Lab 2 - Linguistics Data, Stat 215A, Fall 2021

Sahil Saxena

October 07, 2021

1 Introduction

Dialects, or linguistic variations, may contain interesting insights about the social structure of a group. As seen in the survey of American dialects, collected by Bert Vaux, dialects can vary according to geography, social class, sex, and age. Consequently, they can potentially reveal the underlying social identity and shared history of groups with similar dialects. The data is a result of surveying over 47 thousands individuals in all states and districts of the USA. The results in this report mostly support traditional ideas of American dialects; using dimensionality reduction and clustering methods, we can learn more about the relationship between dialect groups and regional geography.

2 The Data

The linguistics data is collected by the Harvard Dialect Survey in 2003. The survey consisted of 122 questions that explored the variations in the phonetic and lexical differences in the English language across the USA. This report focuses on analyzing only the 67 questions that explore lexical differences (e.g. "Do you cut or mow the lawn or grass?").

Two different representations of the dataset exist: LingData and LingLocation. LingData includes the categorical responses to 67 survey questions by 47,471 respondents, while also containing location data of each respondent, through state, city and zip code. The latitude and longitude (of the center of each zip code) measurements are also added to each observation.

LingLocation is a binary encoding of the categorical data described above, binned into one degree longitude by one degree latitude squares. This partitions the map of US into 781 cells; the data for each cell is encoded as the sum of binary response vectors for individuals who belong to the cell.

With this data, we can gain insights into how specific responses to specific questions vary over different regions in the USA.

2.1 Data Cleaning

Compared to the previous lab, data quality from this survey is much higher, with fewer missing values and inconsistencies.

For LingData, 1,020 observations were removed for which either longitude or latitude information was missing. This data could have been corrected by inserting the real longitude and latitude for these cities. However there are many missing value zip codes and many discrepancies in the observations that had missing lat and long, such as fictional or overly broad city and town names (e.g. "Chicagoland"). Furthermore, 107 and 98 observations were removed from Hawaii and Alaska, respectively, to restrict the analyses to the 48 contiguous states. The number of samples from these two states is too sparse to impact the entire data set (only 1.06% of total).

Also, some observations have missing responses to some of the survey questions; with a large sample size, a stringent filter was applied to remove any observations with any missing values. Remaining are a total of

39,051 observations for further analysis.

For LingLocation, cells with extreme lat/long values, i.e. longitude < -150 or latitude > 50, (cells from Alaska and Hawaii) were removed, leaving a total of 758 cells. There are also some cases where the number of people who responded to a question in a cell did not add up to the total number of people in that cell. Since this type of missing data was more difficult to filter out, these observations were left in the sample.

Consequently, LingData is used for the main analysis and LingLocation is primarily used to validate these results in the Stability section.

2.2 Exploratory Data Analysis

Here we explore a few survey questions and their potential relationship to geography. Figure 1 shows the distribution of responses, across the USA, to 2 questions: 73 - What is your general term for rubber-soled shoes worn in gym class? and 103 - What do you call the thing from which you might drink water in a school?

Figure 1.Q73 shows the distribution of the two most common answers, which are “sneakers” and “tennis shoes” (removing other answer choices to make interpretability easier). Those who answered “sneakers” are concentrated in the Northeast region while “tennis shoes” is more commonly used throughout the rest of the country; some overlap does exist, which indicates people may have moved from one region to another or began adapting to different terminology in their native regions.

Figure 1.Q103 shows the distribution of the most common answers, which are “drinking fountain”, “water fountain”, and “bubbler.” The term “water fountain” seems to be more commonly used in the West, with a high concentration in southern California. On the other hand, “drinking fountain” seems to be more favored in the Northeast and the Southeast. The Midwest seems to use both terms at similar frequencies. Interestingly, individuals who responded “bubbler” form small isolated groups: Wisconsin, Massachusetts, and Rhode Island; this seems to support the idea that some terminology evolves from local/regional “slang” rather than an official dictionary.

Experimenting with linked brushing (in other/brushing_experiment.html) shows that response to question 73 does not help predict the response to question 103. However, adding another question might give more insight. Specifically, if someone answered “drinking fountain” for question 73 and ”tennis shoes” for question 103, then for question 110 (What do you call the night before Halloween?), it is unlikely to call the night before Halloween “mischief night.” This suggests that the responses to some groups of questions may be correlated with each other.

