Machine learning-based ensemble model predictions of outdoor ambient sound levels

Katrina Pedersen, Mark K. Transtrum, Kent L. Gee, Brooks A. Butler, Michael M. James, and Alexandria R. Salton

Citation: Proc. Mtgs. Acoust. **35**, 022002 (2018); doi: 10.1121/2.0001056
View online: https://doi.org/10.1121/2.0001056
View Table of Contents: https://asa.scitation.org/toc/pma/35/1
Published by the Acoustical Society of America

Volume 35 http://acousticalsociety.org/

176th Meeting of Acoustical Society of America
2018 Acoustics Week in Canada
Victoria, Canada
5-9 Nov 2018

Computational Acoustics: Paper 2pPA6

Machine learning-based ensemble model predictions of outdoor ambient sound levels

Katrina Pedersen, **Mark K. Transtrum, Kent L. Gee,** and **Brooks A. Butler**
Physics and Astronomy, Brigham Young University, Provo, UT, 84602; katrina.pedersen@gmail.com; mktranstrum@byu.edu; kentgee@byu.edu; brooks.butler93@gmail.com

Michael M. James and Alexandria R. Salton
Blue Ridge Research and Consulting, LLC, Asheville, NC, 28801; michael.james@blueridgeresearch.com; Alex.Salton@blueridgeresearch.com

Outdoor ambient sound levels can be predicted from machine learning-based models derived from geospatial and acoustic training data. To improve modeling robustness, median predicted sound levels have been calculated from an ensemble of tuned models from different supervised machine learning modeling classes. The ensemble is used to predict ambient sound levels throughout the contiguous United States. The training data set consists of 607 unique sites, where various acoustic metrics, such as overall daytime L_{50} levels and one-third octave frequency band levels, have been obtained. Data for 117 geospatial features, which include metrics such as distance to the nearest road or airport, are used. The spread in the ensemble provides an estimate of the modeling accuracy. Results of an initial leave-one-out and leave-four-out validation study are presented.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0001056

1. INTRODUCTION

A. GEOSPATIAL ACOUSTICS

An environment's soundscape is a collection of all the ambient noises in a geographic area. Modeling an environment's soundscape is a challenging problem because ambient noise is the accumulation of many effects, including diverse sources, barriers to propagation, etc., and can vary with time of day or season. However, the ability to accurately characterize a region's soundscape has potentially broad applications.

The study of natural soundscapes 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.[1,2] The natural soundscape is an important part of a visitor's experience,[2–4] and unnatural contributions to the soundscape, such as recreational motorized noise, have been shown to negatively impact visitor experience.[5]

Accurate characterization of soundscapes is also important to ecological studies. Soundscapes play an integral role in a species' habitat, particularly for animals that respond to sound such as birds,[6–9] marine life,[10–14] and anurans (i.e., frogs and toads).[15] Changes in a soundscape may affect animal behavior by inhibiting communication or masking predator noise, and may be accompanied by other changes in behavior such as times of activity or communication patterns.[8,9,11,12,14] The natural acoustic complexity of a region (i.e., not including anthropogenic noise) correlates with higher biodiversity.[13,16]

In areas dominated by anthropogenic sources, sound levels correlate with depression and anxiety,[17] as well as hypertension.[18–20] Public health studies have found that increased noise may be associated with changes in blood pressure, heart rate, and stress.[18,20] Elevated sound levels also affect mental health, impairing reading comprehension, recognition memory, and motivation in children.[21] In adults, sound that varies significantly in pitch, timbre, or tempo over time has been shown to impair cognitive function.[22] Accurate models of anthropogenic soundscapes will be increasingly important to public health studies as data continue to accumulate.[20] An accurate soundscape model also may hold commercial applications for real estate and urban development and have implications for social justice.

This paper describes an ambient soundscape model developed using machine learning algorithms that take geospatial data as inputs. Previously, Mennitt et al.[23–25] used machine learning to relate geospatial features and acoustic metrics. Elsewhere, linear and nonlinear land-use regression models have been used to map urban environmental noise.[26] In our team's approach we used an ensemble of machine-learned models to estimate both the ambient sound levels as well as the uncertainty in the model predictions. In addition to quantifying model confidence, uncertainty estimates provide additional insights into the behavior of the model and will be useful to guide future acoustic data collection efforts. Our team found that with limited acoustic measurement data available, model uncertainty can be high, but that with targeted data collection strategies, there is the potential for improvement.

B. VALIDATION AND 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*[27] for further explanation than what is provided in this paper of validation methods in machine learning.) 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 Leave-One-Out (LOO), k-fold, or a holdout validation method. A broader question is that of transferability, i.e., the ability of a model to accurately make predictions for novel inputs that are statistically different from the training set.

In LOO cross validation, one data point is omitted from the training set and used to validate the learned model that is based on the remainder of the training set. The procedure is then repeated for each data point and statistics are collected on the predicted errors for each data point. 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 for each data point); however, when data are limited this is a reasonable tradeoff.

When large amounts of labeled data are available, it is common to use holdout validation methods. In these approaches, the data is partitioned into a training and validation set. However, for limited data sets, as is presently the case for ambient soundscapes, each data point may contain unique information about the input/output relationship. Randomly omitting instances from the training set 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.

Validation 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. The ability of a model to make accurate predictions for inputs that are statistically different from those on which it is trained is called *transferability*. Transferability estimation is an important, open problem in machine learning.[28] One of the main objectives of this paper is to explore the transferability of machine-learned models for ambient soundscapes.

