K-Means Clustering Analysis of Geospatial Features: Optimizing the Acquisition of Training Data

Brooks Butler

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

Bachelor of Science

Kent Gee and Mark Transtrum, Advisors

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

K-Means Clustering Analysis of Geospatial Features: Optimizing the Acquisition of Training Data

Brooks Butler
Department of Physics and Astronomy, BYU
Bachelor of Science

Data acquisition for machine learning training data sets can be an expensive and time consuming process. For the BYU Acoustics Research Group and the Blue Ridge Research Company, the acquisition of outdoor ambient acoustic data for a single location can take up to a week to measure accurately. Since the purpose of our research is to create an accurate model for predicting ambient acoustic noise levels across the continental United States area, we would like to maximize the benefit of new data added to our existing training data set while minimizing the amount of new data needed. A joint K-means Clustering analysis was used to measure the statistical similarity between the geospatial feature values of our training and input data set for the North Carolina region. Using this analysis we were able to identify which kinds of geographic locations are under-sampled in the NC area. We can conclude from our analysis that our current training set does not fully represent the geographical area over which we are trying to make predictions. Additionally, targeted sampling of locations based on cluster assignment will improve the statistical similarity of our training and input data sets.

Keywords: Machine learning, K-means clustering, Geographical information systems (GIS), Acoustics, Community Noise

ACKNOWLEDGMENTS

EDITS MADE:

Most of the changes that I made to this thesis since my submission of the final draft were in my results section. I received valuable feedback on how to make the significance of my results more clear to the reader by assigning interpretive labels to the cluster maps where I could definitively say what kinds of factors were influencing the assignment of that cluster label. Although in other cases it seemed like there were cluster assignments that were simply just fillers for those not assigned to the other prominent features so I left those unlabeled since those required further analysis.

I also made small polishing edits throughout where I have made some simple grammar mistakes and clarified one point in my overview discussing what predictive power meant.

Contents

Table of Contents **iv**

List of Figures **v**

1 Introduction **1**
1.1 Motivation 1
1.2 Previous work 3
1.3 Overview 3

2 Geospatial and Ambient Acoustic Data **5**
2.1 The nature of geospatial feature data 5
2.2 Ambient acoustic noise level measurements 7
2.3 Applications through machine learning 9

3 Clustering Methods **11**
3.1 Unsupervised Machine Learning 11
3.2 K-Means Clustering Analysis 11
3.3 Scaling of Input Feature Data 14
3.4 Calculation of Cross-entropy 15
3.5 Feature Ranking 15

4 Results and Analysis **19**
4.1 Cluster Maps 19
4.2 Feature Importance 20
4.3 Input Map vs Training Data 21

5 Colnclusions **28**
5.1 Optimization of Training Data Acquistition 28
5.2 Future Work 29

Bibliography **30**

Index **31**

List of Figures

2.1 A map of several geospatial features. 6

2.2 Example of how L10, L50, and L90 acoustic metrics are calculated. 8

2.3 Example of ambient acoustic L50 spectral data. 9

2.4 Flow chart of supervised machine learning model. 10

3.1 A global view of machine learning practices. 12

3.2 An example of two dimensional cluster group 12

3.3 A flow chart of the K-means clustering algorithm 13

3.4 A plot of average centriod density over number of clusters 14

3.5 Feature maps of higher importance in determining cluster assignment. 17

4.1 A cluster map of North Carolina with magnified portion of Ashville, NC. 20

4.2 A map of the NC area as shown by Google 21

4.3 A visualization of NC clusters 1-3 separated from each other. 22

4.4 A visualization of NC clusters 4-6 separated from each other. 23

4.5 A visualization of NC clusters 7-9 separated from each other. 24

4.6 A visualization of NC clusters 10-12 separated from each other. 25

4.7 Urban gradient at Ashiville, NC. 26

4.8 Map of the Ashiville city limits and surrounding area. 26

4.9 A comparison of cluster representation between the NC input map and our Training set. 27

Chapter 1

Introduction

Data acquisition and processing are both essential steps in every field of the physical sciences. Specifically in the field of machine learning, the primary goal of data acquisition is to make predictions on large sets of input data using relatively small sets of training data. There is, however, a cost related to the acquisition of training data, and in some cases the amount of time and effort needed to obtain a single point of training data can be substantial. This leads us to an important question: How can we maximize the benefits gained from adding new training data to an existing training data set while simultaneously minimizing the number of additional data points needed?

1.1 Motivation

In recent work done at BYU, in conjunction with the Blue Ridge Research Company (BBRC) based in Ashville, NC, efforts have been made to develop a machine learning model that uses geospatial features from the continental United States (CONUS) region to predict ambient acoustic sound levels at any given location in the United States. This model relies on finding obvious and non-obvious connections between CONUS geospatial data as inputs and given training points of measured sound level data at various locations across the CONUS region as outputs; then make

sound level predictions across the entire area itself. As would be expected, the more training data that you can give to the model, the more likely it will be able to predict an accurate sound level at the more numerous unmeasured locations.

Unfortunately, the acquisition process of ambient acoustic sound level data is relatively expensive and time consuming. In order to obtain a measurement of high enough fidelity, a weeks worth of recorded data is required for each individual location. When taken into the context that robust machine learning models can require sets of training data in the range of up to 10^4 to 10^6 data points[reference here], it becomes evident that obtaining measurements in these amounts would take considerably more time and resources than would be realistically available.

This raises a new question: With the amount of time commitment needed for a single data point, how do we choose which new locations to sample from? Choosing where to acquire new data can be an ambiguous process, especially since the effects on model improvement and diversification can be difficult to quantify. In a scenario where obtaining training data is relatively cheap, the answer would be to continuously sample the input population at random and eventually the training data and input population should look statistically similar. However, due to the high cost of acquisition, high dimensional, and diverse nature of geospatial data, the challenge arises in choosing new locations that are not already represented statistically in the current training data set.

By analyzing properties of the input data population combined with the features of our current training data set, using a K-Means clustering algorithm, we can glean quantitative insights into how closely the training data represents the input population. These insights can then inform the process of acquiring new training data whose addition will then better represent the input data population.

1.2 Previous work

There has been extensive work done in the field of data science using unsupervised machine learning [1] [2]. However, most studies involving clustering analysis are focused on a specific sets of data and published work on machine learning methods used specifically on geospatial data is sparse.

There is, however, a large community devoted to the collection of geospatial data using satellite imaging, weather monitoring systems, statistical topography, and other means under the field of geographical information systems (GIS) [3] [4]. The CONUS geospatial data used in this analysis was collected from these various GIS databases and formatted by the BBRC for the purpose of developing a supervised machine learning model to predict ambient outdoor noise levels. The work contained in this thesis was done in conjunction with the machine learning model development in hopes of informing future data sampling of outdoor noise levels for overall model improvement.

An overview of clustering methods and unsupervised machine learning will be covered in detail in a later chapter.

1.3 Overview

The focus of this thesis will be 1) an exploration into the nature of geospatial data and its statistical properties, 2) an explanation of the numerical methods used to perform the clustering analysis, 3) providing physical insights on the acoustic implications of the clustering analysis, and 4) to show how a statistical comparison of cluster occurrence can inform which new locations to sample acoustic data from based on interpretation of the clustering results.

Initial results show a significant disparity in the statistical similarity between the input geospatial data and our current training data set. Additionally, the input data clusters appear to group themselves according to prominent geospatial features based on the standard deviation of the cluster centers.

This allows us to rank geospatial features in order of predictive importance for cluster assignment and can help lead to feature down selection in the future.

Chapter 2

Geospatial and Ambient Acoustic Data

Since we are using geospatial feature data as inputs to predict outdoor ambient noise levels, a general understanding on the nature of geospatial data is required before addressing the clustering methods used to group geographic sites together. In this chapter the goal will be to understand what is meant by the term geospatial feature layer and how such data is collected and compiled. We will also touch on the nature of ambient acoustic data as the outputs of our model and the process by which we make predictions at unmeasured sites.