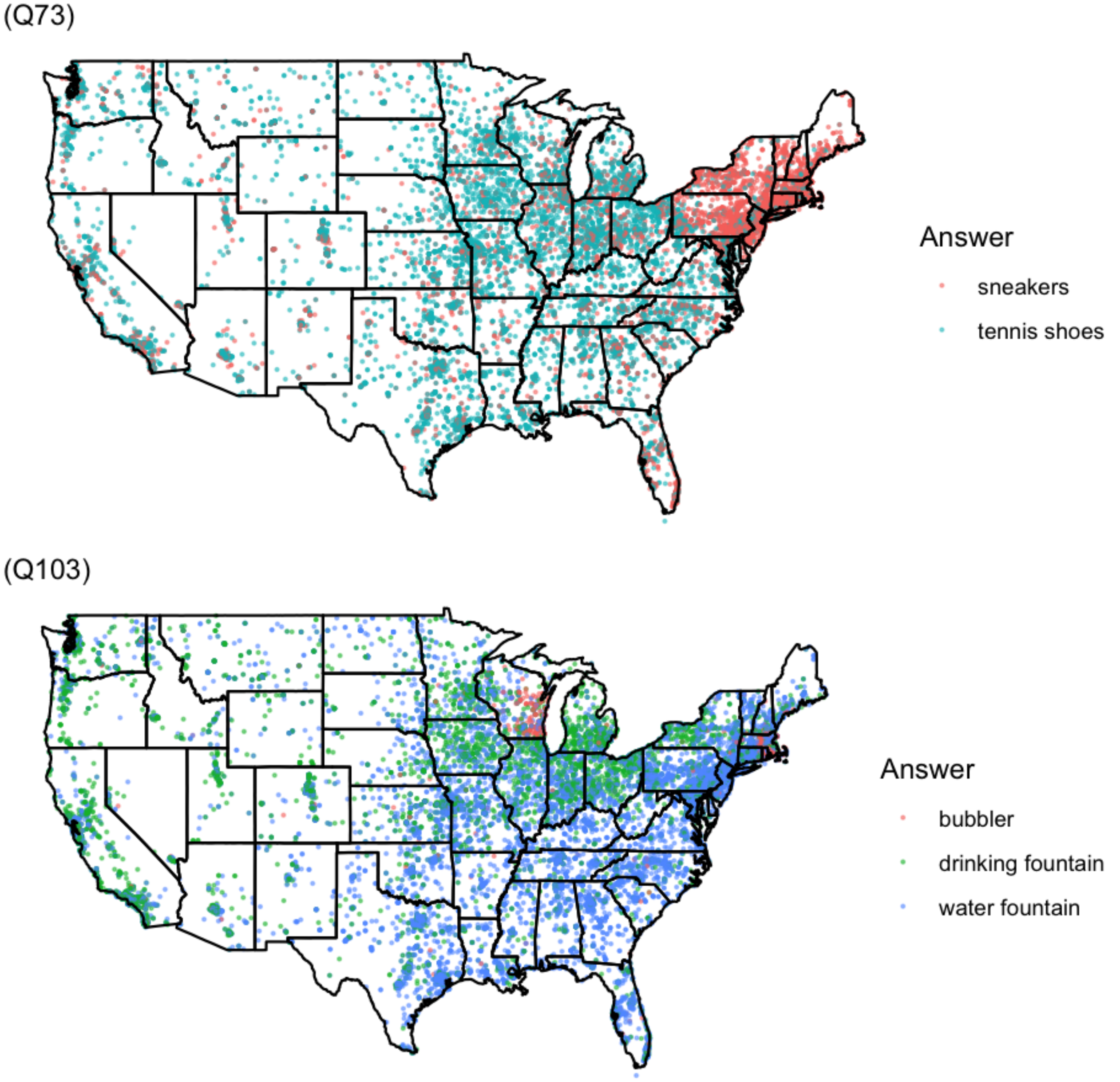

Figure 1: Distribution of responses in the contiguous states. (Q73) What is your general term for rubber-soled shoes worn in gym class? (Q103) What do you call the thing from which you might drink water in a school?

3 Dimension reduction methods

Through the dimension reduction method of PCA, Figure 2 shows the projection of the binary representation of LingData onto the first two principal components. Here, three loose clusters according to longitude can be seen, which may be relating the linguistics results to regional geography. Although these clusters are colored only based on longitude, this provides a starting point for how to cluster the data (see more in Clustering). Furthermore, Figure 3 shows that the first ten principal components explain only a modest proportion (about 20%) of the total variability, which indicates that the underlying cause for linguistic variations is not simple; in other words, it is not accurate to portray the data's complexity in just a few features.

