Diffusion maps,” *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.

[111] A. W. Long and A. L. Ferguson, “Landmark diffusion maps (L-dMaps): Accelerated manifold learning out-of-sample extension,” *Appl. Comput. Harmon. Anal.*, vol. 47, no. 1, pp. 190–211, 2019.

[112] Bureau of Transportation Statistics, United States Department of Transportaion, "National Transportation Noise Map," https://maps.dot.gov/BTS/-NationalTransportationNoiseMap/.