K-Means clustering of inputs to a geospatial model for optimizing acoustic data collection

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

Citation: Proc. Mtgs. Acoust. **35**, 055008 (2018); doi: 10.1121/2.0001299

View online: https://doi.org/10.1121/2.0001299

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 November 2018

CANADIAN ACOUSTICAL ASSOCIATION
ASSOCIATION CANADIENNE D'ACOUSTIQUE

Signal Processing in Acoustics: Paper 1pSP11

K-Means clustering of inputs to a geospatial model for optimizing acoustic data collection

Brooks A. Butler, Katrina Pedersen, Casie Maekawa, Kent L. Gee and Mark K. Transtrum
Department of Physics and Astronomy, Brigham Young University, Provo, UT, 84602; brooks.butler93@gmail.com; katrina.pedersen@gmail.com; casie.maekawa@juabsd.org; kentgee@byu.edu; mktranstrum@byu.edu

Michael M. James and Alexandria R. Salton
Blue Ridge Research and Consulting, LLC, Ashville; michael.james@blueridgeresearch.com, alex.salton@blueridgeresearch.com

Outdoor ambient acoustical environments may be predicted through machine learning using geospatial features as inputs. However, collecting sufficient training data is an expensive process, particularly when attempting to improve the accuracy of models based on supervised learning methods over large, geospatially diverse regions. Unsupervised machine learning methods, such as K-Means clustering analysis, enable a statistical comparison between the geospatial diversity represented in the current training dataset versus the predictor locations. In this case, 117 geospatial features that represent the contiguous United States have been clustered using K-Means clustering. Results show that most geospatial clusters group themselves according to a relatively small number of prominent geospatial features. It is shown that the available acoustic training dataset has a relatively low geospatial diversity because most training data sites reside in a few clusters. This analysis informs the selection of new site locations for data collection that improve the statistical similarity of the training and input datasets.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0001299

1. INTRODUCTION

Ambient noise, and its effects, have been the focus of study across many disciplines including psychology[1-3], medicine[4-6], urban planning[7], and ecology[8]. One method for predicting outdoor ambient noise—developed by Mennitt et al[9,10]—utilizes machine learning on sampled ambient acoustic data from across the contiguous United States (CONUS). This method can be described in two parts: (1) The collection of outdoor acoustic data and geospatial features (e.g., roads, nighttime lights, etc.) across CONUS, and (2) training a machine learning model to learn the relationship between outdoor acoustic environments and geospatial features, enabling predictions across a larger geographic area. Details on creating a machine learning model for predicting ambient acoustic soundscapes have been published by Pedersen et al.[11]

The focus of this paper is the optimal experimental design (OED) of part (1), informed by part (2); namely, the optimal collection of new acoustic data at targeted locations such that the uncertainty of predictions made is minimized. OED can also be classified as a type of active learning in the field of machine learning and data acquisition[12,13]. The goal of OED is to use the output of a given learned model to inform where additional data is required to optimize model performance[14-16].

The challenge posed by our acoustic dataset is a low ratio of available training data when compared with the size and geospatial diversity of acoustically unmeasured areas across CONUS. The effect of this sparse training set is a high amount of uncertainty in predictions made at locations where little to no training data at geospatially similar locations is available. This uncertainty in our model predictions can be a guide to where more training data is needed. However, for a more complete understanding of where our training data struggles to represent areas geospatially, we performed a comparative analysis to measure the overall geospatial similarity between our training data set and the entire CONUS area using a joint k-means clustering analysis.[17,18] The clustering analysis identified where our training data lacked enough data points in certain geospatial clusters and that these under-sampled clusters correlate with areas of high uncertainty in our model predictions. Using this clustering analysis as a guide, we simulate adding new acoustic data and attempt acoustic measurements in areas labeled as under-sampled geospatial clusters in Utah. We find the addition of this data helps to reduce the uncertainty of our model significantly in other areas across CONUS that are geospatially similar. As such, this targeted data acquisition and implementation of OED should significantly improve the ability of our model to make predictions with far fewer additional data points and for a fraction of the cost of random data sampling.

2. GEOSPATIAL DATA

We compile geospatial data from the National Parks Service (NPS) Natural Sounds and Night Skies and Inventory and Monitoring Divisions. The data consists of 117 geospatial feature layers, with each geospatial layer representing a different physical metric across the CONUS defined by latitude and longitude coordinates. These values are measured in a variety of units such as proportions of land cover, ratios, area densities, frequency of flight observations, distances, measured light intensity from satellite imagery, etc. More generally, our geospatial layers can be sorted into six categories: topography, climate, land cover, hydrology, anthropogenic, and position. An example for each category of geospatial layer can be seen in Figure 1. For more details on the collection of feature data see Pedersen et al.[11]

To account for numerical range differences which are inherent in a variety of physical units, we apply a standard scaling method to ensure equal weighting for each feature layer. This process performs an affine transformation such that each geospatial layer has a mean of zero and a standard deviation of one. For simplicity, this is the only scaling method we use for our data. However, other scaling methods may prove useful in future research.