2.1 The nature of geospatial feature data

A single geospatial feature layer is compiled by picking a single metric to describe a given point, or area, with a quantifiable value given a specific latitude and longitude. These values can be measured in a variety of units such as ratios, proportions of land cover, area densities, frequency of flight observations, distances, measured light intensity from satellite imaging, weather data, etc. Using these values we can then plot a pixel map across a given area using latitude and longitude as axes as shown in Figure 2.1.

While each feature layer is distinct, many geospatial features can be highly correlated to each

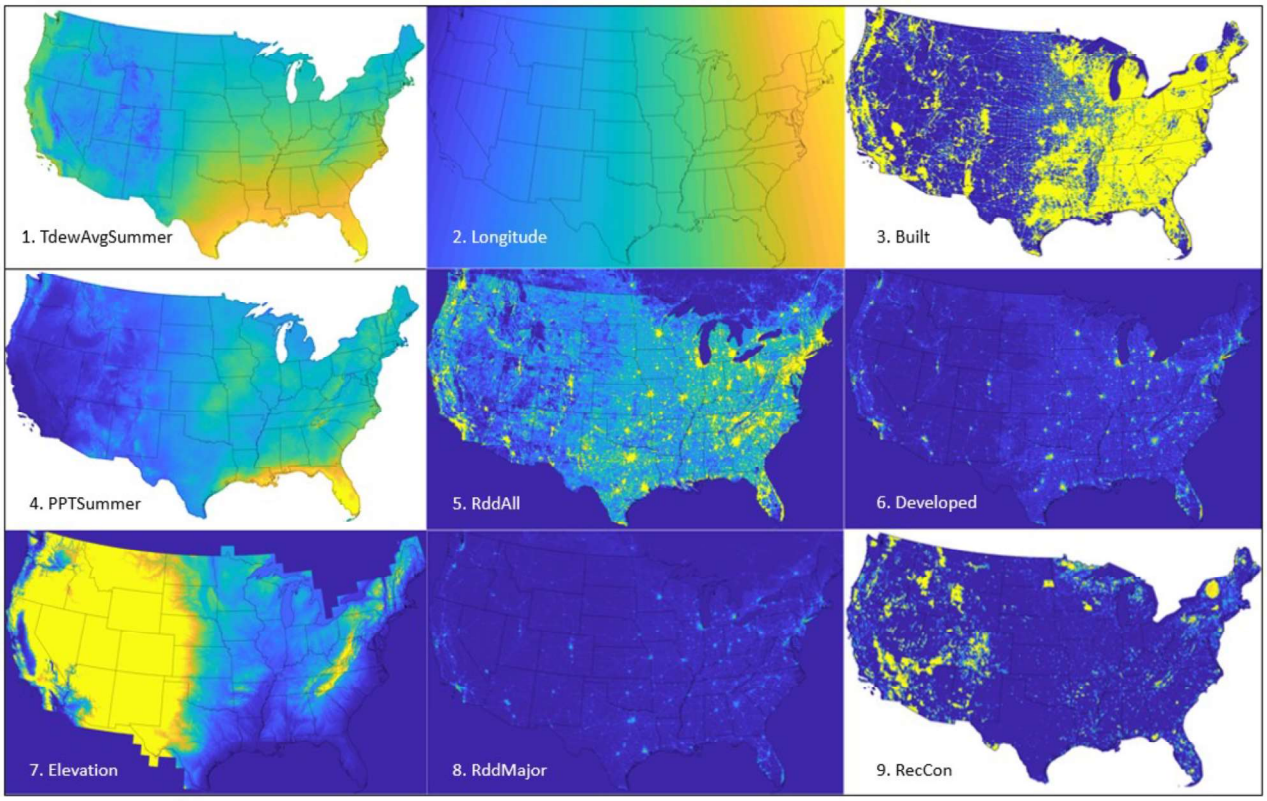

Figure 2.1 A map of several geospatial features. These nine geospatial feature layers represent a few of the 100 feature layers that we use to make predictions on ambient noise levels across the CONUS area shown. The x and y axis of each map are latitude and longitude, the x axis being displayed by the Longitude feature map above (top middle), with color representing a relative numerical value at a given location. Features such as Summer dew point temperature (PPTSummer, middle left) and Elevation (bottom left) are more descriptive of the climate for a given area, while the sum of all road lengths in given area (RddAll, center) and proportion of developed landcover (Developed, middle right) are more indicative of human population levels. For more detailed descriptions on what other feature layers represent see Table 3.1.

other due to common physical factors. For example, road density and mean upward radiance at night are closely related by the factor population levels, although not exactly the same. Similarly, elevation and dew point temperature (PPTSummer) show a strong negative correlation due to related meteorological factors.

However, it is possible for false correlation to arise due to the size of the area of analysis, such as with the high correlation between all of the geospatial feature layers displayed in Figure **??** and longitude (displayed in the upper middle feature map). This is intuitively a poor feature to use for predicting sound levels on a global scale, since geographic areas vary widely and unpredictably across different longitudes, but when analyzed only over the CONUS(Continental United States) area we can see a general correlation that shows a strong difference between the eastern and western areas of the United States that is fairly consisted across many features.

Also of note are the various levels of data resolution and accuracy for geospatial features depending on how each type of data was collected. In some cases several different area resolutions were used to describe the same feature ranging from $200m^2$ to $5km^2$ [add figure showing different levels of resolution for the same geospatial feature]

2.2 Ambient acoustic noise level measurements

Even though the actual ambient acoustic measurements are not used in the clustering analysis of this thesis, a general understanding the process involved for the acquisition of ambient acoustic data is helpful for a full understanding of the challenges involved in selecting and sampling training data site locations.

The first step in taking ambient acoustic data is setting up a sound level meter that is independently powered and protected against the outdoor elements in the location of interest. Then over an extended period of time, usually at least one week, the sound level of the location is sampled over a

set interval of time and averaged to extract the various noise metrics commonly used in community noise measurements. Different noise metrics include L10, L50, L90, daytime vs nighttime levels, frequency band energy, etc.

The L10, L50, and L90 levels are calculated based off of the percentage of time a certain sound level was exceeded during the length of the recording. For an example of one such spectral measurement see Figure 2.2. For each exceedance level metric we can also measure the average frequency band energy. An example of One-third Octave (OTO) band frequency levels for an L50 measurement at one of our training sites, and accompanying predictions from various machine learning models, is shown in Figure 2.3.

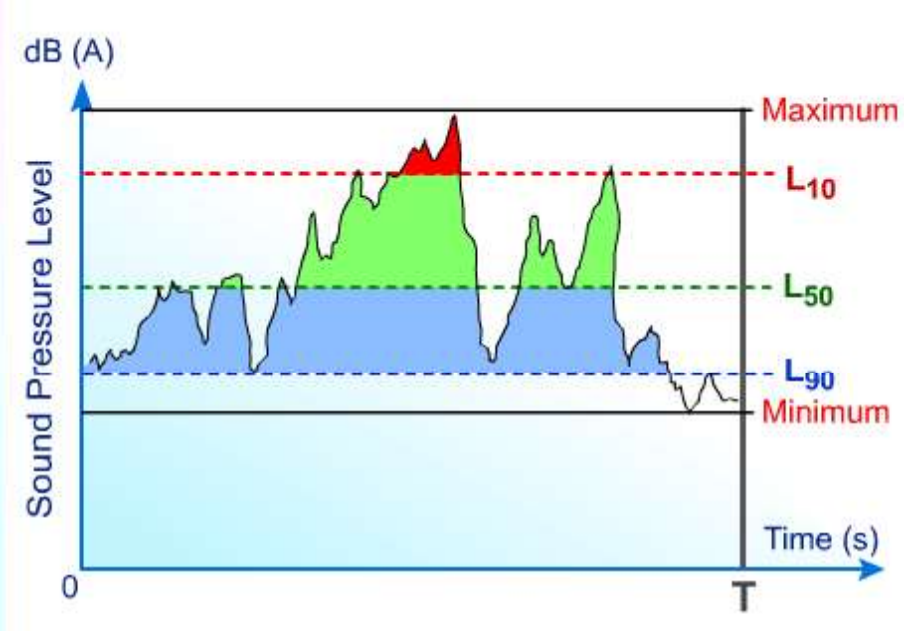

Figure 2.2 Example of how L10, L50, and L90 acoustic metrics are calculated.