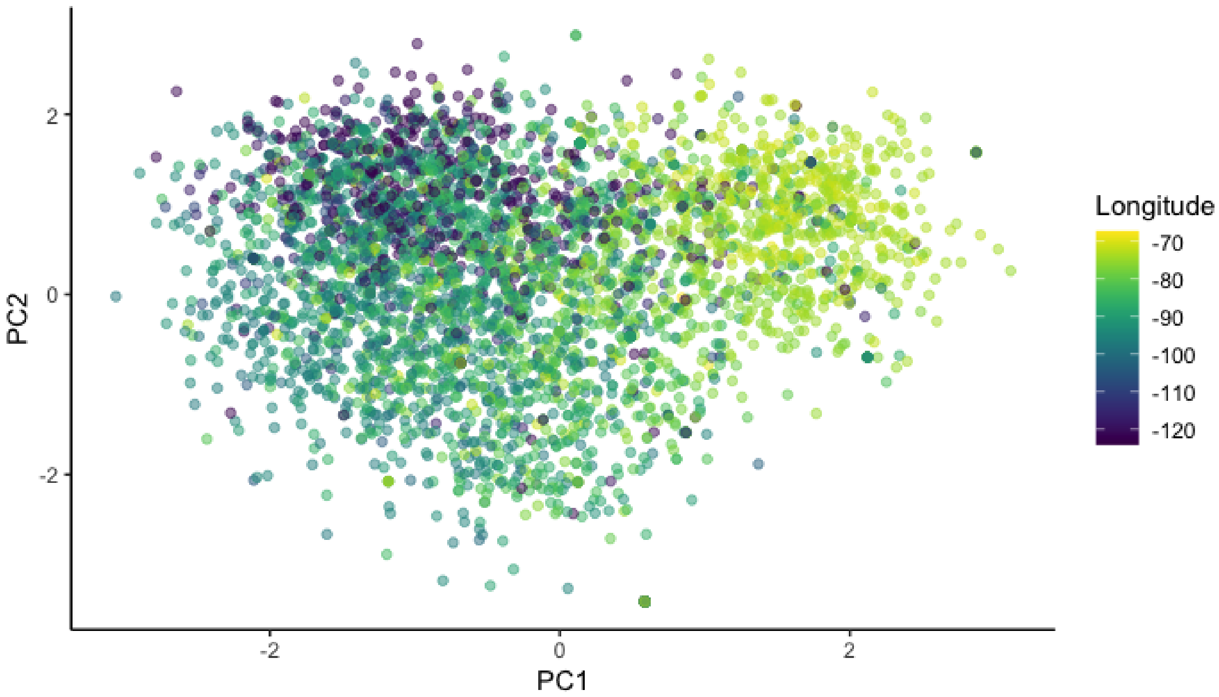

Figure 2: Projection of LingData onto the first two principal components.