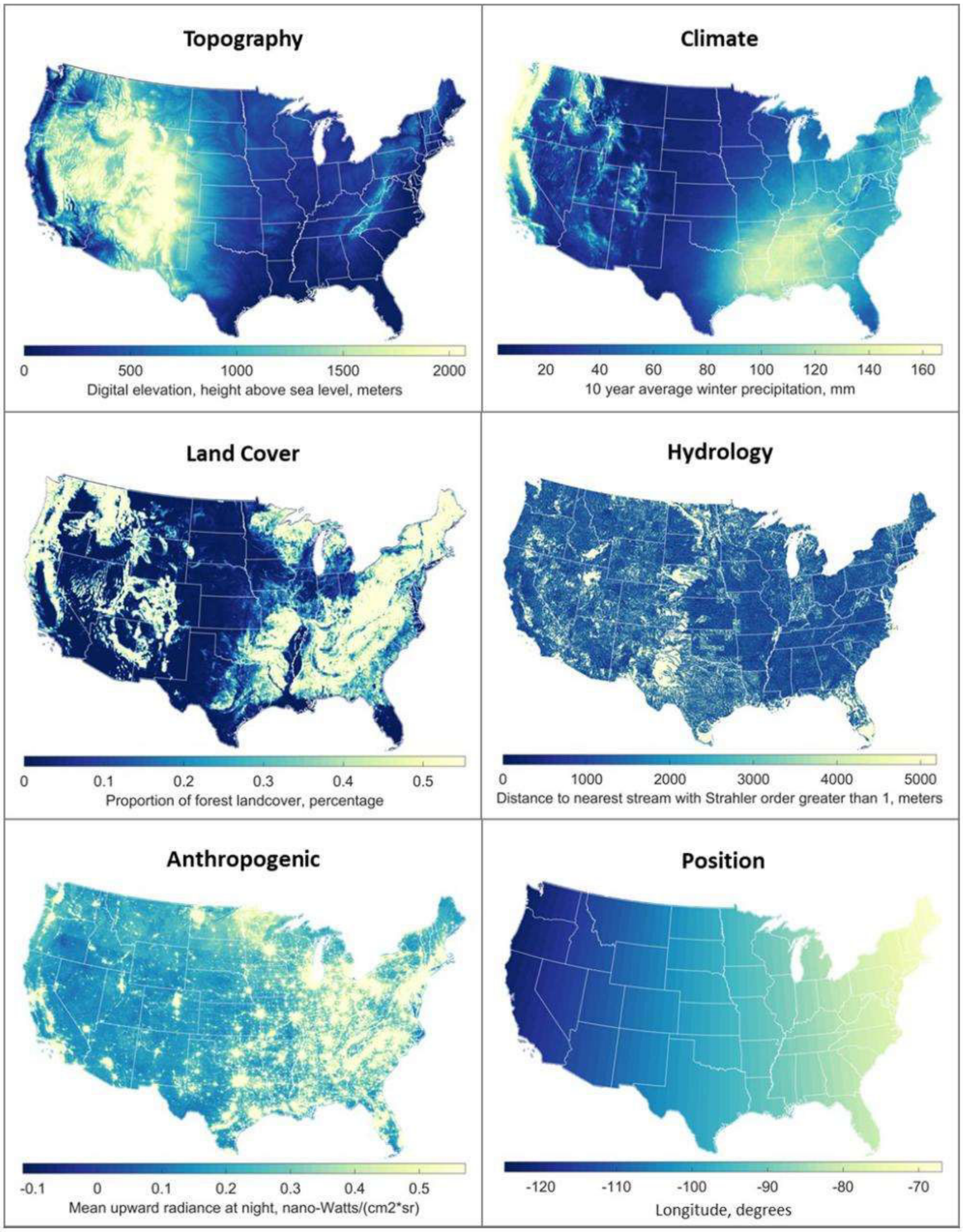

Figure 1. Examples of the six categories of geospatial data layers found in our input data set.