Our team's approach is inspired by techniques from the uncertainty quantification community. Uncertainty Quantification (UQ) has existed as long as probability and statistics and is the science of identifying, quantifying, and reducing uncertainty when predicting Quantities of Interest (QoIs).[29] 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.[29] Aleatoric uncertainty, or statistical uncertainty, is inherent to a problem, and hence cannot be reduced and is generally represented in terms of probabilities.[29] Epistemic, or systematic, uncertainty originates from an incomplete knowledge or missing physics in a model and can be reduced through better modeling methods.[29] Following Kennedy et al.,[30] 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, our team 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.[30]

Because our team used a data-driven, machine learning approach, our model necessarily omitted 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. 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 form an equivalence class. Although each of these functions is consistent with the training data, they may disagree in their predictions under novel conditions. Our team lacks confidence in our model's predictions because we do not know which model from the equivalence class is correct. Therefore, our team's goal is to characterize the range of predictions in this equivalence class of functions.

C. PAPER OVERVIEW

In this study, our team chose a single function from each of six classes of machine learning models. Models from each model class were trained to predict ambient sound levels throughout the Contiguous United States (CONUS). From each class, our team selected the best model (as measured by the LOO cross validation error) and considered their acoustic predictions on new spatial and frequency domains.

The range of predictions from the ensemble of models serves as a surrogate for the accuracy of a single model. Because our team's ensemble consisted of the single best model from each machine learning class, we interpreted 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. In principle, this ensemble could be extended to account for other types of uncertainty. The ensemble not only provides a method of quantifying uncertainty, but also aids in directing further data collection efforts. Our team demonstrate the performance of our approach on a leave-four-out validation study.

2. METHODS

A. DATA

A machine learning model was developed using a database of geospatial and acoustic data points (hereafter referred to as the "soundscape database").

The soundscape database's geospatial data contains CONUS layers for 117 geospatial features from the NPS Natural Sounds and Night Skies and Inventory and Monitoring Divisions database.[31,32] The geospatial features can

be classified into six categories: topography, climate, land cover, hydrology, anthropogenic, and position. An example CONUS layer for the mean upward radiance at night (using a 270 meter area of analysis) is shown in Figure 1.

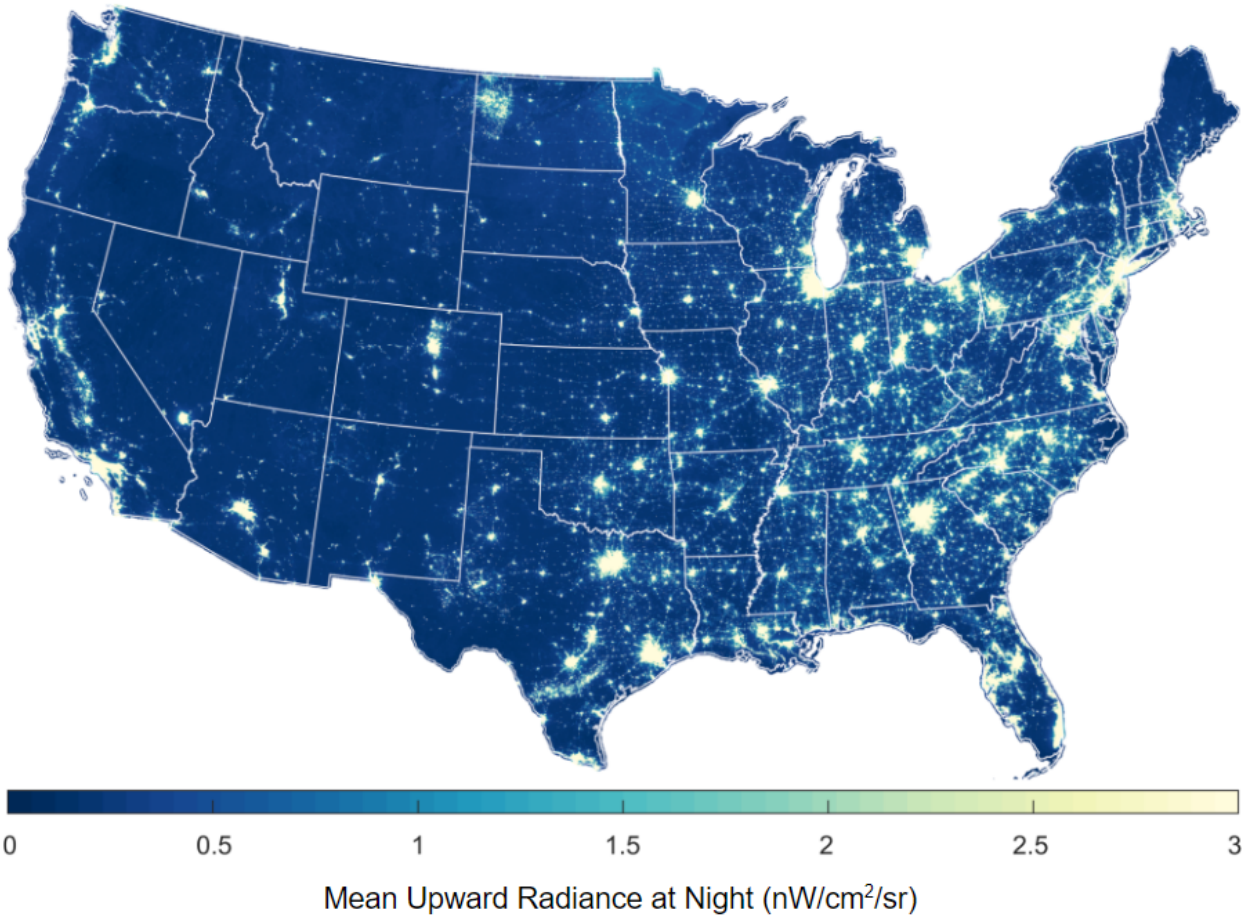

Figure 1: Mean upward radiance at night with a 270 meter area of analysis.