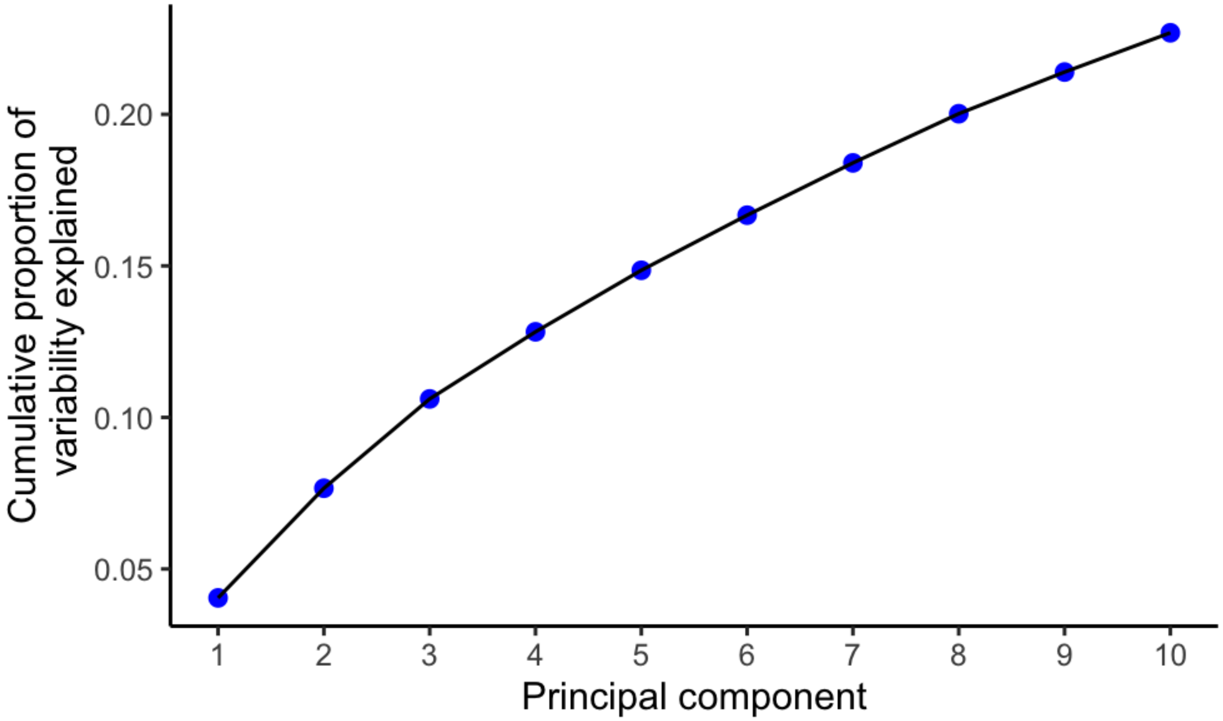

Figure 3: Amount of variability explained by the first 10 principal components.

The data was centered as covariance matrix calculation in PCA already "centers" the data. The PCA was tried with and without sacling. With scaling, the first 10 components cumulatively accounted for about 12% of the variance. However, without scaling, the first 10 components cumulatively accounted for more than 20% of the variance. Clearly, PCA done without scaling has less representative dimensions; this could be due to the fact that the observations are already on the same "scale" as they are binary. With naturally unstandardized data (if the data was not binary encoded), then PCA will project the first component in the direction of the questions with higher covariance, which does not make sense. Given that answer choices are just mapped to numbers, this does not provide a meaningful interpretation.

It is not a good idea to do PCA on the original dataset because the values are inherently insignificant. Answer choices are mapped to a particular number, so for different questions, many individuals will have the "same" answer number; this does not really make sense. By encoding in binary, every answer choice is treated as its own dimension. From there, the dimensionality reduction takes place.

4 Clustering

K-means and NMF clustering approaches are performed on the data.

4.1 K-Means

To further study the potential relationship between dialect groups and geography, k-means clustering on the first two principal components was performed. The number of clusters, k, was set to 3, because it gave the highest average silhouette value. Fig 4 shows the distribution of the cluster members on the map of the US. Here three general geographical clusters are present: 1) the Northeast, 2) the Southeast, and 3) a super region including the Midwest, the Southwest, and the West. The geographical clusters are not divided extremely well and definitely exist on a continuum. There is a fair amount of overlap between clusters on border regions; there are also many outliers, such as the members of the "Northeast cluster" present in Florida.