3. CLUSTERING ANALYSIS

Clustering is a form of unsupervised machine learning that attempts to sort data into groups, or clusters, by measuring the geometric similarity of data groups[17,18]. Clustering analyses are also useful in other acoustic data applications such speech analysis[19] and room acoustics[20]. However, in the case of geospatial acoustics, we performed clustering on the geospatial components of our data to locate sites where new ambient acoustic data should be taken.

We use a k-means clustering analysis on our geospatial layer data in hopes of objectively identifying which geospatial sites are unique from one another according to the values of their geospatial features. The goal of k-means clustering is to partition data into K clusters based on geometric proximity of data objects to each other, where K is a predetermined number of how many clusters we expect to see in the data. Each cluster is assigned a centroid, defined as the center point of each cluster, which is then moved iteratively until all centroids have reached a certain threshold of stability.[21,22]

While it is easier to visually discern possible cluster groupings in two or even three-dimensional data, higher dimensional data—such as our 117 geospatial feature layers—can prove difficult to intuitively visualize. There are, however, computational methods for determining the optimal number of clusters K. One such method is called an "Elbow analysis"[23] which is done by calculating the average distortion (d_K) for each number of clusters K,

$$d_K = \frac{1}{N}\sum_{i=1}^{N}(X_i - c_K)^2$$

where N is the number of data points and $(X_i - c_K)^2$ is the squared distance between the data X_i and its nearest cluster center c_K. Plotting the average distortion value for each number of clusters K creates a distortion curve, which monotonically decreases until $K > N$. We can use this curve to determine at which point we begin to see diminishing returns in the descriptive power of adding more clusters to the analysis.

We determine that a reasonable number for geospatial clusters occurs around $K = 12$, being around the inflection point of the distortion curve as shown in Figure 2. While we do not believe that only 12 discrete unique geospatial clusters exist across the Contiguous United States, we do believe 12 clusters sufficiently generalize the different types of geographic regions that should also be observed in our training data set. However, future research may explore finer resolution clustering by performing a hierarchical clustering analysis.[24]

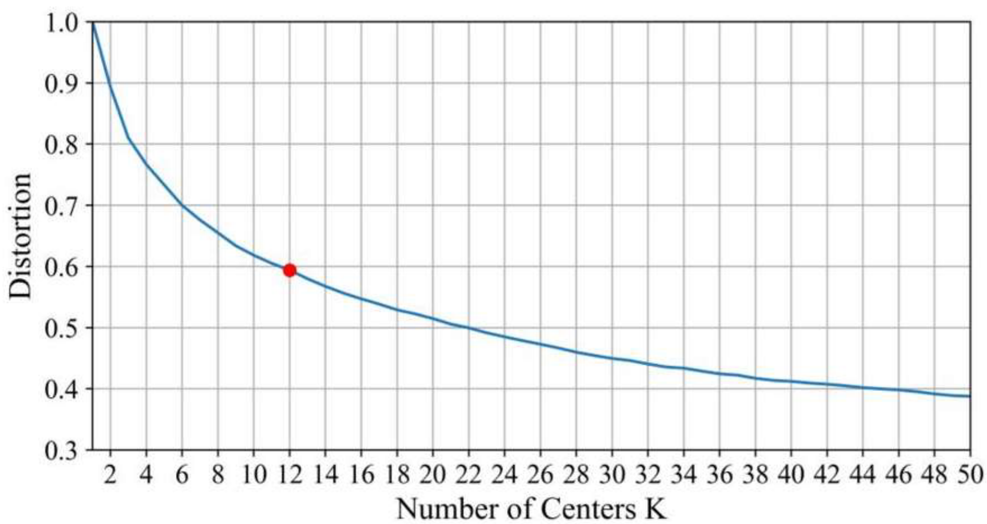

Figure 2. Elbow analysis performed on geospatial data set, showing the average distortion calculated for each number of clusters K. The red dot at K= 12 represents approximately the point of diminishing returns where adding more clusters does not significantly further describe the data set.

Geospatial clusters are most easily observed when projected onto the two-dimensional space of latitude and longitude, effectively creating a "cluster map" of the CONUS area. This cluster map using 12 clusters for CONUS is shown in Figure 3, with cluster colors being arbitrarily assigned to cluster numbers for maximum contrast. A more zoomed in map of the Utah area is shown in Figure 4. By performing a cross correlation analysis of each geospatial feature compared with each predicted cluster, we can gain some insight into which features contribute the most to the assignment of each cluster. The top five correlated features for each cluster are show in Table 1.