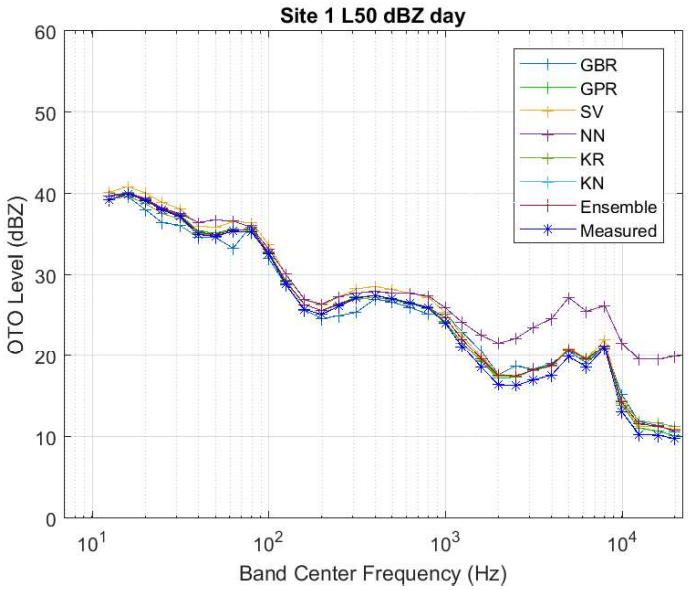

Figure 2.3 Example of ambient acoustic L50 spectral data.

2.3 Applications through machine learning

Using the relationships between geospatial features and measured sound levels we can predict ambient noise levels through supervised machine learning models. Each observed point where ambient acoustic data has been taken has a set of geospatial feature values attached to it for that specific location. By feeding the measured noise level outputs with the attached geospatial inputs into a machine learning algorithm the model can learn the correlations between the ambient noise levels and geospatial feature values. An example of a general supervised machine learning model is shown in Figure 2.4

Our challenge lays in deciding which kind of sites we should sample from next to further diversify our training data. We can better visualize how closely our training data statistically matches our input data through a joint clustering analysis of the geospatial features associated with each training point in conjunction with the entire geospatial feature space that we are trying to make predictions over.

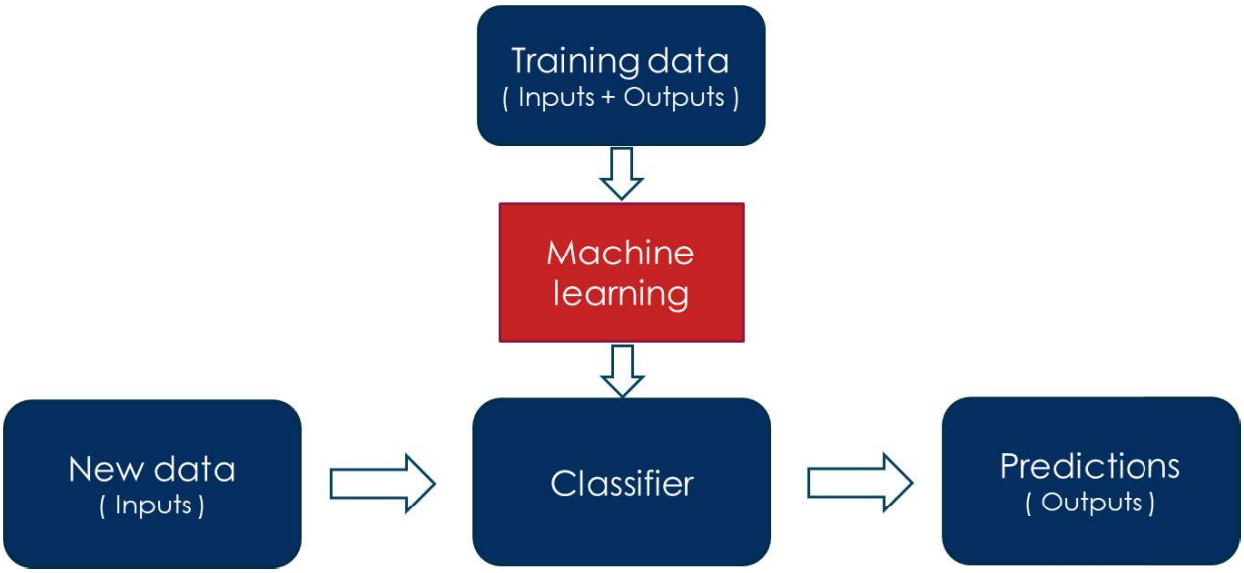

Figure 2.4 Flow chart of supervised machine learning model.

Chapter 3

Clustering Methods

3.1 Unsupervised Machine Learning

The field of machine learning includes many methods which are used to make predictions and draw interpretations from various sets of data. A very broad overview of machine learning categories is displayed in Figure 3.1. The K-Means clustering analysis used in this thesis falls under the category of unsupervised machine learning; meaning that instead of giving an algorithm a set of training outputs to learn from in order to make predictions you provide only the inputs and let the algorithm categorized the data on its own. This allows us to use only the input data to find patterns in the data without knowing the "true" values of the outputs.

For a review on principles of statistical machine learning see Jain et al, 2000 [5].

3.2 K-Means Clustering Analysis

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. Each cluster is assigned a centroid , which is a point denoting the center of each cluster,

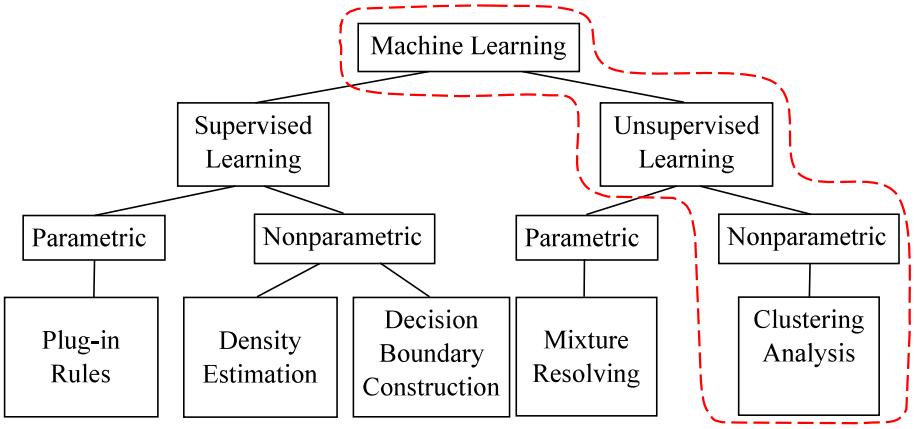

Figure 3.1 A global view of machine learning practices.

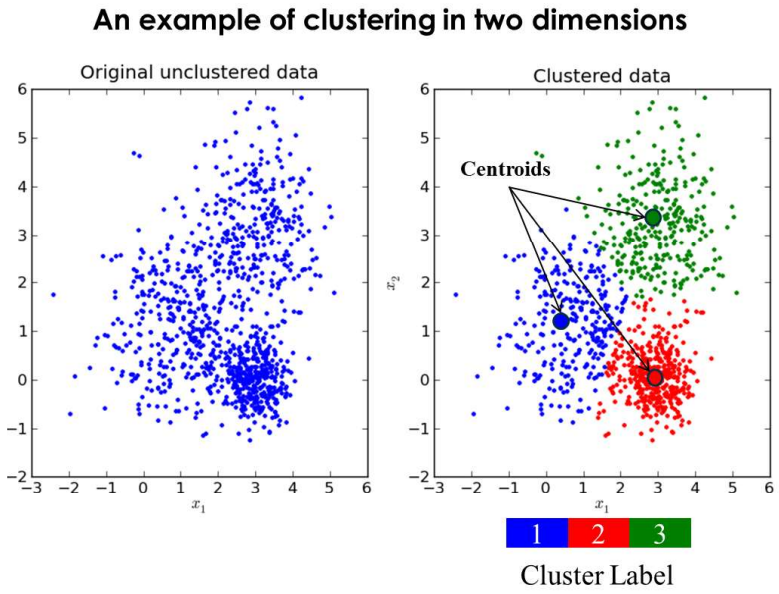

Figure 3.2 An example of a two dimensional data set that can be grouped into clusters.

which is moved iteratively until all centroids have reached a stable position according to a set threshold [1] [6]. An example of two dimensional clustering is shown in Figure 3.2 where $K = 3$. Additionally, a simplified flow chart of the K-means clustering algorithm is shown in Figure 3.3.