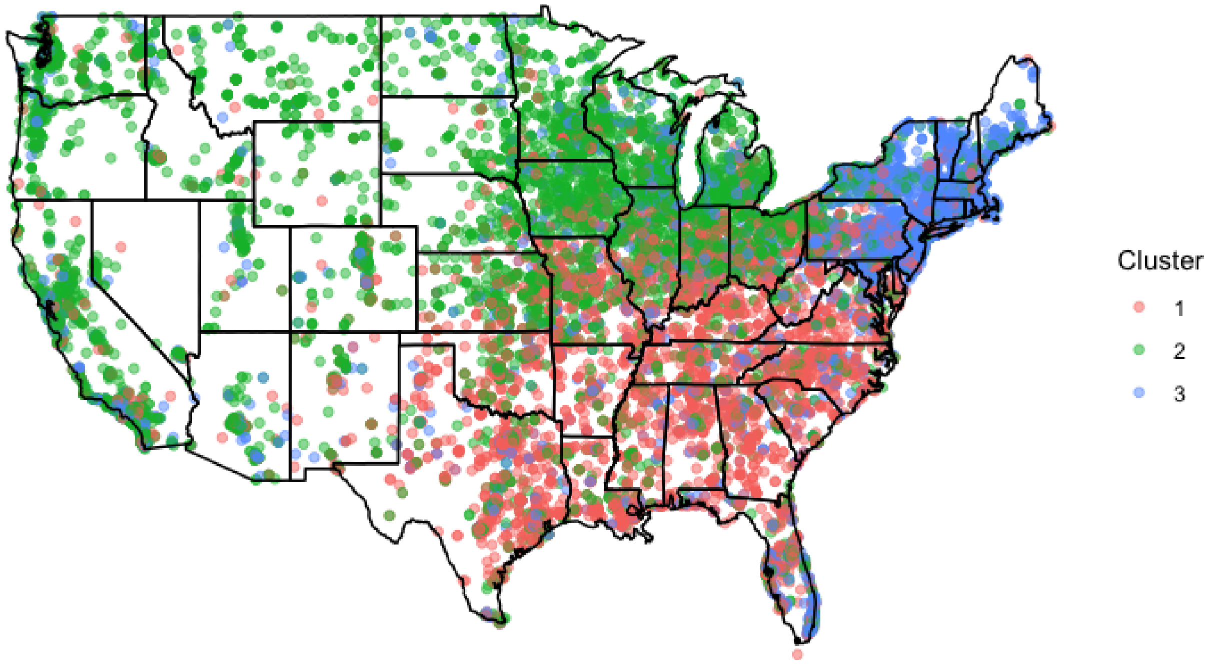

Figure 4: Distribution of clusters, from the k-means clustering with $k = 3$ used on the first 2 principal components. The clusters correspond to three geographical regions.

4.2 NMF Clustering

Using rank-2 matrix factorization, recursive bipartitioning split the data into k = 3 clusters. Min_samples, one parameter of the algorithm, was chosen to be 8000 to force about 3 clusters; with this, a fair comparison can be made with the clusters from the k-means. Fig 5 shows the distribution of the cluster members on the map of the US. The three main clusters are similar to the ones from k-means: 1) the Northeast, which here also includes the southeast, 2) the southern Midwest, and 3) a super region including the northern Midwest and the West. The geographical clusters are not divided extremely well and definitely exist on a continuum, more so than in k-means. There is more overlap between clusters on border regions and many outliers, such as the members of the "Northeast cluster" present in California

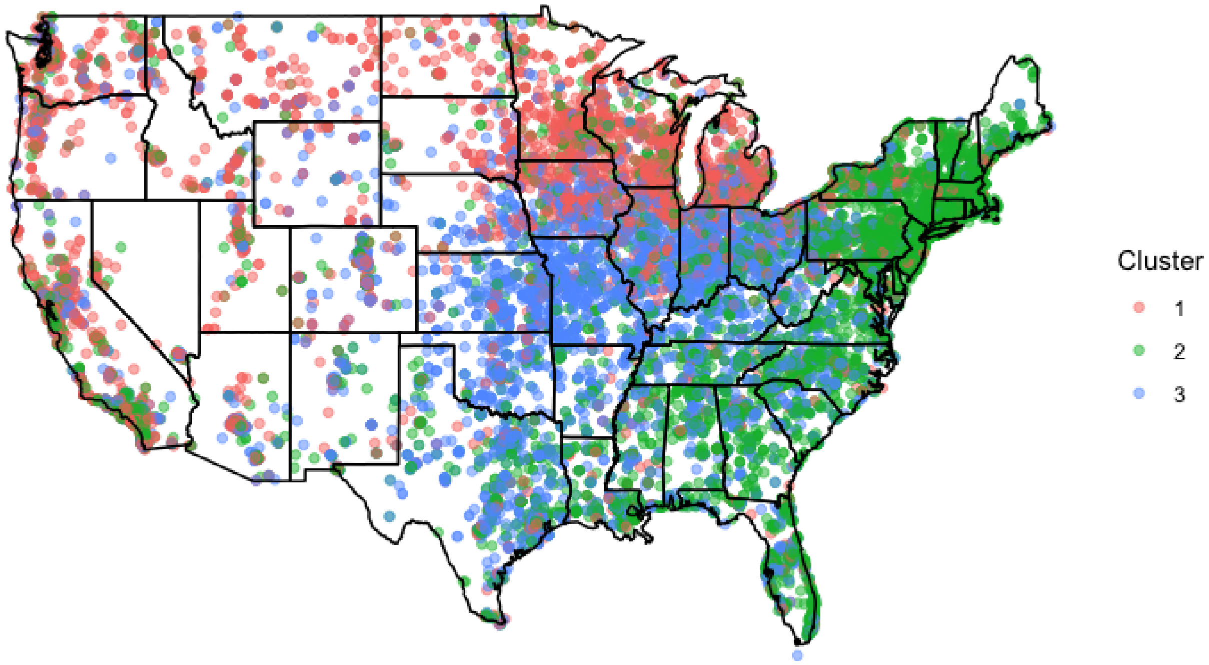

Figure 5: Distribution of clusters, from the nmf clustering with $k = 3$. The clusters correspond to three geographical regions.