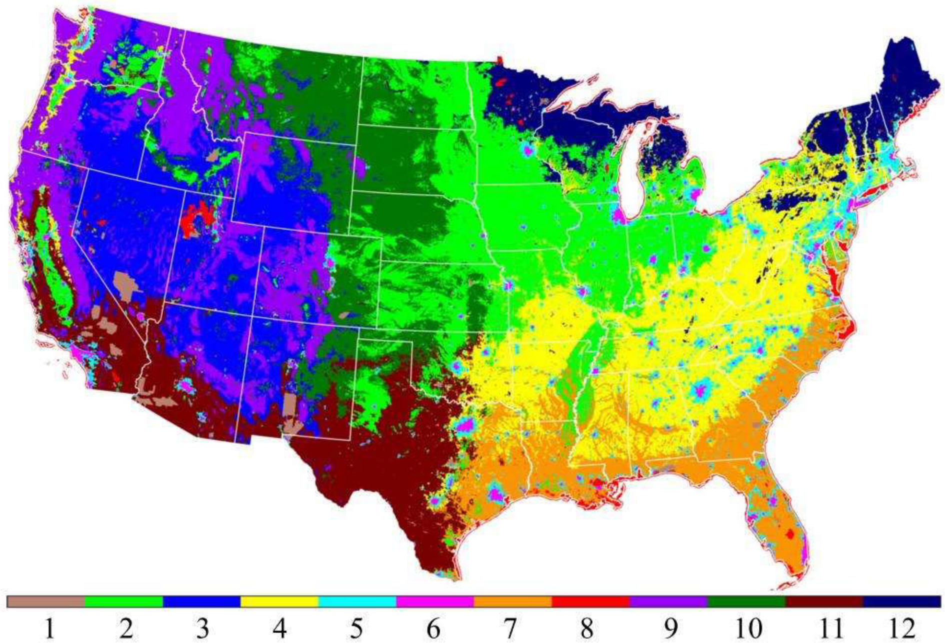

Figure 3. Cluster map of the contiguous United States for K=12 clusters. Each color represents a different cluster label and is assigned arbitrarily.

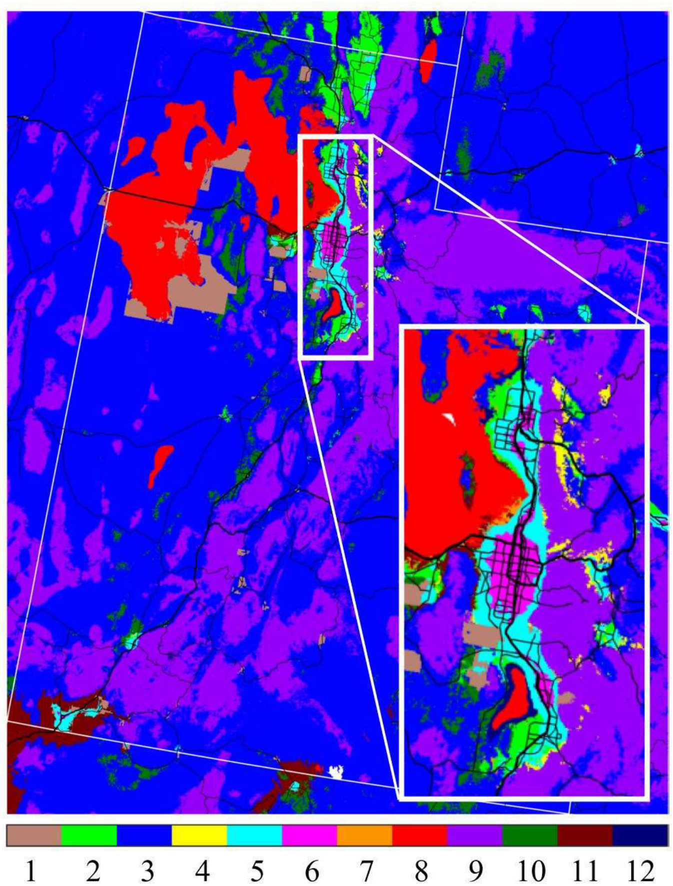

Figure 4. An enlarged section of the cluster map showing Utah and the Wasatch Front. Black lines indicate major roads and highways while white lines indicate state boundaries. Each color represents a different cluster label and is assigned arbitrarily.

These correlations motivate our human interpretation of geospatial cluster assignments. For example, we see that cluster 2 correlates most with areas relating to cropland, cultivated land, and land used for grazing; all of which are related to land used for farming in more rural areas. Additionally, clusters 5 and 6 correlate most with features related to concentrated population and land modified for human use. These correlation results lead us to believe that a 12-cluster model is a good generalization for determining significantly different geospatial areas.

Table 1. The top 5 correlated features for geospatial clusters according to the absolute value of the calculated correlation coefficients, where a value of 1 represents perfect correlation and a value of -1 perfect negative correlation.