The soundscape database's acoustic data contains measurements from 607 distinct training sites; compiled from multiple sources including Blue Ridge Research and Consulting, LLC's internal soundscape data, the NPS soundscape database,[31,32] and a 1974 Environmental Protection Agency (EPA) study.[33] The acoustic measurements from each site are summarized on a seasonal basis using several statistical measures including the NN% time exceeded level (L_{NN}; e.g., L_{50} and L_{90}) and the equivalent sound level (L_{eq}). These acoustic metrics were integrated over three durations: daytime (7 AM to 7 PM), nighttime (7 PM to 7 AM), and hourly (024 hour). Spectral L_{NN} data were also calculated on a one-third octave band daytime and nighttime basis. Note, measurement data during all four seasons is not available at every site. Therefore, the model results are presented for the summer acoustic metrics as the largest number of acoustic training sites (502 of 607) contain summer data.

The raw acoustic measurement data are the result of considerable effort in terms of number of recorded hours. However, the summary statistics on which our team performed machine learning 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.[34,35] Although the formalism of machine learning can be applied to our data set, it is unclear whether the predictions will be accurate away from the training sites. In this limited data regime, statistical validation measures (such as LOO) do not reflect the actual confidence our team has in model predictions and more sophisticated uncertainty quantification techniques are needed.

B. COMPUTATIONAL METHODS

Our team implemented a computational pipeline to facilitate the machine learning process, summarized in Figure 2. First, we load training data for all acoustic metrics and geospatial features. A subset of the geospatial features is selected for learning, along with target acoustic metrics, i.e., QoIs. All data are scaled (normalized) to remove artifacts of measurement units. Our team used a "standard scaler," which normalizes each feature to have zero mean and a standard deviation of one. After this initial pre-processing, models from a selected set of machine learning classes are trained. The result of the pipeline is one or more models trained on the selected QoIs using the specified geospatial features.

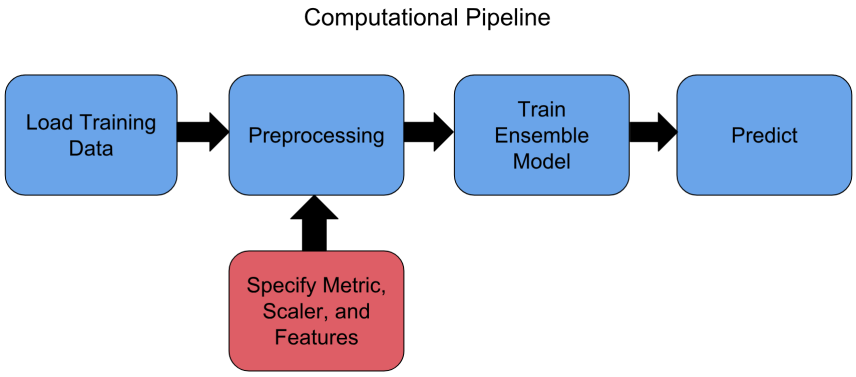

Figure 2: Flowchart of the computational pipeline.

This pipeline enabled our team to explore different validation metrics and a wide variety of machine learning models. As an initial investigation, we explored six machine learning model classes: Gradient Boosted Regression trees (GBRs), Neural Networks (NNs), K-Nearest Neighbors (KN), Support Vector machines (SVs), Kernel Ridge regression models (KRs), and Gaussian Process Regression models (GPRs). Algorithms for the six machine learning model classes were implemented using the Python library scikit-learn.[36] To compare initial model performance, LOO cross validation was used to calculate the Root-Mean-Square Error (RMSE) and Median Absolute Deviation (MAD) for each model. Hyperparameters were optimized for each model to minimize the MAD LOO cross validation errors.

3. RESULTS

A. MODEL SELECTION AND PREDICTIONS

A comparison of the prediction errors from the six model classes is presented in Table 1. The residuals given by the LOO cross validation errors are non-Gaussian and the MAD is typically about half the value of the RMSE. 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.[24,25] The results in Table 1 show that the MAD and RMSE of the LOO cross validation errors are comparable for each model class and the difference in errors is statistically indistinguishable. Therefore, each of the model classes give an adequate fit to the available training data.

Our team discovered that using an ensemble of the models, defined as the median predicted value of the six optimized model classes, reduces the model's sensitivity to outliers. The ensemble model predictions appear more physically accurate in the extrapolation regimes.

The ensemble model's summer daytime L_{50} predictions are presented in Figure 3.

B. UNCERTAINTY ESTIMATES

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

Table 1: Fit and LOO cross validation errors for six different machine learning models.