4.3 Clustering Analysis

As shown in Q103, individuals who answered "water fountain" lie in the northeast region but also flow into the southeast. This is a pattern that is represented in the NMF clustering result (green color). Also, the answer "sneakers" for Q73 support the geographic split shown in the k-means clusters; specifically the blue cluster matches the region who answered "sneakers". These questions serve as examples, but also clear evidence, of how these geographic linguistic clusters are forming.

The mathematical model behind the dimension reduction strategy (PCA) seem to make sense for these clusters. As seen in Figure 2, there are 3 loose clusters between the first two components based on longitude (east-to-west) which covers a lot of the variation in the clusters generated from the clustering algorithms.

As mentioned earlier, it is important to note that the total variability in the data cannot be explained by a few of principal components. This suggests that the underlying cause for linguistics dataset is very complex and/or the relationships between the variables may not be linear as assumed by PCA. In the latter case, it may be worthwhile to try transforming the variables so that they have more linear relationships or using a nonlinear dimension reduction method.

The NMF approach to clustering has a theoretical advantage in that is using the Cosine distance rather than Euclidean distance as a metric. This implies that it should handle outliers and sparsity better than k-means, which uses Euclidean distance. However, both models have a significant amount of overlap in their clusters.

5 Stability of findings to perturbation

First, the k-means clustering algorithm is run four times with different starting points. Figure 6 shows that each run of k-means produced very similar clusters and our finding, therefore, is stable under different starting points.

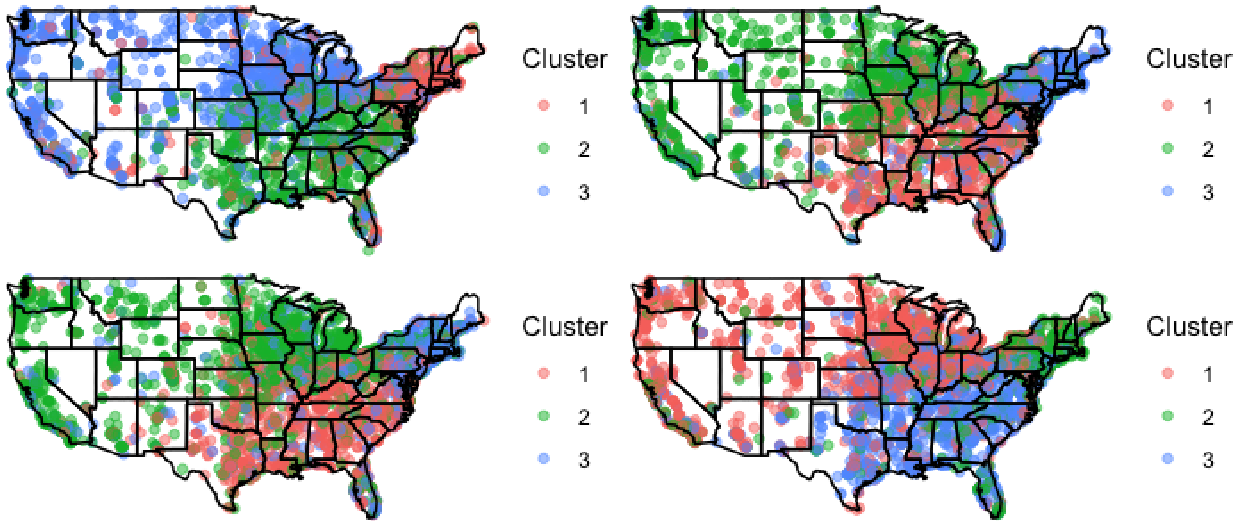

Figure 6: The results of four runs of k-means with different starting points. All four runs show similar geographical clusters to before. The mapping from color to cluster is arbitrary.

Another stability check is to remove data and see how algorithms respond. First, half of the samples were removed; PCA and k-means were run on the first two principal components. As seen in Figure 7, the three geographical clusters are still present on the downsampled data.