Cluster #	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
1	Institutional (0.89)	Built (0.32)	Extractive (-0.16)	Shrubland (0.15)	Summer Max Temp (0.10)
2	Cropland (0.86)	Cultivated (0.80)	Grazing (-0.32)	Distance to Coast (0.31)	Winter Max Temp (-0.27)
3	Shrubland (0.64)	TdewAvgSummer (-0.52)	Elevation (0.51)	Summer Precipitation (-0.44)	Grazing (0.42)
4	Deciduous (0.70)	Exurban High (0.61)	Timber (0.55)	Forest (0.52)	Annual Precipitation (0.39)
5	Suburban (0.63)	Exurban Low (0.63)	Developed (0.52)	Built (0.50)	RddAll (0.43)
6	Urban Low (0.76)	VIIRSMean (0.72)	VIIRSMinimum (0.71)	Major Roads (0.69)	Developed (0.67)
7	Wet (0.62)	Wetlands (0.58)	TdewAvgAnnual (0.49)	Annual Min Temp (0.44)	Summer Precipitation (0.42)
8	Natural Water (0.70)	Water (0.64)	Extractive (-0.28)	Barren (0.15)	TdewAvgWinter (0.14)
9	Evergreen (0.79)	Slope (0.52)	Forest (0.51)	Elevation (0.51)	Summer Min Temp (-0.48)
10	Herbaceous (0.81)	Grazing (0.42)	Distance to Coast (0.34)	Winter Precipitation (-0.31)	Winter Min Temp (-0.29)
11	Shrubland (0.50)	Annual Max Temp (0.47)	Annual Min Temp (0.36)	Grazing (0.34)	MilitarySum (0.29)
12	Mixed Forest (0.53)	Winter Max Temp (-0.32)	Annual Max Temp (0.31)	Forest (0.30)	Deciduous (0.29)

4. TARGETED ACOUSTIC DATA COLLECTION

By comparing the occurrence of cluster assignment across the CONUS area with the occurrences in our training data we see that several geospatial clusters are significantly under-represented in our training set of acoustic data, as shown in Figure 5. The goal of our data acquisition approach is to maximize the benefit of adding the fewest possible data points to our training data set. In our case, we quantify the benefit added by data through the reduction of uncertainty in model predictions, where uncertainty is measured by the standard deviation of predictions made by an ensemble of machine learning models as described by Pedersen et al.[11] Instead of sampling data randomly to reduce the overall uncertainty of acoustic predictions, we hypothesize that a more targeted approach to adding data will yield targeted reductions in model uncertainty.

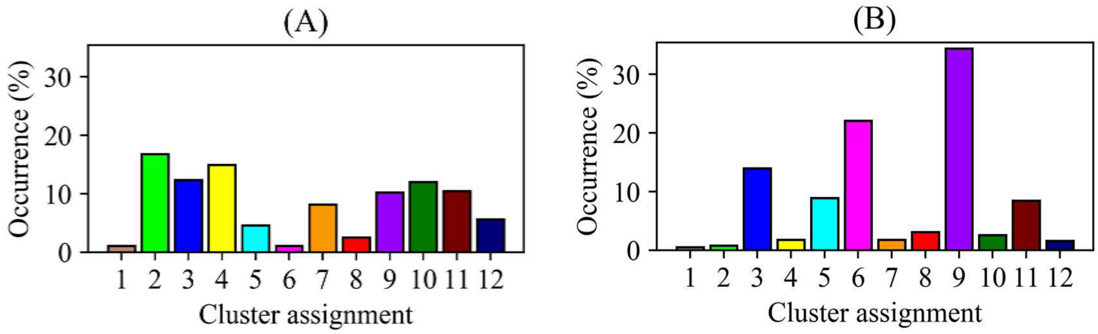

Figure 5. The occurrence of cluster assignment across the contiguous United States (A), and in our training data set (B).

We took acoustic data in two rural locations along the Wasatch Front metropolitan region in the north-central part of Utah and ran a series of tests to measure uncertainty reduction in predictions made across each geospatial cluster. Because of the discrete nature of data clustering, it is possible to take data in regions that lay near the border of several clusters when considering the values of their geospatial features, making it sometimes difficult to precisely identify geospatial cluster regions while taking field measurements. Consequently, while our intention was to take data in under-sampled cluster regions, the actual data taken in this study ended being narrowly located in Cluster 3. However, when measuring the distance of each site in the feature space to each cluster centroid, as shown in Figure 6, we see that the next closest cluster centers are Clusters 9, 10 and 11. The effect of these data, and proximity to multiple clusters, are discussed further in our results.

