Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

24978168

October 5, 2017

1 Introduction

Dialects, or linguistic variations, contain valuable information about the social identity of a group. Dialects may vary according to geography, social class, sex, and age, and thus have a potential to reveal the underlying social identity and shared history of groups of individuals who use similar dialects. In this report, we will explore the Dialect Survey data collected by Vaux [1], which studied the variations in the English language by surveying over 47 thousands individuals in the United States. In particular, we will use dimensionality reduction and clustering methods to gain insight into the relationship between dialect groups and geography.

2 Kernel Density Estimation and Smoothing

Before we dive into the linguistics dataset, we will first explore the effects of kernel function and bandwidth on density estimation. For this purpose, we will use the Redwood dataset from lab 1 [2].

2.1 Kernel Density Estimation

Fig 1 shows the density of the temperature distribution estimated with a Gaussian kernel. We can see that a small bandwidth, h, causes low bias and high variance, where the estimated density follows the observed values closely. As we increase the bandwidth, however, the estimated density follows the data less closely, resulting in high bias and low variance.

Fig 2 shows the effects of using different kernel functions on density estimation. As we can see, Gaussian, cosine, and triangular kernel functions gave similar results whereas the rectangular (i.e. uniform) kernel produced a slightly less smooth estimate.

2.2 LOESS

Now we turn to exploring the effects of bandwidth on LOESS. Similar to kernel density estimation, Fig 3 shows that a small bandwidth resulted in a wiggly fit that followed the data more closely whereas a higher bandwidth produced a smoother fit. The degree of polynomials had a similar effect; as we can see in Fig 4, a higher degree of polynomials produced high variance and low bias, whereas a lower degree produced lower variance but higher bias.

3 The Data

The second part of this report explores the linguistics data collected by the Harvard Dialect Survey in 2003. The survey consisted of a series of questions that explored the variations in the phonetic and lexical differences