While it is easier to discern possible cluster patterns in two dimensional data, higher dimensional

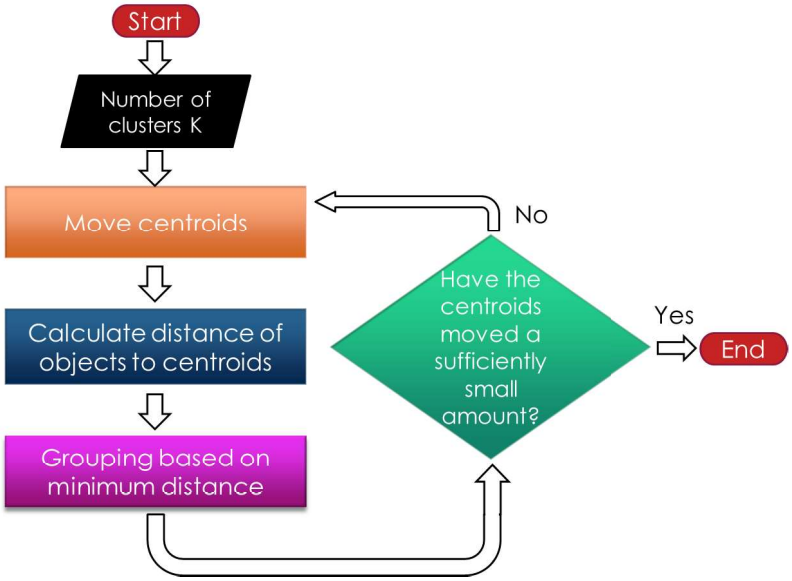

Figure 3.3 A flow chart of the K-means clustering algorithm.

input data groups can be harder to visualize. Since our current geospatial data set has over 100 feature layers, choosing the number of clusters K to give the algorithm is less intuitive than visually assigning clusters to a two dimensional data set. Instead we can perform a simple analysis to determine approximately the optimal number of clusters by determining the point of diminishing returns for the average centroid density for each number K up to a certain limit as shown in Figure 3.4.

This method is not meant to give an exact analytical answer as to the optimal number of clusters, rather it is meant more to give a general idea as to the proper range of K that we should be using. The number $K = 12$ was chosen since it appears around that point that we begin seeing diminishing returns on the average cluster centroid density.

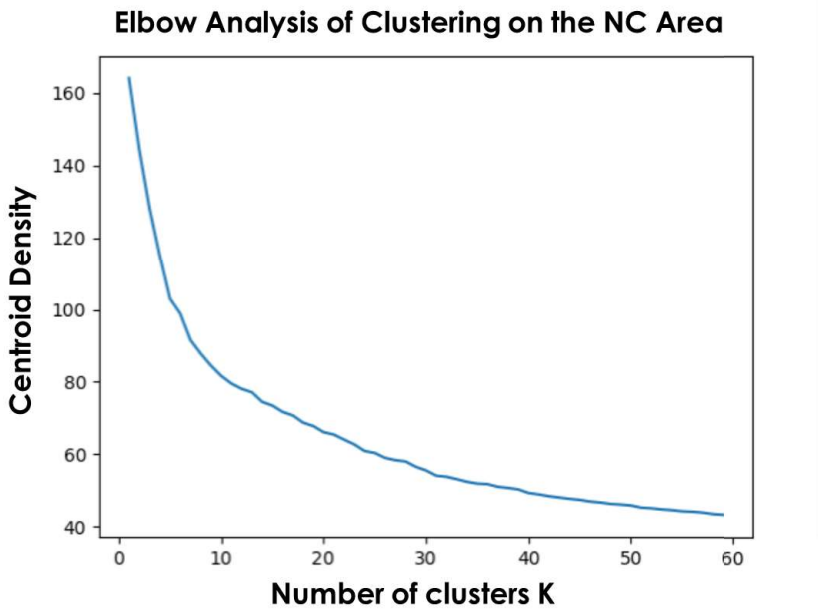

Figure 3.4 A plot of average centriod density over number of clusters .

3.3 Scaling of Input Feature Data

Many of our geospatial feature layers have different units of measurement ranging from percentages and proportions to distances and temperature values. Consequently, the numerical values can differ from layer to layer by orders of magnitude, such that we found it necessary to scale our data to avoid unintentional weighting on features with generally higher numerical values than others. We achieved this using two kinds of scaling methods: standard scaling and logarithmic scaling.

The process of standard scaling involves subtracting the mean of each feature from itself and then scaling it according to the variance of each feature. This allows our algorithm to more easily compare features by how much variation occurs within each feature the same mean zero scale.

Logarithmic scaling sets limits to what range we think a certain feature might be relevant intuitively by simply putting the feature on a log scale relative to the maximum value of relevance. Since there are several distance based features, such as distance to coastline or to nearest airport,

that we would expect to become negligible at large distances with little difference between 100 miles and 500 miles when it comes to predictive importance. Additionally, proximity to noise sources such as roads become drastically more important for determining noises levels in a way that is intuitively non-linear.

3.4 Calculation of Cross-entropy

We can calculate similarity between our training data set and input map through a metric commonly used in data science called cross-entropy . After a joint clustering of the two data sets we can calculate the percentage of occurrence for each cluster in each set by dividing the total number of data points assigned to a given cluster by the total number data points in the set, effectively create two probability density functions (PDF). We will call these two PDFs f_i, for our input map, and g_i, for our training data.

While comparing these two PDFs and their differences is informative in itself, to obtain a more analytical measure of their similarity we can then calculate the cross-entropy between the two functions as follows

$$E = \sum_{i=1}^{K} f_i \ln\left(\frac{f_i}{g_i}\right) \tag{3.1}$$

The units of this calculation are arbitrary, but if a cross entropy value is lower compared to a previous iteration of added training data it denotes that the two PDFs have become more similar to each other than they were before the new data was added. For more information and background on cross-entropy and related calculations see Gokcy et al, 2002 [2].

3.5 Feature Ranking

By analyzing the standard deviation of the cluster centriods across each geospatial feature we can get an idea as to the feature's predictive importance in cluster label assignment. If a feature has high

variation over the range of cluster centroid locations it is indicative of the feature being spread more across many clusters. Consequently, a value change in said feature would be more likely to change the cluster label of a given point. A list of the top 20 ranked features according to variance across centroid locations from a clustering analysis on the NC area is shown in Table 3.1.

Additionally, for reference later in the analysis of the clustering results, geospatial maps of several of the top ranked features are shown in Figure 3.5.