Model Class	Leave-One-Out MAD (dBA)	Leave-One-Out RMSE (dBA)	Fit MAD (dBA)	Fit RMSE (dBA)
GBR	3.5	6.0	0.08	1.2
NN	3.7	6.3	3.4	5.7
KR	3.6	6.3	0.3	1.4
KN	3.7	6.6	0.0	1.2
GPR	3.6	6.2	2.1	4.0
SV	3.4	6.2	1.3	4.2

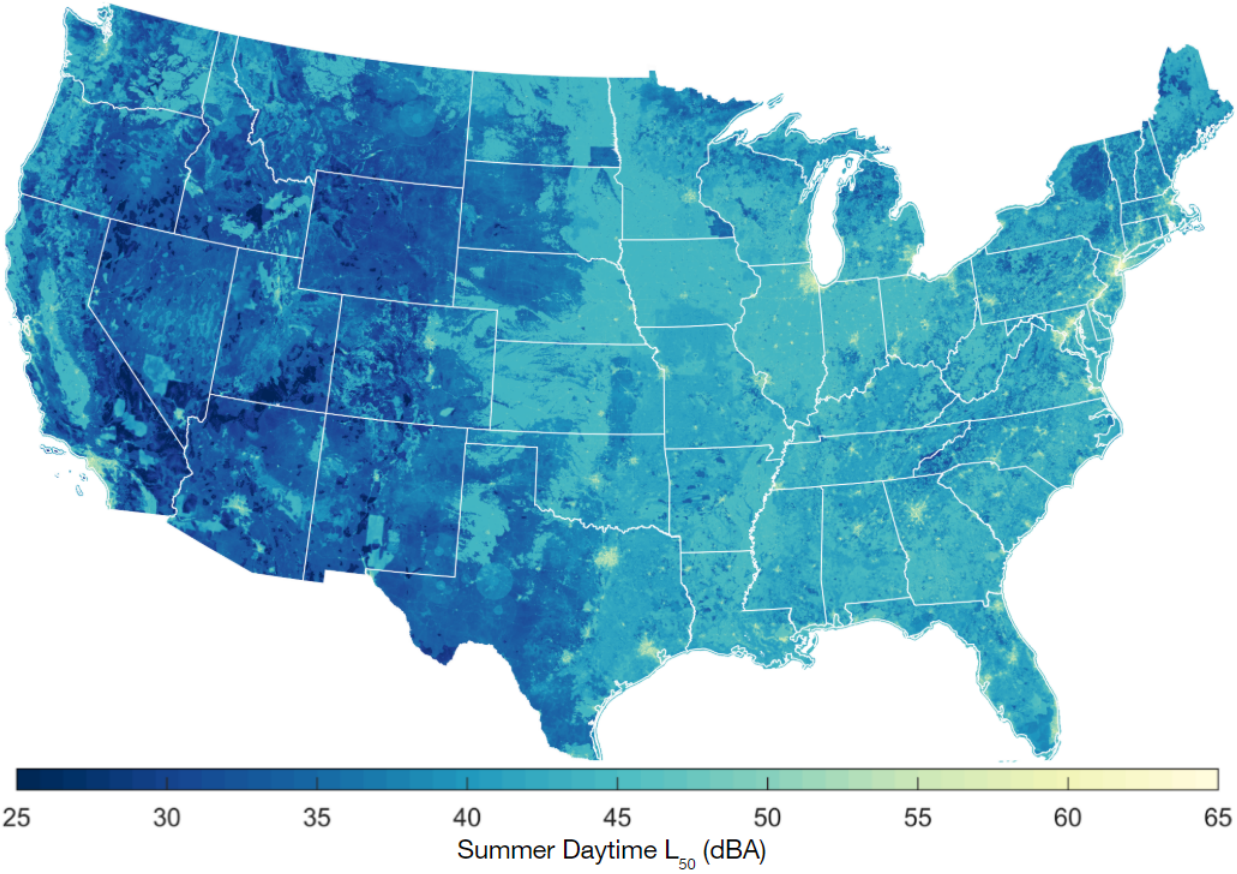

Figure 3: Ensemble model predictions for the A-weighted summer daytime L_{50} for CONUS.

The standard deviation of the ensemble model's predictions was used as a measure of uncertainty. The variation among ensemble members is a surrogate estimate of the accuracy for any individual model. Because our team's 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, called "structural uncertainty." Figure 4 shows the standard deviation of ensemble model predictions. The confidence intervals provided by the standard deviation of ensemble predictions are similar to those generated by a GPR model, which also uses an ensemble of models, or functions.

The uncertainty estimates are useful for guiding data collection efforts. When selecting geographic sites for additional training data, one should target areas with large uncertainty as these are the sites that will provide the most constraint on the model predictions. Presumably these sites have geospatial features that are underrepresented by the

sites in the training set.

The uncertainty analysis can also guide feature reduction strategies. For example, as can be seen in Figure 4, sites with large uncertainty include areas around Iowa and Illinois, and the unusual large circular regions in western Texas, eastern New Mexico, and eastern Montana. These circular regions correlate with a single geospatial feature: the maximum upward radiance at night using a 69,120 meter area of analysis. Note, the geospatial feature set also includes the maximum upward radiance at night using a 270 meter area of analysis. This suggests that the 69,120 meter area of analysis data may not be adding any predictive power to the model while increasing the extrapolation error. In future studies, this feature could be removed to improve model accuracy when extrapolating outside the training set.