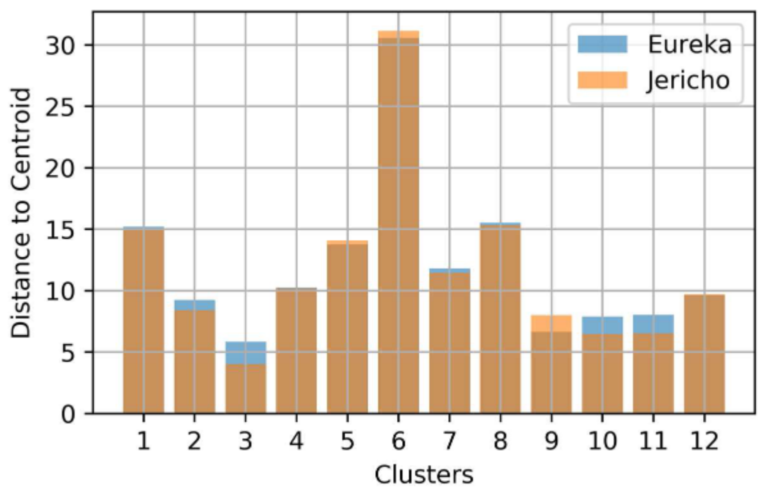

Figure 6. The calculated distance in feature space of the two added sites to our dataset, named the Eureka (Blue) and Jericho (Orange) sites.

The measurements were made using a Larson Davis 824 sound level meter with Type-1 half-inch free-field microphone. The meter was kept inside weather resistant casing to protect against heat and moisture. The microphone was placed on a stake approximately 5 ft above the ground and covered with a wind screen to prevent wind noise contamination. An example of a typical data acquisition setup is shown in Figure 7. Additionally, we chose microphone placement to be sufficiently far from potential sound sources (e.g., immediately adjacent to a road) to hopefully prevent measurements from being influenced by individual sound events. We collected measurements for several days at each site to provide enough averaging for both daytime and nighttime ambient sound levels

Figure 7. Data acquisition setup for collecting measurements of outdoor ambient sound.

5. RESULTS

To measure the effect of these new data on our model predictions, we incorporate the summer daytime L50 measured levels from each location into the training dataset. A script is generated to load in the training dataset of 502 measurement sites, not including the two new locations, and randomly leave out ten sites. After the ten random sites are removed, the remaining data is scaled using a standard scaler and the six ensemble models are trained on the data. Predictions are made for a low-resolution map of CONUS sound level predictions, and the standard deviation of ensemble model predictions is calculated. Then, two sites from the ten removed sites are added back into the training data before scaling was applied. The data is then scaled, and predictions and uncertainties are calculated again. The two sites that were added from the original removed ten sites are then removed again, and this entire process is repeated by sampling without replacement from the remaining eight sites until no sites remain. Lastly, data from our two new targeted locations are added to the set of 492 sites and predictions and uncertainties calculated. This process is performed 100 times, randomly ordering the 502 sites each time.

For each run, the mean uncertainty of predictions from each cluster is calculated. Histograms are generated to show changes in the mean uncertainty by adding two random sites from the training dataset or by adding the new targeted data sites. Our results show that adding only two data sites, when targeted for specific geospatial clusters, measurably reduced the uncertainty of predictions made in their related clusters while leaving the uncertainty of predictions made in other clusters unaffected. We show this in Figure 8, where the uncertainty in clusters 3, 10, and 11 is reduced while cluster 9 is unaffected, which is characteristic of all other clusters. It should be noted that even though the cluster label for both sites is Cluster 3, the proximity of each site to the cluster centers of Clusters 10 and 11 is enough to reduce the uncertainty model predictions for those regions. Additionally, even though the data are also close in the feature space to Cluster 9, we see no effect due to the already large proportion of data in Cluster 9 already present in our dataset.