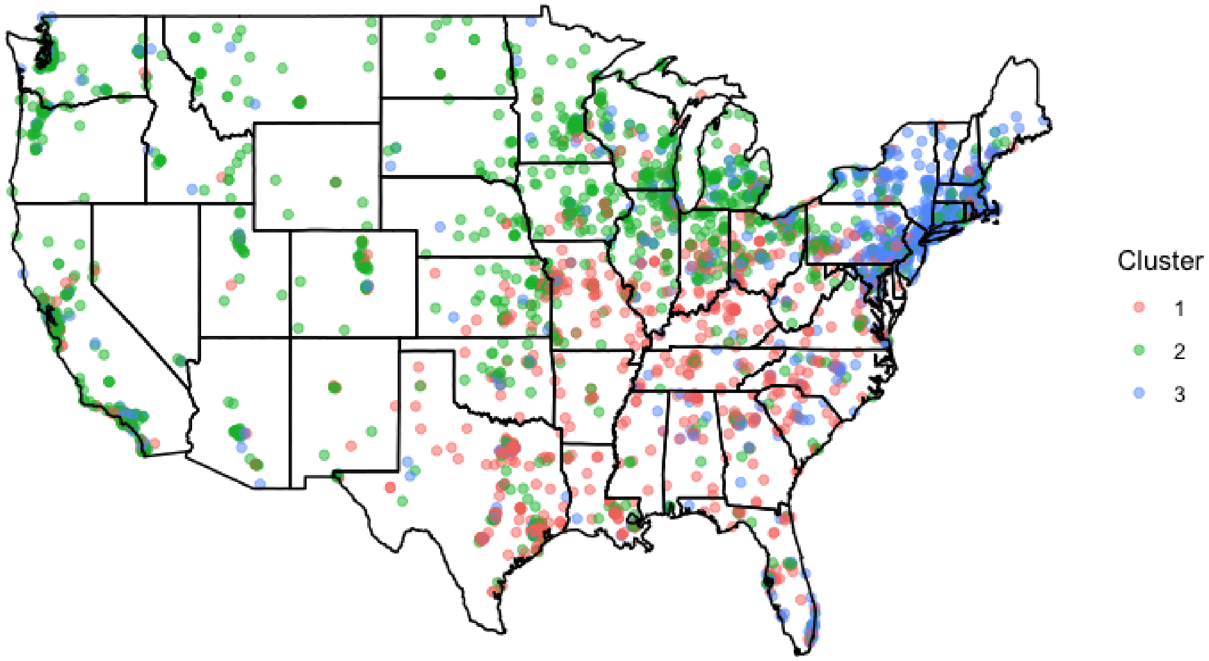

Figure 7: The result of k-means clustering after removing 50 percent of the observations; the three geographical clusters found on the original data are still present here.

To investigate the effects of removing questions from the dataset, questions 73 and 50 were removed, which had the highest loadings for the first two principal components, respectively. Figure 8 shows that there are still similar geographical clusters, implying that many other survey questions in the perturbed dataset are separate the groups in a similar way.

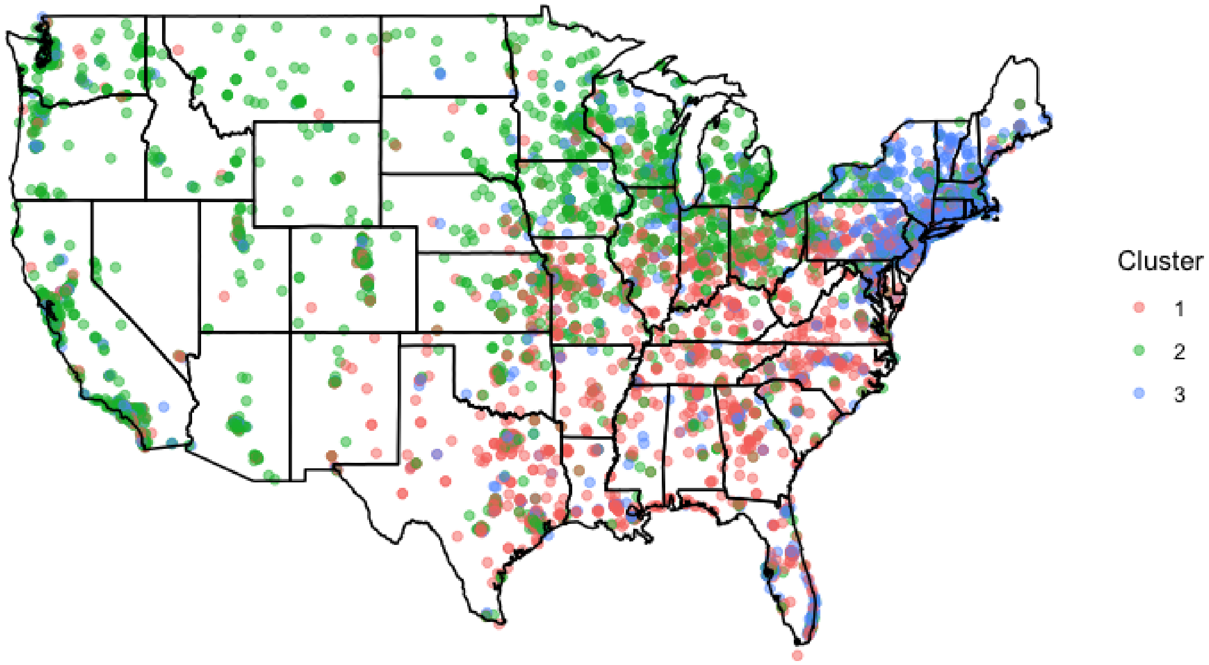

Figure 8: The result of k-means clustering after removing questions 50 and 73, which had the highest PCA loading for the first and second principal components, respectively; the three geographical clusters found on the original data are still present here.