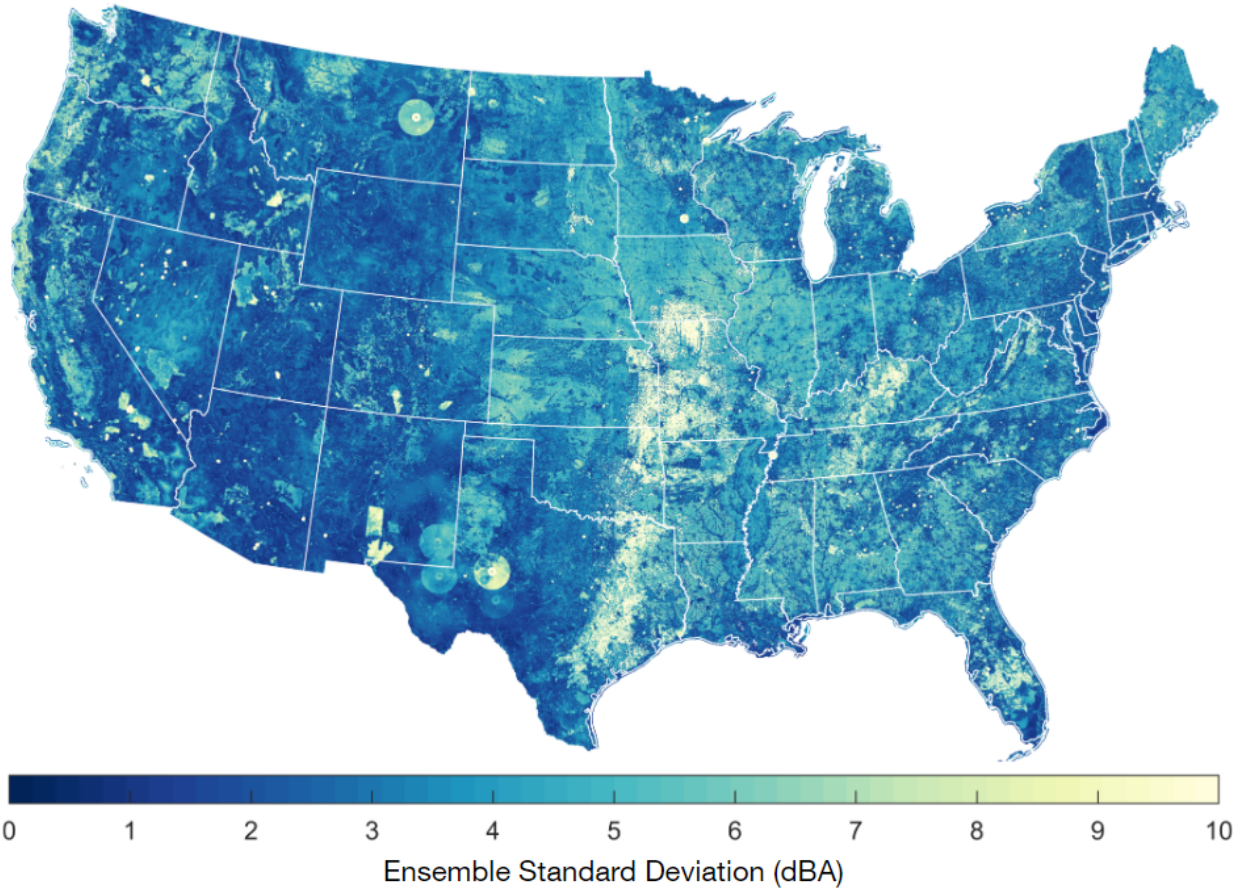

Figure 4: Standard deviation of ensemble model predictions for the A-weighted summer daytime L_{50} for CONUS.

4. LEAVE-ONE-OUT AND LEAVE-FOUR-OUT VALIDATION STUDY

Four sites were selected to be removed from the training set as part of a more extensive validation study. These sites were chosen to be unique from one another and illustrate strengths and weaknesses of the ensemble model approach. We present the results of this validation study for two of the four sites: (1) a site at Gilmore Meadow in Acadia National Park which is representative of the large number of national park training sites, and (2) a private home in a residential area in Asheville, North Carolina where significant insect noise occurs during the evening hours.

Figure 5 presents a comparison of the measured versus predicted daytime spectra levels for the two aforementioned validation sites, where the predicted levels in the left column of figures used the full training set and the predicted levels in the right column of figures used a training set that omitted the target prediction site (i.e., LOO). Similarly, Figure 6 presents the measured versus predicted *nighttime* spectra levels. The results presented in Figures 5 and 6 indicate that the model predictions are somewhat sensitive to the removal of a single training site as LOO predictions are not as close to the measured values as full model predictions. The results shown for these two validation sites are representative of

the results from all four validation sites.

Note, the NNs for all four sites are the only model that struggled to fit the measured data when the full training data set was used. This is likely because the NNs were very shallow (to prevent overfitting). The median ensemble LOO predictions were fairly stable and resilient to abrupt or drastic changes in one or two individual models' predictions. For example, at the Acadia National Park site, the KN algorithm tended to underpredict and the NN tended to overpredict, but the ensemble did a fairly good job of matching the measured values. However, there are some exceptions to the accuracy of the ensemble LOO predictions. In particular, none of the models were able to accurately predict the high-frequency daytime levels at the site in Acadia National Park or high-frequency nighttime levels at the site in Asheville. Our team speculates that the discrepancy in LOO model predictions and measured values in Asheville and Acadia is due to insect noise. It is possible that the geospatial data is insufficient to account for insect noise. It is also possible that the training data do not include sufficient examples of insect noise.