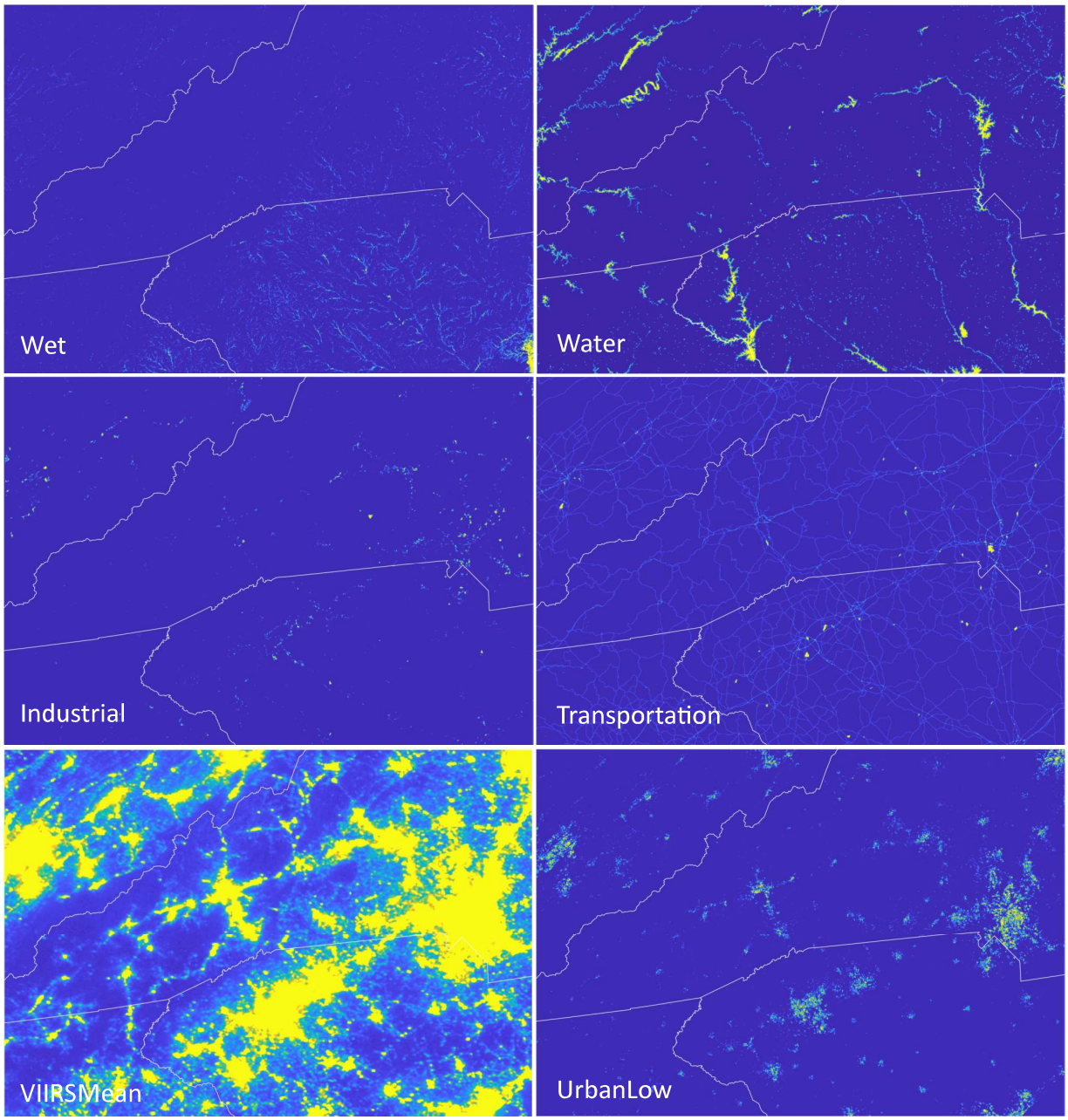

Figure 3.5 Feature maps of higher importance in determining cluster assignment.

Table 3.1 A list of the top twenty ranked geospatial features out of a total of one hundred features. Rank is determined by the standard deviation of all cluster centers across a given feature space, with a larger standard deviation over a single feature indicating higher predictive power in determining which cluster a data point is assigned to.

Rank	Feature Name	Description
1	Wet	Degree of human modification from wet land use
2	Water	Proportion of water (only) landcover
3	Industrial	Degree of human modification from industrial land use
4	Transportation	Degree of human modification from transportation land use
5	VIIRSMean	Mean upward radiance at night
6	VIIRSMinimum	Minimum upward radiance at night
7	VIIRSMaximum	Maximum upward radiance at night
8	UrbanHigh	Degree of human modification from high urban land use
9	UrbanLow	Degree of human modification from low urban land use
10	WaterNat	Degree of human modification from natural water land use
11	Commercial	Degree of human modification from commercial land use
12	Deciduous	Proportion of deciduous landcover
13	RddAll	Road density, sum of road lengths (all roads) divided by area of interest
14	PhysicalAccess	Travel time given transportation infrastructure and off-trail permeability
15	Shrubland	Proportion of shrubland landcover
16	FlightFreq	Total weekly flight observations
17	Cultivated	Proportion of cultivated landcover
18	Timber	Degree of human modification from timber land use
19	Grazing	Degree of human modification from grazing land use
20	Forest	Proportion of forest landcover

Chapter 4

Results and Analysis

There is much interpretation that can be draw from the clustering results over North Carolina, but for the sake of brevity the results will be displayed in full with a few comments from the author as to the meaning of the clustering patterns and the implications it has for the acquisition of data in the future. Additionally, it should be noted that these results are localized solely to the NC area and that clustering results may change when analyzed in the context of the entire continental United States (CONUS) region.

4.1 Cluster Maps

Since visualizing clusters in over 100 dimensions is non-intuitive and abstract we can instead project our cluster labels down into the two dimensional space of latitude and longitude to effectively give us a cluster label map of our chosen geographic area of analysis. This dimension reduction allows us to visualize how a given physical site location would be labeled in comparison to other sites on the map.

Figure 4.1 shows the results of clustering the geospatial features across North Carolina into twelve clusters, with black dots denoting locations of training sites used in the predictive machine

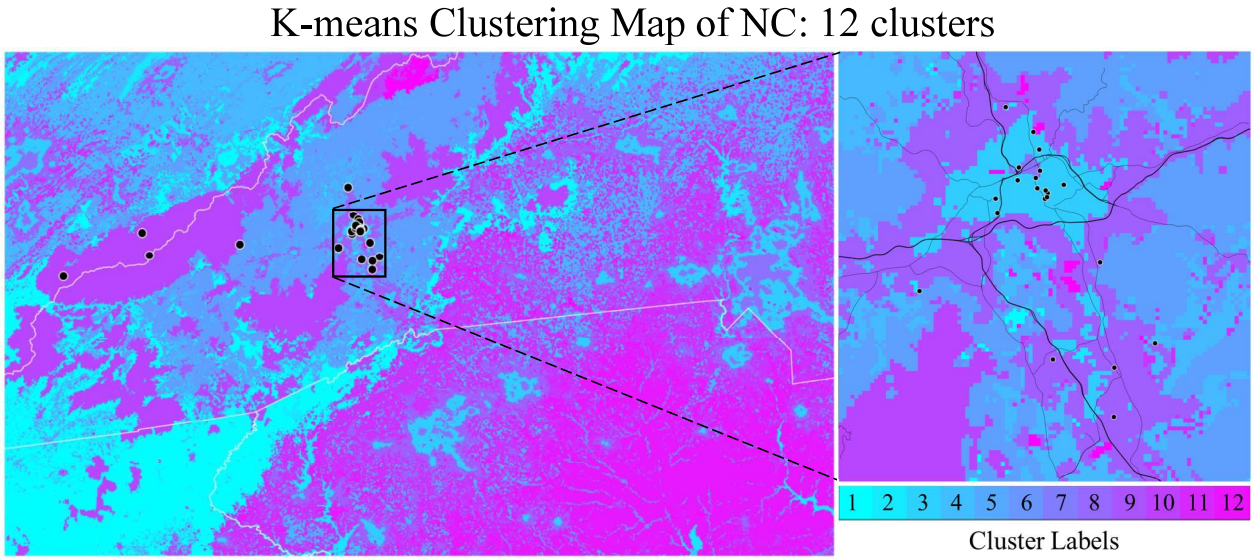

Figure 4.1 A clustering map of North Carolina with magnified portion of Ashville, NC. The black points represent sites where ambient acoustic data has been sampled as training points for the supervised machine learning model. Black lines represent major roads while white lines represent state boundaries. Each individual cluster is represented by an arbitrary and discrete integer color value from 1-12.

learning model. In this case colors are arbitrary and meant solely to differentiate one cluster label from another. For reference, a road map of the NC area with locations of major cities labeled in Figure 4.2 has been provided for comparison with the individual cluster maps shown in later figures.

Additionally, each cluster label was mapped separately for easier distinction of the organization of each cluster label on the input map as shown in Figures 4.3, 4.4, 4.5, and 4.6.