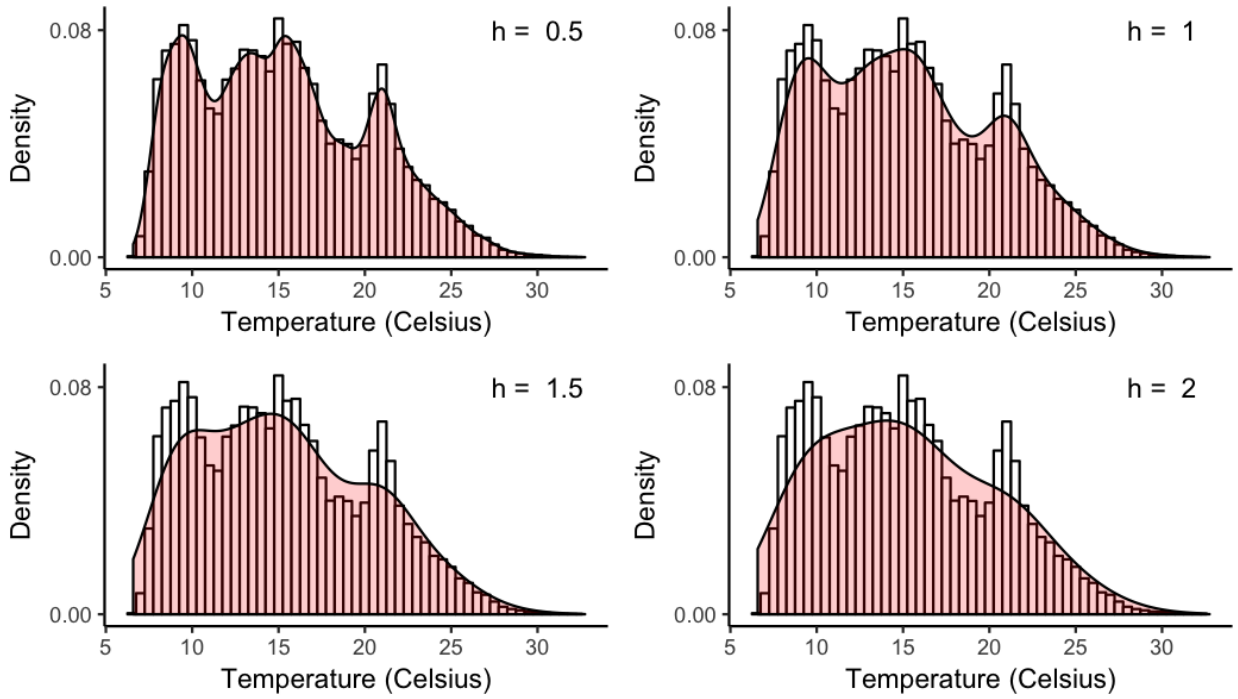

Figure 1: Effects of bandwidth, h, on kernel density estimation. Each panel shows the histogram of the data and the estimated density for a particular bandwidth, h. A Gaussian kernel was used to estimate the density.

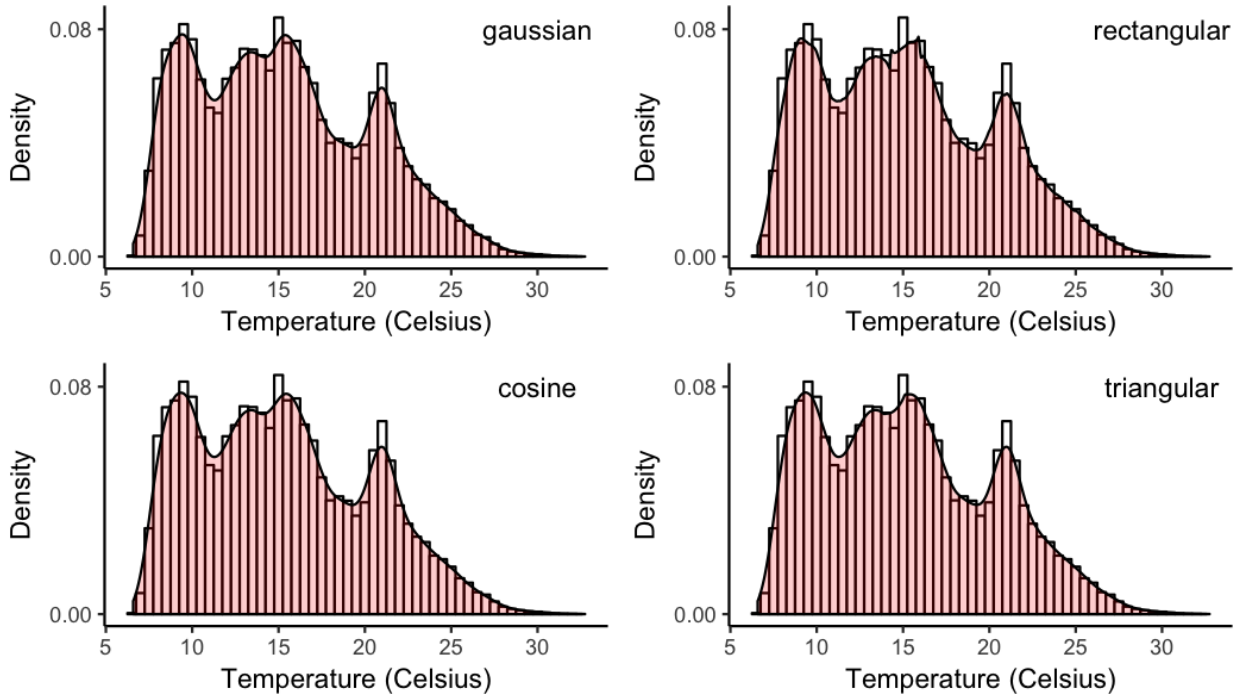

Figure 2: Effects of different kernel functions on kernel density estimation. Each panel corresponds to the estimated density for a particular kernel function. The bandwidth was set to .5 for all kernel functions.