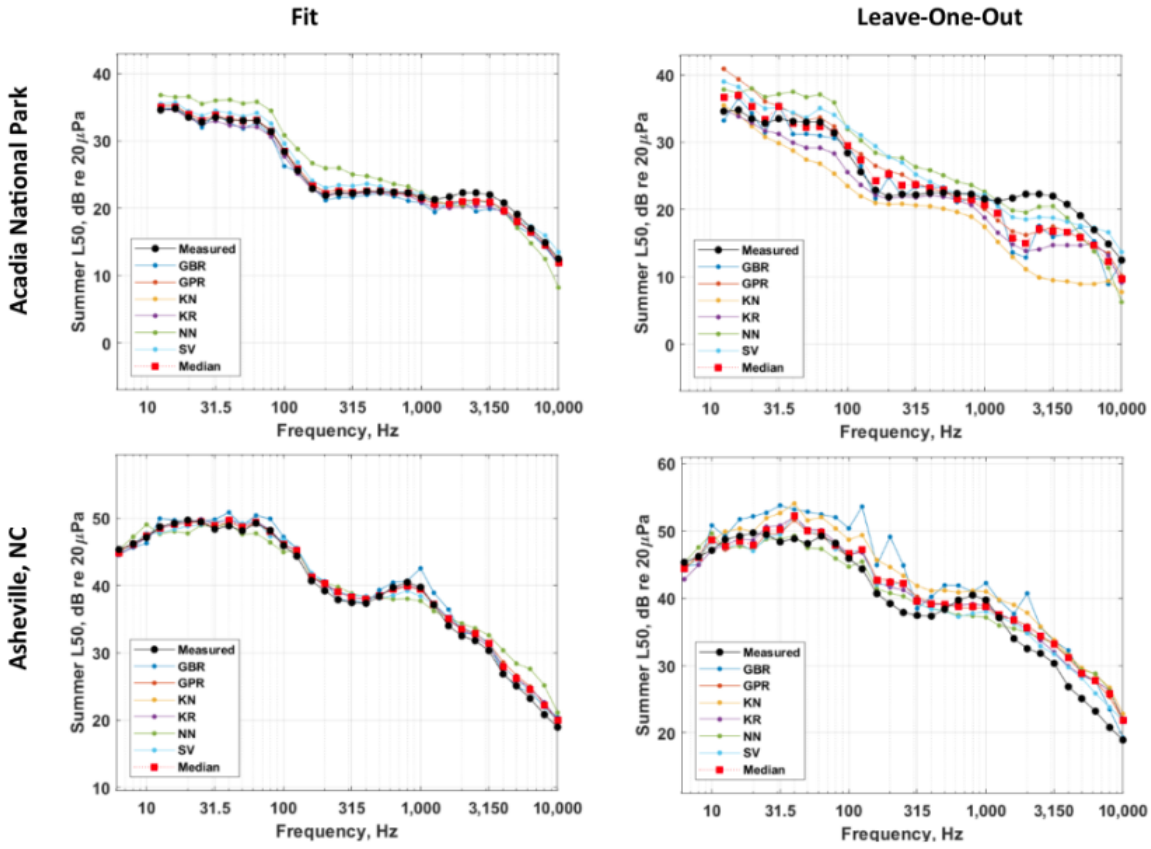

Figure 5: Daytime predictions and measured levels at the first two chosen validation sites as a function of frequency. Plots on the left show daytime model predictions when the complete training data set is used. Plots on the right are similar, but were created from predictions using a leave-one-out training data set in which all models were trained using the complete training data set, excluding the site of interest.

Figure 7 presents a comparison of the measured versus predicted daytime and nighttime spectra levels for the Gilmore Meadow and Asheville validations sites, where the predicted levels used a training set that omitted all four validation sites (i.e., leave-four-out). Note that the leave-four-out and LOO predictions are very similar. In both the leave-four-out and LOO plots, the GBR model has a tendency to predict values which oscillate as a function of frequency. This is more pronounced in the leave-four-out analysis and could be due to the propensity of decision tree-type models to overfit the training data. The ensemble, however, does not have this characteristic.