4.2 Feature Importance

By comparing the cluster map to several of the higher ranking feature maps shown in Figure 3.5 we can see several clusters grouping themselves strongly according to a select handful of features. For example, when compared with the geospatial feature of mean upward radiance at night (VIIRSMean) at least three separate clusters appear to form in a gradient like pattern according to the level of

Figure 4.2 A road map of the NC area showing major cities for comparison with the cluster maps generated by our K-means clustering of geospatial features.

brightness in an urban location at night shown in Figure 4.7. When compared to a map of the Ashiville city limits shown in Figure 4.8 it is apparent that the clustering algorithm finds different levels of urbanization to be important in determining cluster label, with the most urban at the center surrounded by a semi-urban area near the downtown area, then a larger suburban area surrounding the semi-urban area.

Features such as bodies of water (Wet and Water), shown in Figure 3.5, are also strong indicators of a specific cluster assignments and almost exclusively follow the features definition for clusters 2 and 6 shown in Figures 4.3 and 4.4.

4.3 Input Map vs Training Data

By comparing the occurrence of each cluster label in the NC area compared with the occurrence of each cluster label in our training data, shown in Figure 4.9, a general idea can be obtained as to how

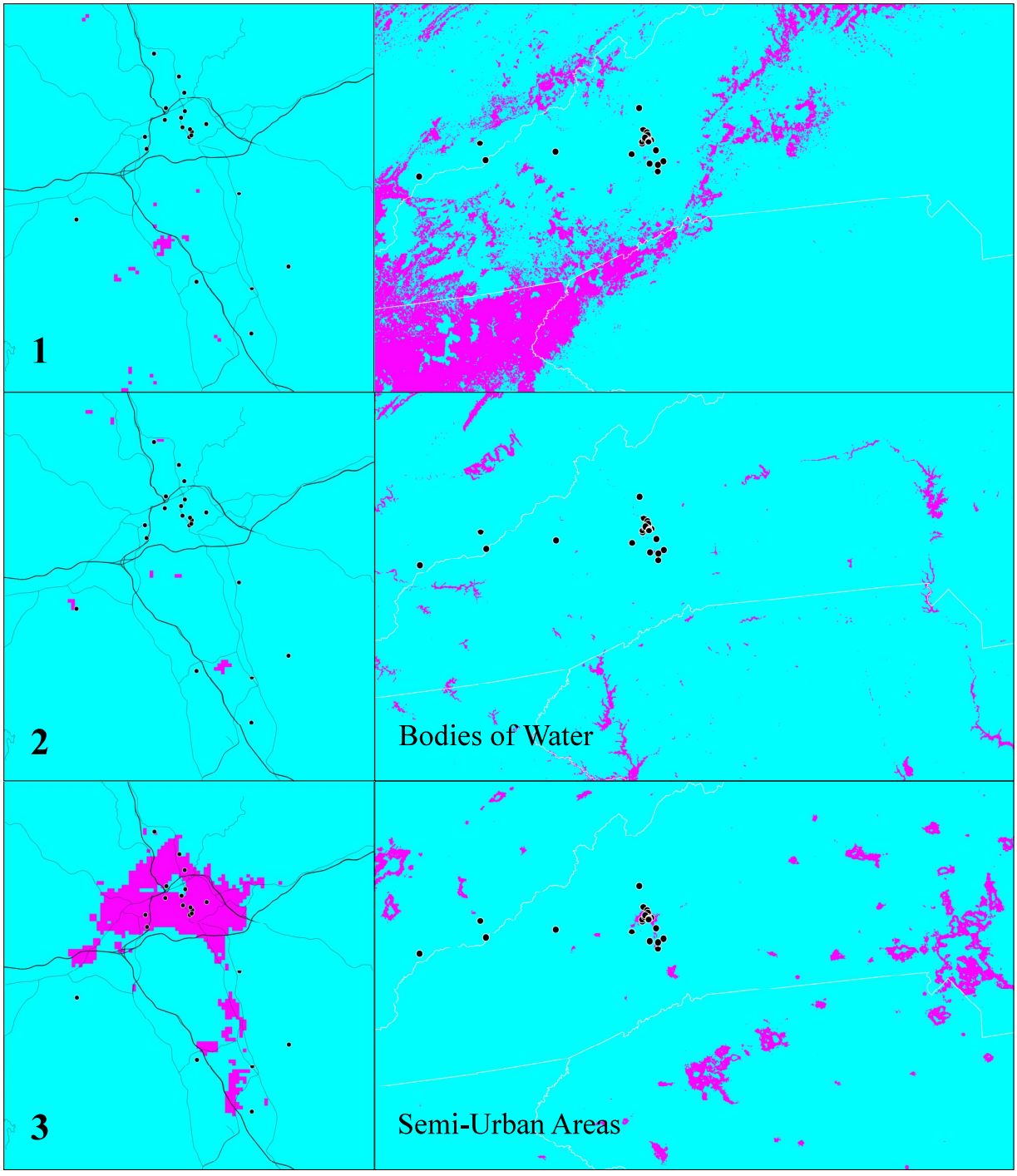

Figure 4.3 A visualization of NC clusters separated from each other. A zoomed in portion of Ashville, NC (left) and full map of the NC area (right) are show for cluster labels 1-3, with the cluster label number being shown in the bottom left corner of each pair. Some clusters more obviously follow certain features and when appropriate interpretive labels are given to the clusters on the right.

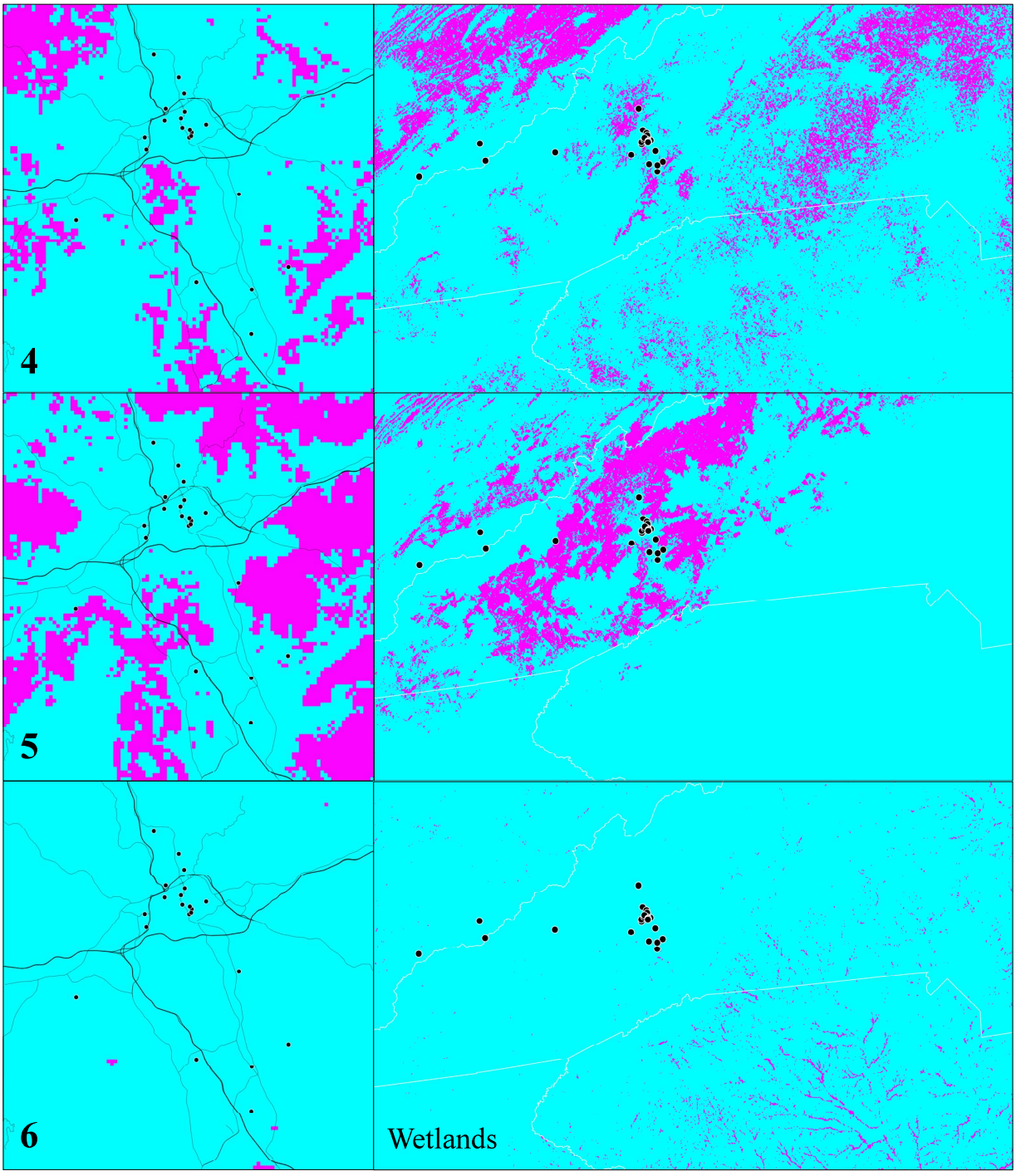

Figure 4.4 A visualization of NC clusters separated from each other. A zoomed in portion of Ashville, NC (left) and full map of the NC area (right) are show for cluster labels 4-6, with the cluster label number being shown in the bottom left corner of each pair. Some clusters more obviously follow certain features and when appropriate interpretive labels are given to the clusters on the right.