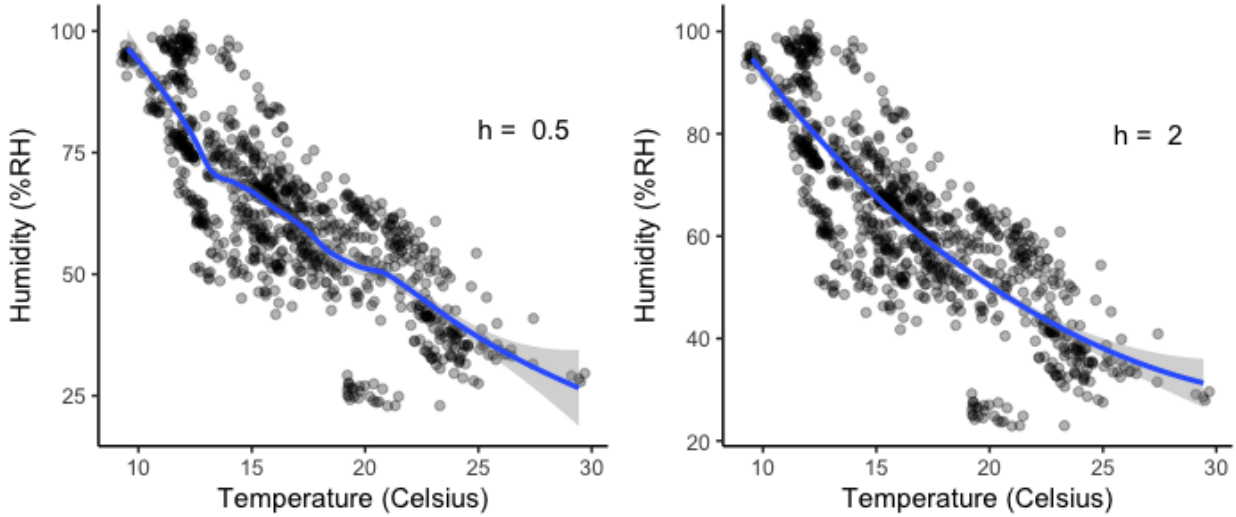

Figure 3: Effects of bandwidth on LOESS. The degree of polynomial was set to 1.

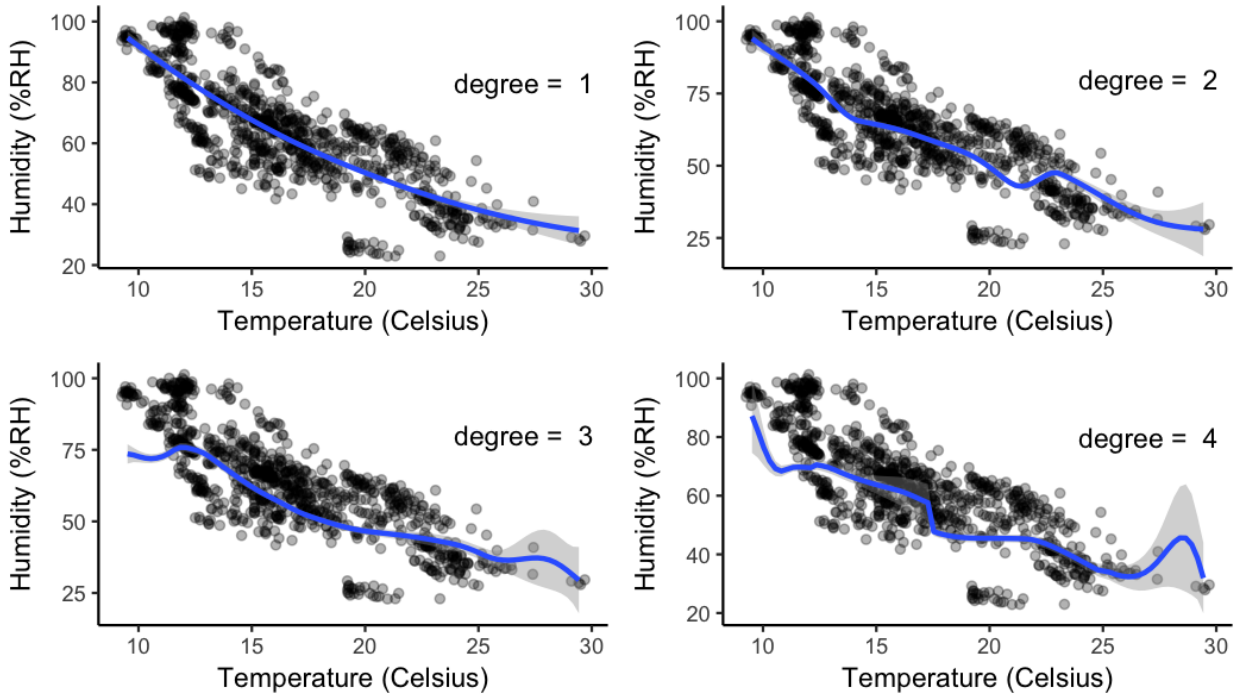

Figure 4: Effects of degree of polynomials on LOESS. The bandwidth was set to 2.