6 Conclusion

The three realms of data science are data, algorithms and analysis, and future data. In this report, the first two realms are heavily focused on. The size of total data in this lab is large and on the spatial dimension, which always sets up interesting anaylsis. However, data exploration is an important step in making further analysis; by first cleaning the data, visualizing it (over space), and then changing the encoding of the data, analysis was made possible. Next, algorithms were used to analyze the data and highlight potential patterns and relationships. Specifically, dimension reduction was done through PCA and clustering methods were used (K-means and NMF); it was interesting to see how two different approaches (purely distance-based versus heirarchical) lended relatively similar clusters. This is an indication of a potential true relationship between certain geographic regions and a linguistic dialect group.

Finally, a reality check for this analysis is to see whether re-running the analysis on a different encoding of the data would produce similar results. PCA and k-means was ran on LingLocation, in which bins the observations in one degree latitude by one degree longitude squares. Figure 9 shows that the three geographical clusters we find are similar to those shown previously. Thus our finding that linguistic variations relate to three distinct geographical groups is stable under perturbations to the data.

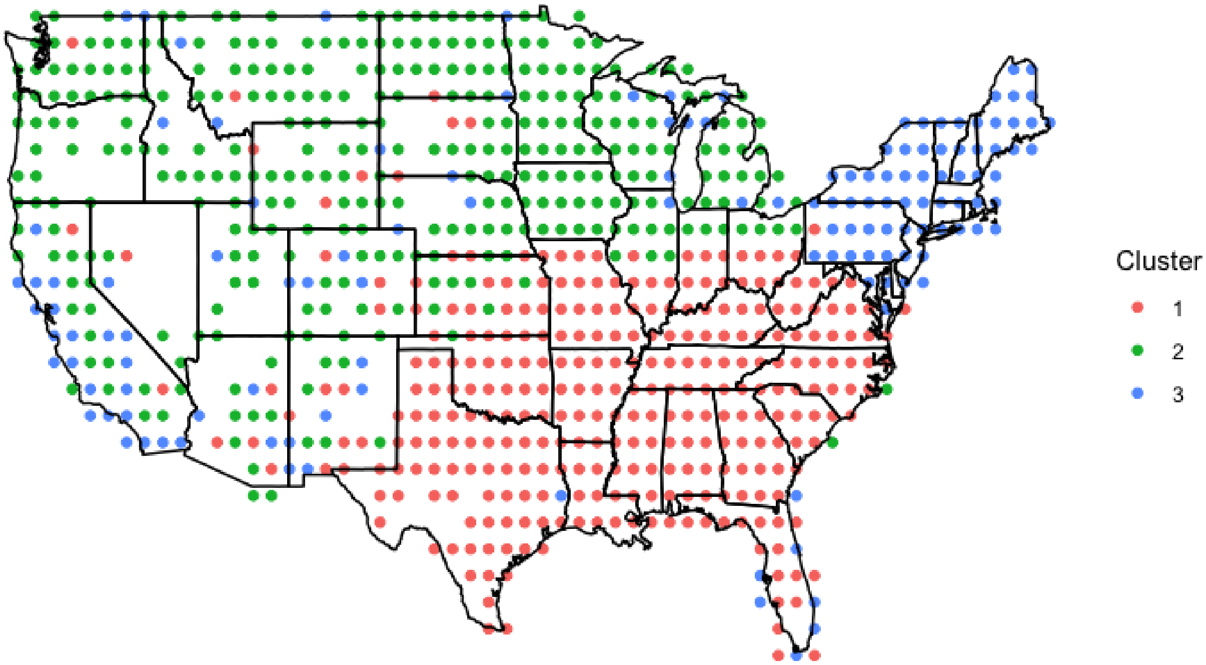

Figure 9: The result of spectral clustering on LingLocation dataset; the three geographical clusters found on the original data are still present here.

7 Academic Integrity Statement

This work is my own and all sources used are properly cited.

8 References

Vaux, Bert. “Harvard Dialect Survey.” dialect.redlog.net. Harvard University. 2003. Web.