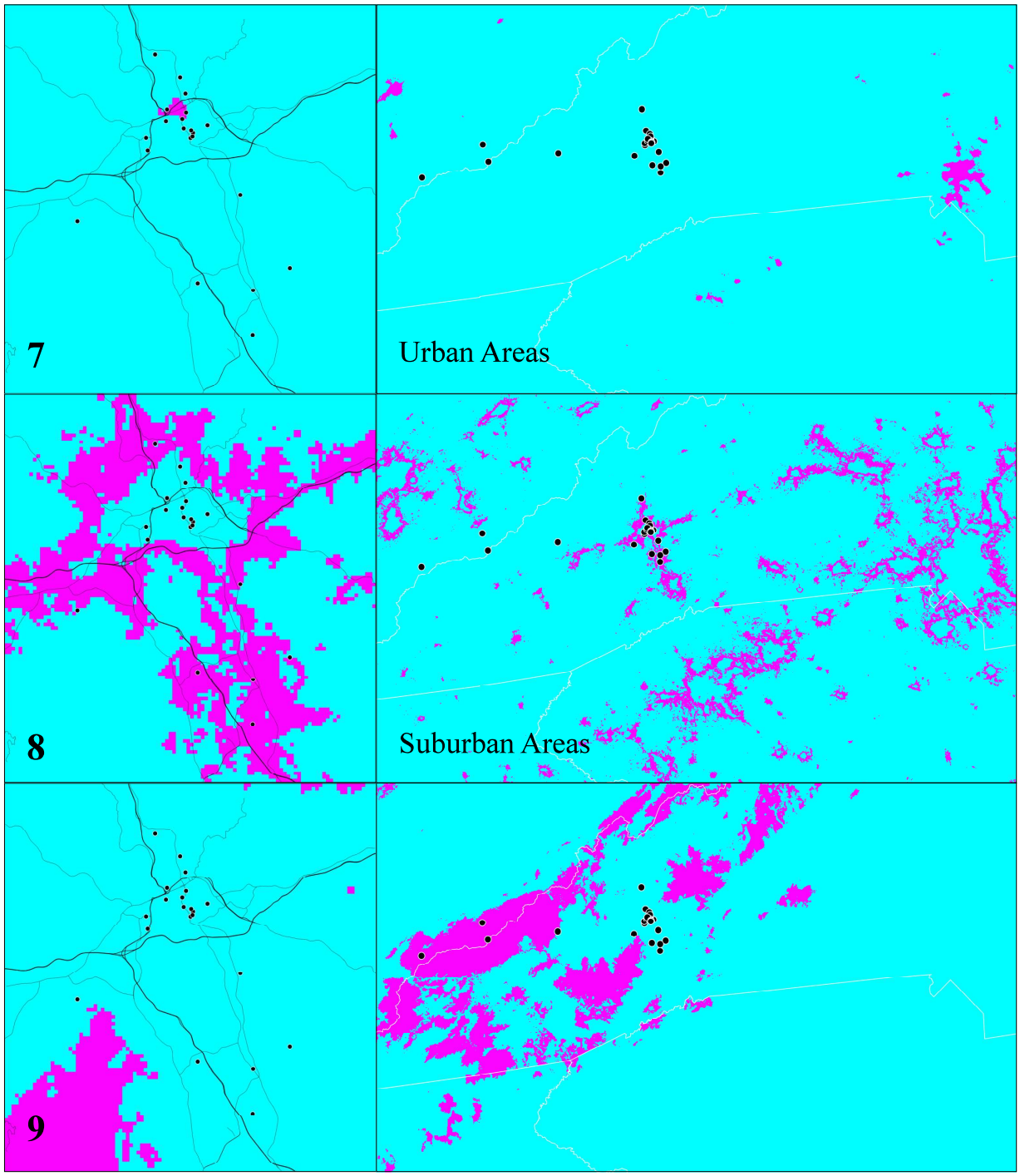

Figure 4.5 A visualization of NC clusters separated from each other. A zoomed in portion of Ashville, NC (left) and full map of the NC area (right) are show for cluster labels 7-9, with the cluster label number being shown in the bottom left corner of each pair. Some clusters more obviously follow certain features and when appropriate interpretive labels are given to the clusters on the right.