in the English language in the United States. In this report, we will focus on analyzing 67 questions that explore lexical differences (e.g. What do you call it when rain falls while the sun is shining?).

We have two different representations of the dataset: LingData and LingLocation. LingData includes the categorical responses to 67 survey questions by 47,471 respondents. It also contains information about the reported location of each respondent, given by state, city and zip code. Also included in the data are the latitude and longitude of the center of each zip code.

The second dataset, LingLocation, is a binary encoding of the categorical data described above. The observations were then binned into one degree longitude by one degree latitude squares, which partitions the map of US into 781 cells. The data for each square is encoded as the sum of binary response vectors for individuals who belong to the cell.

3.1 Data Quality and Cleaning

For LingData, I removed 1020 observations for which either longitude or latitude information was missing. Furthermore, 107 observations from Hawaii and 98 individuals from Alaska were excluded to restrict the analyses to the 48 contiguous states. Furthermore, there were individuals who had missing responses to some of the survey questions. Since we had a large number of observations to work with, I chose to apply a stringent filter that removed any observations for which there was missing data for any of the questions asked. This left us with 39,051 observations for downstream analyses.

Similarly for LingLocation, cells that corresponded to Hawaii or Alaska were removed (i.e. longitude $<$ -150 or latitude $>$ 50). This reduced the original count of 781 cells to 527. Furthermore, there were instances where the number of individuals 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, I chose not to remove such observations. Therefore, I chose to use LingData for the main analyses and use LingLocation primarily to validate those results (see Stability of Findings for further discussion).

3.2 Exploratory Data Analysis

Here we explore a few survey questions and their potential relationship to geography. Fig 5A shows the distribution of the responses to the question: What do call the rubber-soled shoes worn in gym class? The plot shows the distribution of the two most common answers, which were "sneakers" and "tennis shoes". As we can see from the figure, those who answered "sneakers" are concentrated in the Northeast whereas the word "tennis shoes" is more commonly used throughout the rest of the country.

Fig 5B shows the distribution of the most common answers to the question: What do you call the thing from which you might drink water in a school? Here we see a loose partition for those who chose "drinking fountain" versus "water fountain." 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 and were primarily from Wisconsin, Massachusetts, and Rhode Island.

Experimenting with linked brushing (see "/extra/linked_brushing_2questions.html") showed that the response to question 73 does not help predict the response to question 103. Linked brushing with three questions (see "/extra/linked_brushing_3questions.html") showed that we gain a bit more information. For example, if an individual answered "drinking fountain" for question 73 and "tennis shoes" for question 103, then they are not likely to call the night before Halloween as "mischief night." This suggests that the responses to some groups of questions may be correlated with each other.

Now we turn to reducing the dimension of the data. Fig 6 shows the projection of the binary representation of LingData onto the first two principal components. Here, we see that the observations form loose clusters according to their longitudinal location. This suggests that we may able to relate the linguistics data to geography, which will be discussed further in Dimension Reduction Methods. Furthermore, Fig 7 shows that the first ten principal components explain only a modest proportion of the total variability, suggesting that the underlying cause for linguistic variations is not simplistic and its complexity cannot be captured by a few features.