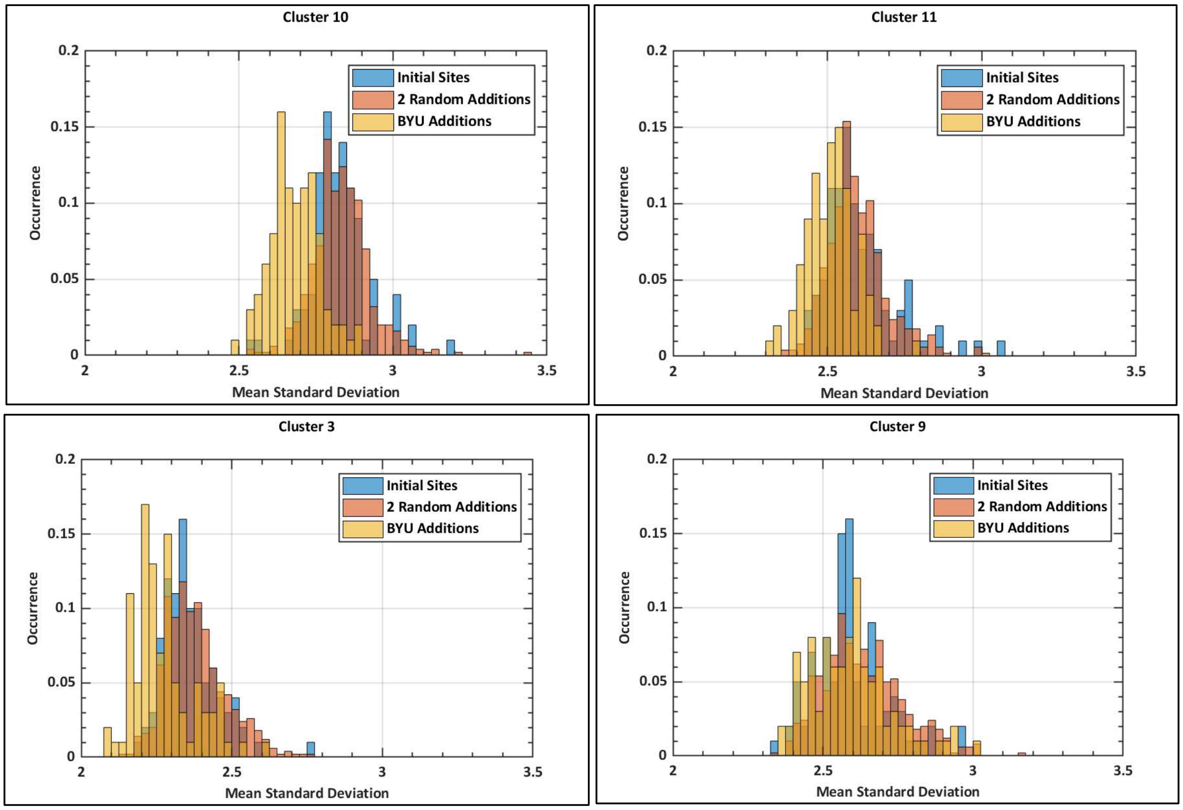

Figure 8. Histograms showing the uncertainty of predictions made using three different data sets over multiple trials. The first set of predictions (blue) consist of our initial data with two random sites left out of the model. The second set of prediction (red) add the two random sites back into the model, while the third set of predictions (yellow) add instead the two sites in Utah labeled as Cluster 3. We see the uncertainty of model predictions are reduced in locations labeled as clusters 3, 10, and 11, while predictions in Cluster 9 is unaffected (which is characteristic of all other clusters).

6. CONCLUSION AND FUTURE WORK

Although this work is preliminary, we have shown a framework for using a targeted data collection approach that can reduce uncertainty of machine-learned acoustic predictions in specific geospatial clusters across all of CONUS. Additionally, this combination of using both supervised and unsupervised methods to drive the optimal experimental design of our model adds another layer of validation which may prove valuable in validating predictions made by future acoustic models. For future work, we can use this same method to guide data collection efforts in under-represented geospatial clusters to reduce model uncertainty while minimizing the number of data points needed to our total training dataset. This targeted data collection will also help to reduce the cost of taking data because data taken in one cluster has an apparent effect on predictions made in that cluster across CONUS, i.e. data taken in Cluster 2 in Utah likely affects predictions made in Cluster 2 across CONUS. Thus, the need to travel to specific locations of model uncertainty in order to collect data may be reduced. Rather, we need only to travel to the nearest location of a cluster where there is a high prediction of uncertainty. The hypothesized benefits of this data collection methodology is the subject of ongoing work.

ACKNOWLEDGMENTS

This research was supported by Blue Ridge Research and Consulting, LLC under a United States Army Small Business Innovative Research program grant, with James H. Stephenson as project monitor.