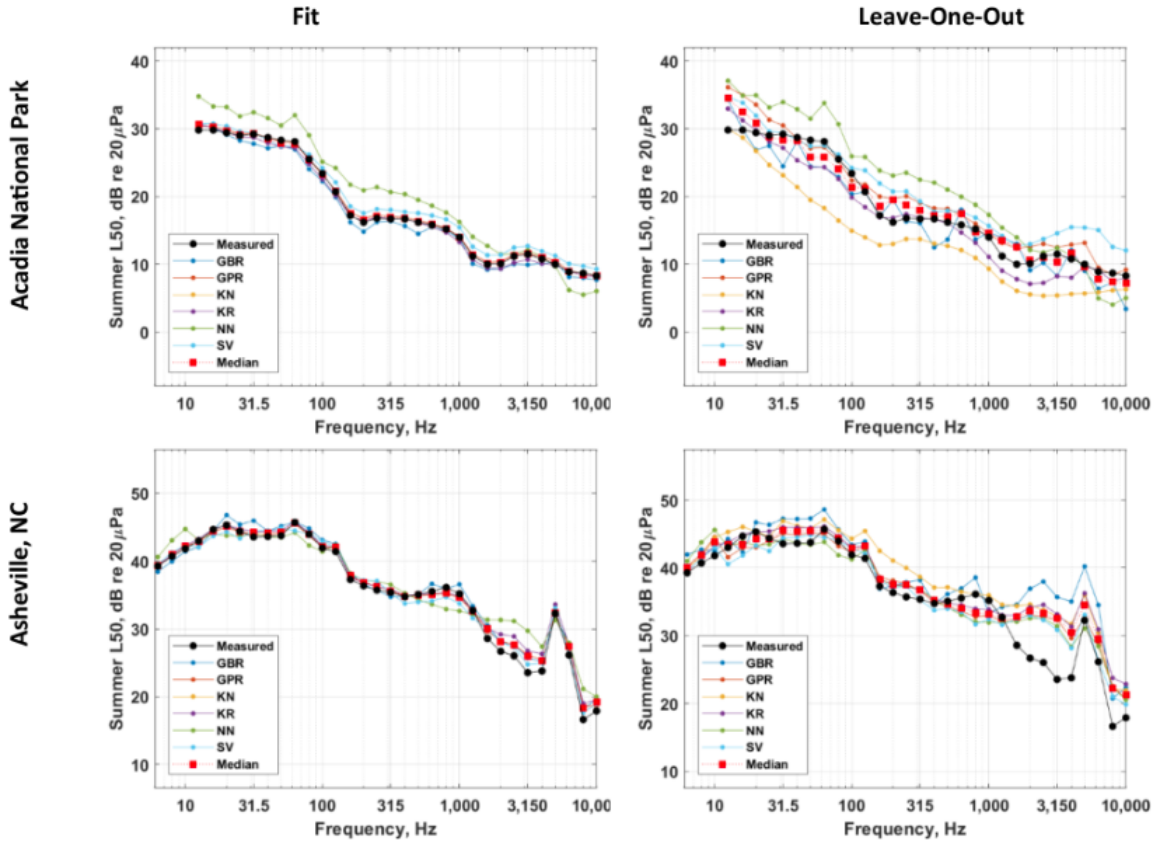

Figure 6: Nighttime predictions and measured levels at the first two chosen validation sites as a function of frequency. Plots on the left show nighttime model predictions when the complete training data set is used. Plots on the right are similar, but were created from predictions using a leave-one-out training data set in which all models were trained using the complete training data set, excluding the site of interest.

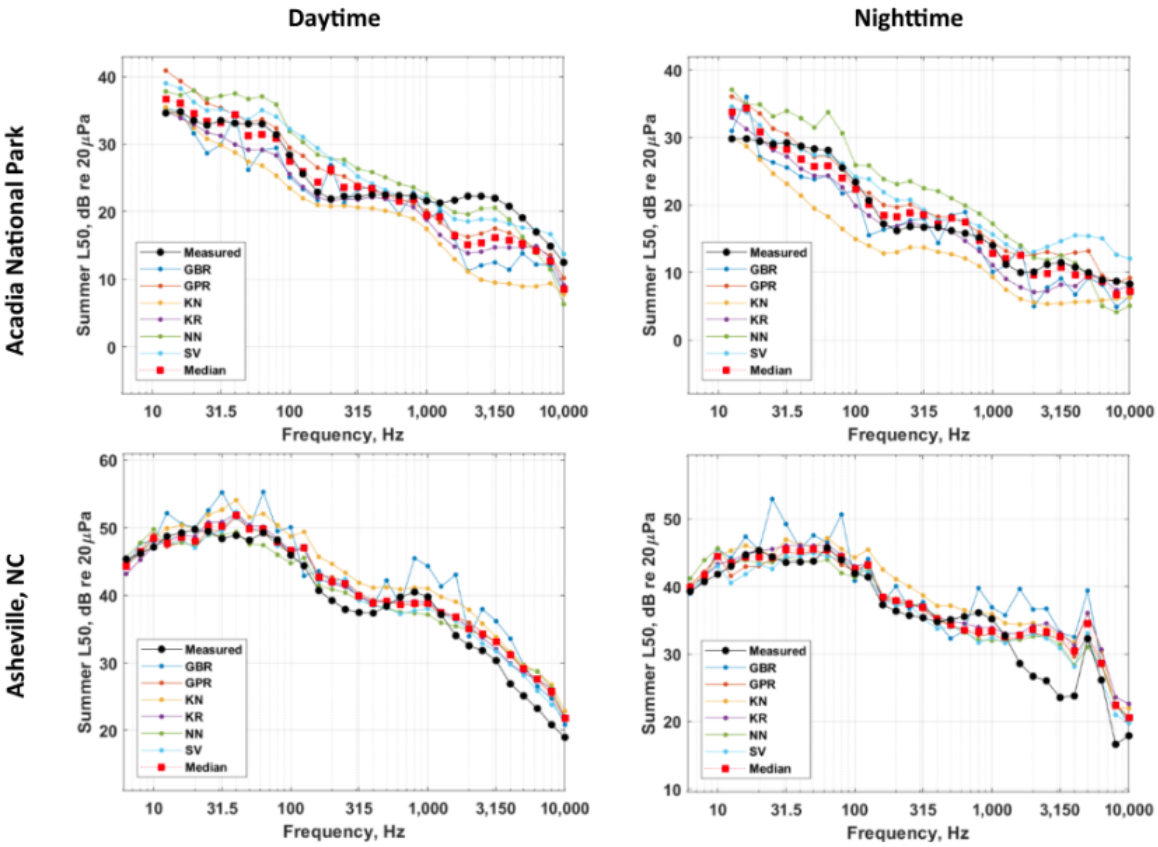

Figure 7: Plots on the left and right show daytime and nighttime model predictions respectively for the first two validation sites when the four validation sites were removed from the training data set. The plots show the predicted and measured levels at the four chosen validation sites as a function of frequency.

5. CONCLUSION AND FUTURE WORK

Our team 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 models, and Gaussian process regression models. The six model classes were 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. Our team's proof-of-concept study used the standard deviation of the ensemble predictions for the summer daytime L_{50} levels throughout the contiguous United States. Additionally, our team performed a leave-one-out and leave-four-out validation study that included spectral predictions.

Beyond estimating model accuracy, there are several advantages for using an ensemble model. Our team found that the ensemble predictions are more robust in extrapolation regions. The uncertainty measures gave an estimate of the model accuracy and can guide future data collection and feature reduction strategies.

This study motivates several future research directions. Our team's ensemble model quantified "structural uncertainty," i.e., uncertainty due to the functional form of the model. More extensive ensembles could be generated to account for other sources of uncertainty. In this regard, the uncertainty estimates reported here are conservative.

Soundscape modeling is an important acoustical question with potentially broad applications. The limitations of the available data, however, demand sophisticated uncertainty quantification tools in order to assess model transferability and improve predictive performance away from the training set. Uncertainty quantification methods will play an important role in guiding future data collection, feature reduction, and model selection.