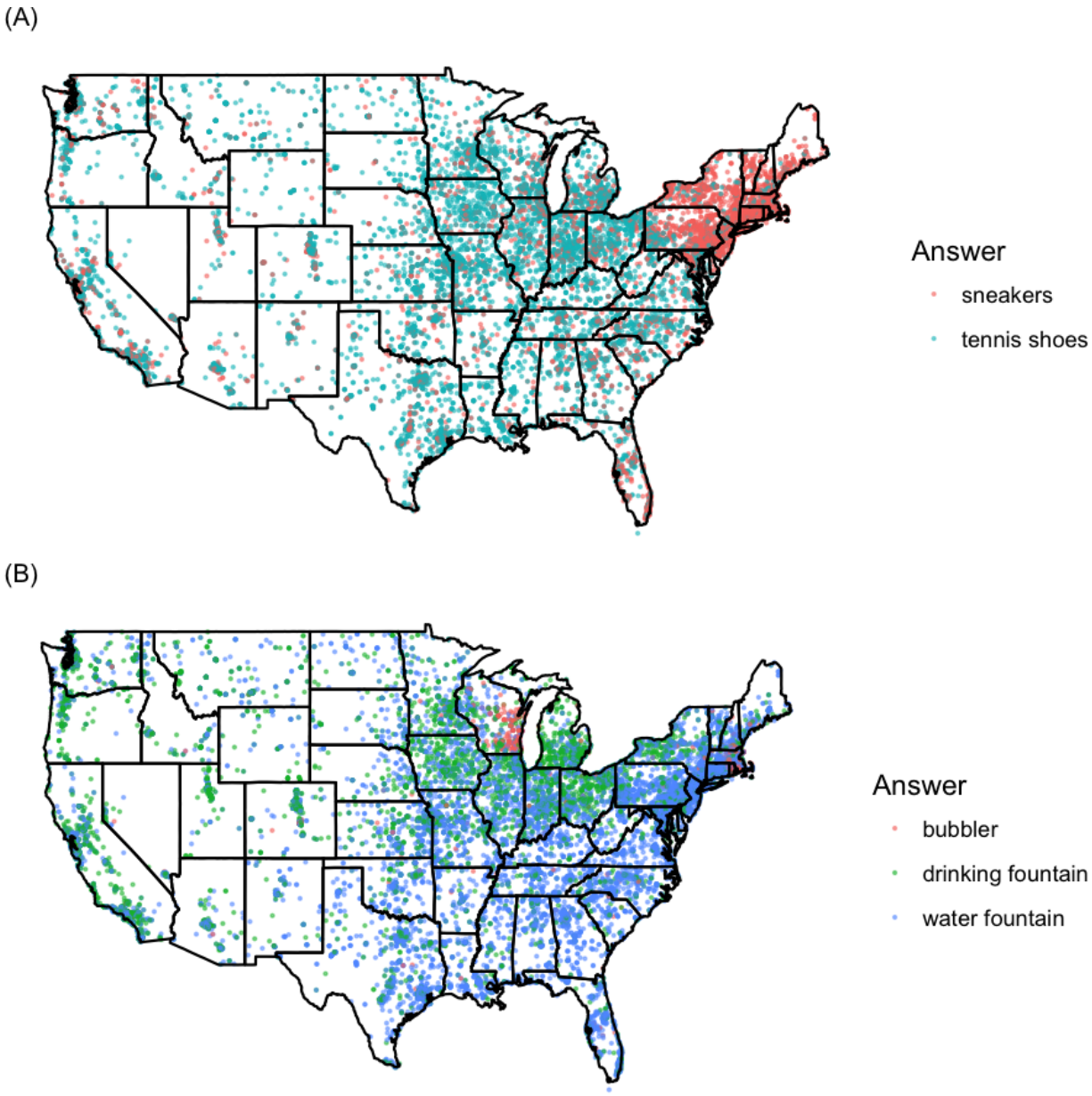

Figure 5: Distribution of responses in the lower 48 states. (A) What is your general term for rubber-soled shoes worn in gym class? (B) What do you call the thing from which you might drink water in a school?