REFERENCES

[1] J. M. Fields, “Effect of personal and situational variables on noise annoyance in residential areas,” J. Acoust. Soc. Am., **93** 2753-2763 (1993); doi:10.1121/1.405851.
[2] G. W. Evans, M. Bullinger, and S. Hygge, “Chronic noise exposure and physiological response: A prospective study of children living under environmental stress,” Psych. Sci. **9**, 75-77 (1998). doi:10.1111/1467-9280.00014.
[3] R. Crombie, C. Clark, and S. A. Stansfeld, “Environmental noise exposure, early biological risk and mental health in nine to ten year old children: A cross-sectional field study,” Environ. Health **10**, 39 (2011); doi:10.1186/1476-069X-10-39.
[4] W. Q. Gan, H. W. Davies, M. Koehoorn, and M. Brauer, “Association of long-term exposure to community noise and traffic-related air pollution with coronary heart disease mortality,” Am. J. Epidem. **175**, 898-906 (2012); doi:10.1093/aje/kwr424.
[5] A. V. Moudon, “Real noise from the urban environment: How ambient community noise affects health and what can be done about it,” Am. J. Prevent. Med. **37**, 167-171 (2009); doi:10.1016/j.amepre.2009.03.019.
[6] M. Kim, S. I. Chang, J. C. Seong, J. B. Holt, T. H. Park, J. H. Ko, and J. B. Croft, “Road traffic noise annoyance, sleep disturbance, and public health implications,” Am. J. Prevent. Med. **43**, 353-360 (2012); doi: 10.1016/j.amepre.2012.06.014.
[7] W. Joo, S. H. Gage, and E. P. Kasten, “Analysis and interpretation of variability in soundscapes along an urban-rural gradient,” Landscape Urb. Plan. **103**, 259-276 (2011); doi:10.1016/j.landurbplan.2011.08.001.
[8] T. C. Mullet, S. H. Gage, J. M. Morton, and F. Huettmann, “Temporal and spatial variation of a winter soundscape in south-central Alaska,” Landscape Ecol. **31**, 1117-1137 (2016). doi:10.1007/s10980-015-0323-0.
[9] 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 (2014); doi:10.1121/1.4870481.
[10] D. J. Mennitt and K. M. Fristrup, “Influential factors and spatiotemporal patterns of environmental sound levels in the contiguous United States,” Noise Control Eng. J. **64**, 342-353 (2016); doi:10.3397/1/376384.
[11] K. Pedersen, K. L. Gee, M. K. Transtrum, B. A. Butler, M. M. James, and A. R. Salton, “Machine learning-based ensemble model predictions of outdoor ambient sound levels,” Proc. Mtgs. Acoust. **35**, 022002 (2018); doi: 10.1121/2.0001056.
[12] Z. Zheng and B. Padmanabhan, “Selectively acquiring customer information: A new data acquisition problem and an active learning-based solution,” Management Sci. **52**, 697-712 (2006); doi:10.1287/mnsc.1050.0488.
[13] B. Osting, C. Brune, and S. J. Osher, “Optimal data collection for informative rankings expose well-connected graphs,” J. Machine Learning Res. **15**, 2981−3012 (2014).
[14] L. Pronzato, “Optimal experimental design and some related control problems,” Automatica **44**, 303-325 (2008); doi:10.1016/j.automatica.2007.05.016.
[15] J. P. Morgan and X. Deng, “Experimental design,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**, 164-172 (2012); doi:10.1002/widm.1046.
[16] C. C. Drovandi, C. C. Holmes, J. M. McGree, K. Mengersen, S. Richardson, and E. G. Ryan, “Principles of experimental design for big data analysis,” Stat. Sci. **32**, 385-404 (2017); doi:10.1214/16-STS604.
[17] A. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” IEEE Trans. Pattern Analysis Machine Intell. **22**, 4-37 (2000); doi:10.1109/34.824819
[18] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Trans. Neural Networks, **16**, 645-678 (2005); doi:10.1109/TNN.2005.845141.
[19] S. S. Chen, and P. S. Gopalakrishnan. “Clustering via the Bayesian information criterion with applications in speech recognition." In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 2, pp. 645-648. IEEE, 1998
[20] S. Bharitkar, and C. Kyriakakis. “A cluster centroid method for room response equalization at multiple locations." In Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575), pp. 55-58. IEEE, 2001.
[21] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” IEEE Trans. Pattern Analysis Machine Intell. **24**, 881-892 (2002); doi:10.1109/TPAMI.2002.1017616.
[22] B. Peralta, P. Espinace, and A. Soto, “Enhancing K-Means using class labels,” Intell. Data Analysis **17**, (2013); doi: 10.5555/2595599.2595605.
[23] P. Bholowalia, A. Kumar, “EBK-Means: A clustering technique based on elbow method and K-means in WSN,” Int. J. Comput. Applic. **105**, 17-24 (2014).
[24] A. K. Jain, M. N. Murty, and P.J. Flynn, “Data clustering: A review,” ACM Comput. Surveys (CSUR) **31**, 264-323 (1999).