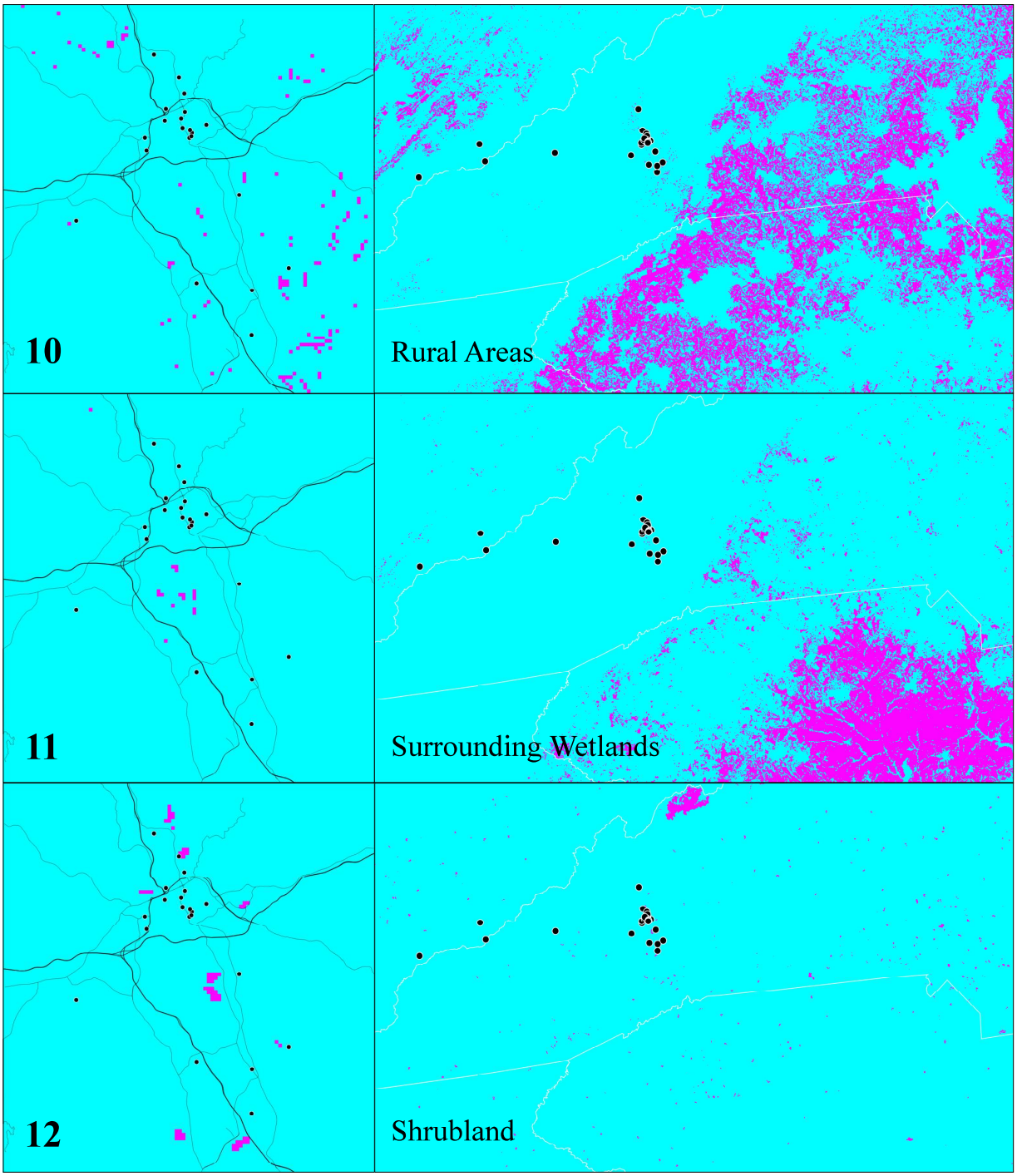

Figure 4.6 A visualization of NC clusters separated from each other. A zoomed in portion of Ashville, NC (left) and full map of the NC area (right) are show for cluster labels 10-12, with the cluster label number being shown in the bottom left corner of each pair. Some clusters more obviously follow certain features and when appropriate interpretive labels are given to the clusters on the right.

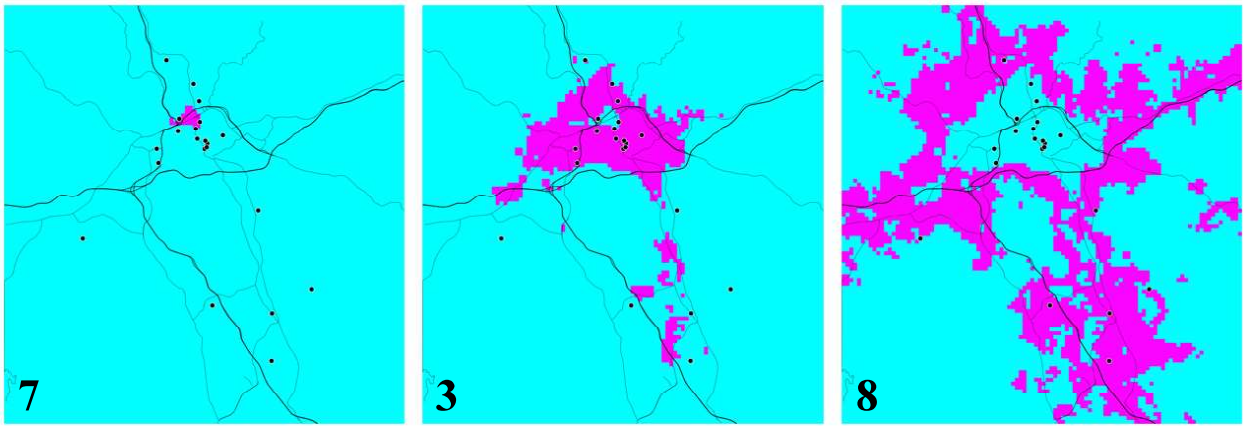

Figure 4.7 Clusters 7, 3, and 8 display different levels of urban population around the city of Ashville as three separate clusters arranged from most the densely populated and urbanized areas to more suburban class areas when compared against the map of Ashville as seen in Figure 4.8.

Figure 4.8 Map of the Ashville city limits and surrounding area.

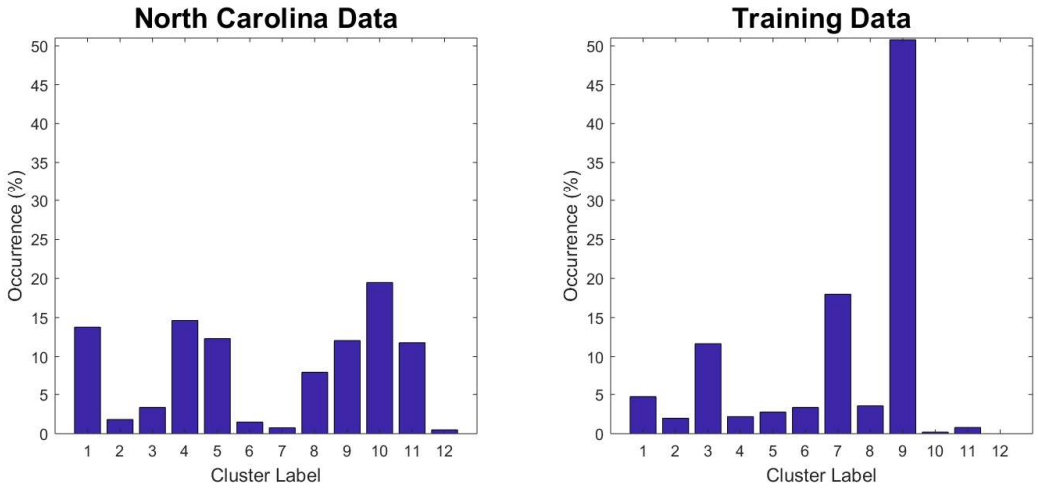

Figure 4.9 The occurrence of each cluster on the NC map by percentage of site locations (left) compared with the occurrence of each cluster in our training set (right).

well our training data represents the area over which we are trying to make predictions.

It is immediately apparent that the training data set has some major statistical difference compared to the the NC area. For example, a large portion of our training sites fall under the cluster label of 9 whereas the NC has a much smaller proportion of sites labeled under the same cluster. There are also clusters such as 10 and 11 that show up frequently in the NC area but are hardly represented at all in our training set. Using this kind of analysis we can infer which kinds of sites need to be sampled from next in order to make our training set more similar to our area of interest.

Chapter 5

Colnclusions

The results of this initial analysis show that the current training set used to make acoustic predictions over the continental United States (CONUS) area may not be fully representative of the full geospatial population. Additionally, out of the 100 features used in the analysis the clustering process seems to pick up on a select handful when determining which label a particular location receives. This feature importance may be useful in informing a reduced input feature set for future supervised machine learning models

5.1 Optimization of Training Data Acquistition

Using a statistical comparison of cluster occurrence between our area on interest and our current training set we can select new sites to sample from based on what is lacking or imbalanced in our training data. This greatly reduces the guess work involved in the selection of new locations to add to our training data by giving an empirical measurement of improvement in terms of statistical similarity.

5.2 Future Work

It is hoped that this same geospatial analysis can be performed over the entire CONUS area, since we are interested in making predictions across all of CONUS and not just NC. There are however issues of memory size in performing clustering calculations on such a large number of sites, but with further optimization and use of larger computing resources, such an analysis is feasible. Once a full analysis is obtained the results can be updated at each addition of new training data.

Bibliography

[1] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE TRANSACTIONS ON NEURAL NETWORKS **16,** 645–678 (2005).

[2] E. Gokcay and J. Principe, "Information theoretic clustering," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE **24,** 158–171 (2002).

[3] L. ANSELIN, "LOCAL INDICATORS OF SPATIAL ASSOCIATION - LISA," GEOGRAPHICAL ANALYSIS **27,** 93–115 (1995), GISDATA (Geographic Information Systems Data) Specialist Meeting on GIS (Geographic Information Systems) and Spatial Analysis, AMSTERDAM, NETHERLANDS, DEC 01-05, 1993.

[4] C. DALY, R. NEILSON, and D. PHILLIPS, "A STATISTICAL TOPOGRAPHIC MODEL FOR MAPPING CLIMATOLOGICAL PRECIPITATION OVER MOUNTAINOUS TERRAIN," JOURNAL OF APPLIED METEOROLOGY **33,** 140–158 (1994).

[5] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE **22,** 4–37 (2000).

[6] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE **24,** 881–892 (2002).

Index

Average centroid density, 13

Cluster centroid, 11
Cross entropy, 15
Cross entropy equation, 15

Geographical information systems (GIS), 3
Geospatial feature layer, 5
Geospatial feature ranking, 16

K-means clustering, 11

L10, L50, L90 acoustic metrics, 8
Logarithmic scaling, 14

One-third Octave (OTO) frequency bands, 8

Standard scaling, 14
Supervise machine learning, 9