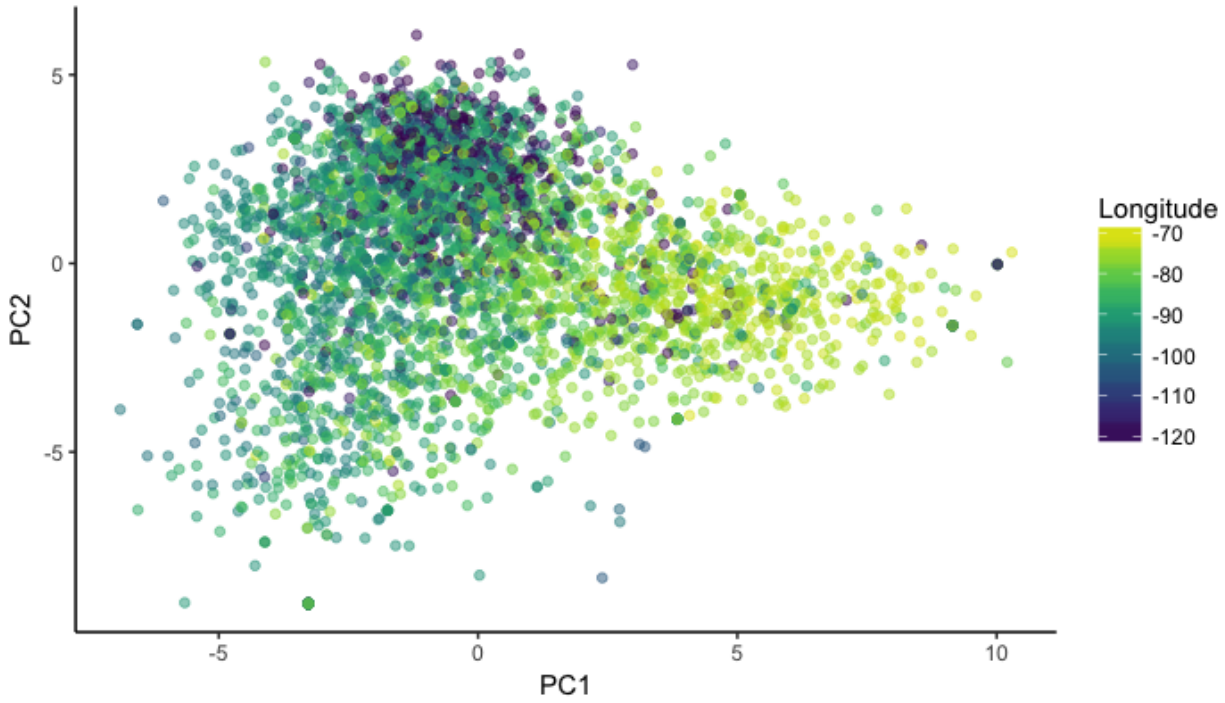

Figure 6: Projection of LingData onto the first two principal components.