6. ACKNOWLEDGMENT

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

REFERENCES

[1] National Park Service, "NPS Director's Order #47: Soundscape Preservation and Noise Management," (2000).

[2] N. A. of Engineering, "Protecting National Park Soundscapes," Washington, DC: The National Academic Press (2013).

[3] C. D. Francis *et al.*, "Acoustic Environments Matter: Synergistic Benefits fo Humans and Ecological Communities," Environ. Manage. **203,** 245–254 (2017).

[4] C. I. Merchan, L. Diaz-Balteiro, and M. Soliño, "Noise Pollution in National Parks: Soundscape and Economic Valuation," Landsc. Urban Plan. **123,** 1–9 (2014).

[5] D. Weinzimmer *et al.*, "Human Responses to Simulated Motorized Noise in National Parks," Leis. Sci. **36,** 251–267 (2014).

[6] C. D. Francis, C. P. Ortega, and A. Cruz, "Noise Pollution Changes Avian Communities and Species Interactions," Current Biology **19,** 1415–1419 (2009).

[7] H. Slabbekorn and W. Halfwerk, "Behavioural Ecology: Noise Annoys at Community Level," Current Biology **19,** R693–R695 (2009).

[8] B. C. Pijanowski *et al.*, "Soundscape Ecology: The Science of Sound in the Landscape," BioScience **61,** 203–216 (2011).

[9] E. P. Derryberry *et al.*, "Patterns of Song across Natural and Anthropogenic Soundscapes Suggest That White-Crowned Sparrows Minimize Acoustic Masking and Maximize Signal Content," PLoS ONE 11 (2016).

[10] G. Buscaino *et al.*, "Temporal Patterns in the Soundscape of the Shallow Waters of a Mediterranean Marine Protected Area," Scientific Reports 6 (2016).

[11] P. A. Hastings and A. Širović, "Soundscapes Offer Unique Opportunities for Studies of Fish Communities," In *Proceedings of the National Academy of Sciences of the United States of America*, **112,** 5866–5867 (2015).

[12] L. Ruppé *et al.*, "Environmental Constraints Drive the Partitioning of the Soundscape in Fishes," In *Proceedings of the National Academy of Sciences of the United States of America*, **112,** 6092–6097 (2015).

[13] F. Bertucci *et al.*, "Acoustic Indices Provide Information on the Status of Coral Reefs: An Example from Moorea Island in the South Pacific," Scientific Reports 6 (2016).

[14] S. M. Haver *et al.*, "The Not-So-Silent World: Measuring Arctic, Equitorial, and Antarctic soundscapes in the Atlantic Ocean," Deep Sea Research Part I: Oceanographic Research Papers (2017).

[15] S. Goutte, A. Dubois, and F. Legendre, "The Importance of Ambient Sound Level to Characterise Anuran Habitat," PLoS ONE 8 (2013).

[16] M. Tennesen, "Gauging Biodiversity by Listening to Forest Sounds," Scientific American (2008).

[17] M. E. Beutel *et al.*, "Noise Annoyance is Associated with Depression and Anxiety in the General Population - the Contribution of Aircraft Noise," PLoS ONE 11 (2016).

[18] T. Bodin *et al.*, "Road Traffic Noise and Hypertension: Results from a Cross-Sectional Public Health Survey in Southern Sweden," Environmental Health **8,** 38 (2009).

[19] L. Jarup *et al.*, "Hypertension and Exposure to Noise Near Airports: the HYENA Study," Environmental Health Perspectives **116,** 329 (2008).

[20] T. Münzel *et al.*, "Environmental Noise and the Cardiovascular System," J. Am. Coll. Cardiol. **71,** 688–697 (2018).

[21] S. A. Stansfeld *et al.*, "Aircraft and Road Traffic Noise and Children's Cognition and Health: a Cross-National Study," Lancet **365,** 1942–1949 (2005).

[22] S. P. Banbury *et al.*, "Auditory Distraction and Short-Term Memory: Phenomena and Practical Implications," Human Factors **43,** 12–29 (2001).

[23] D. J. Mennitt *et al.*, "Mapping Sound Pressure Levels on Continental Scales Using a Geospatial Sound Model," InterNoise13 (2013).

[24] D. Mennitt, K. Sherrill, and K. Fristrup, "A Geospatial Model of Ambient Sound Pressure Levels in the Contiguous United States," J. Acoust. Soc. Am. **135,** 2746–2764 (May 2014).

[25] D. Mennitt and K. Fristrup, "Influential Factors and Spatiotemporal Patterns of Environmental Sound Levels in the Contiguous United States," Noise Control Engr. J. **64,** 342–353 (2016).

[26] D. Xie, Y. Liu, and J. Chen, "Mapping Urban Environmental Noise: A Land Use Regression Method," Environ. Sci. Technol. **45,** 7358–7364 (2011).

[27] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective* (Cambridge University Press, 2011).

[28] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering **22,** 1345–1359 (2010).

[29] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications* (SIAM Computational Science & Engineering Series, Philadelphia, PA, USA, 2014).

[30] M. C. Kennedy and A. O'Hagan, "Bayesian Calibration of Computer Models," J. R. Statist. Soc. **63,** 425–464 (2001).

[31] National Park Service, "Data Store," [Online]. Available: https://irma.nps.gov/DataStore/.

[32] 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).

[33] 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. Environmental Protection Agency (Washington, D.C.) (1974).

[34] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," arXiv preprint arXiv:1609.08144 (2016).

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," In NIPS, 1097–1105 (2012).

[36] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," JMLR **12,** 2825–2830 (2011).