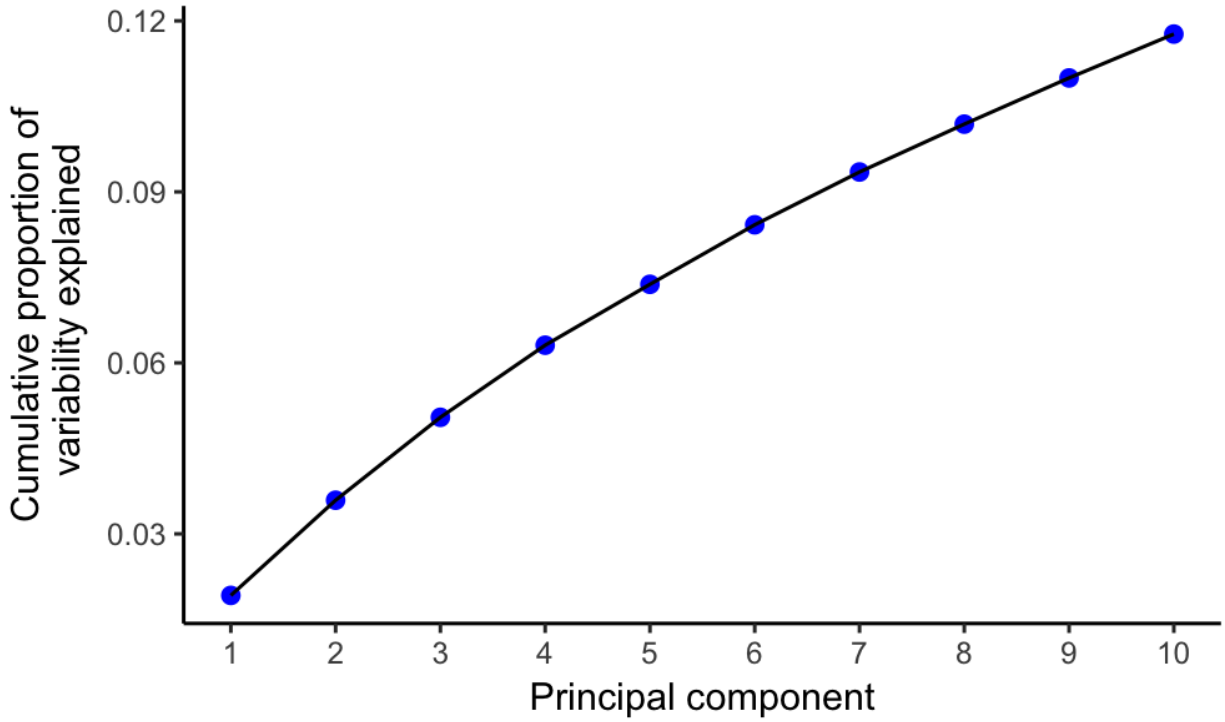

Figure 7: Amount of variability explained by the first 10 principal components.

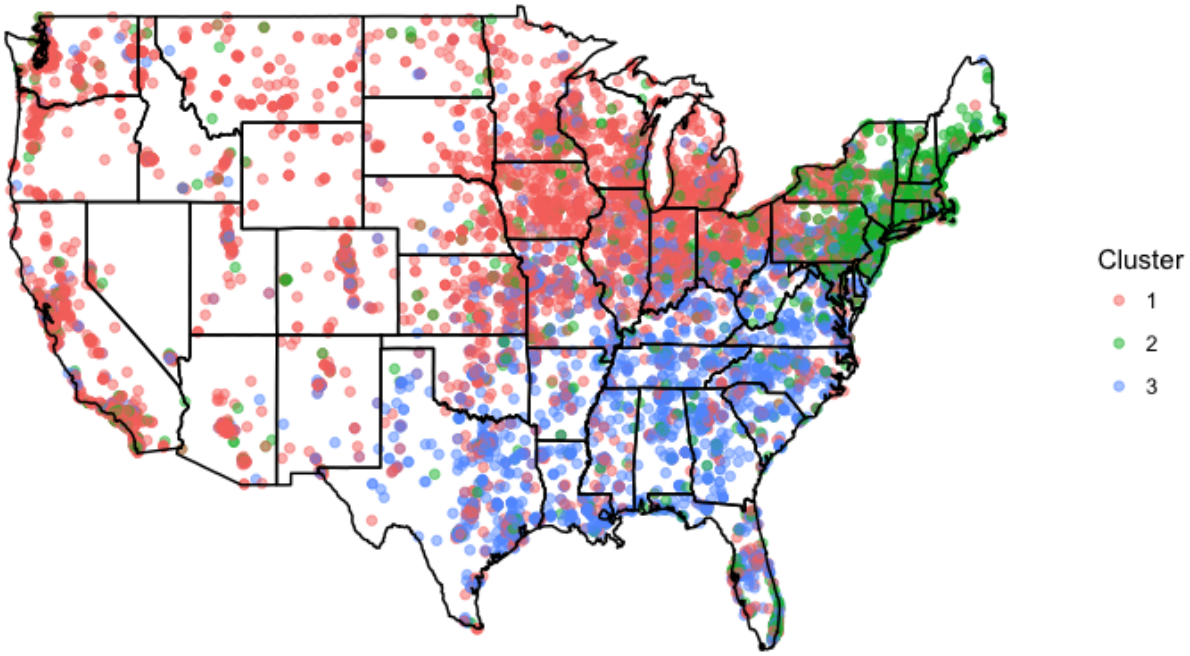

Figure 8: Distribution of the individuals in the US, color-coded by their k-means cluster assignment. k-means clustering with $k = 3$ was used on the first two principal components. The clusters correspond to three geographical groupings. Only 50 percent of the data are plotted for visualization.

4 Dimension Reduction Methods

In order to gain insight into the potential relationship between dialect groups and geography, I performed k-means clustering on the first two principal components computed in the previous section. The number of cluster, k, was set to 3, which gave the highest average silhouette value. Fig 8 shows the distribution of the cluster members on the map of the US. Here we can see three general geographical clusters: 1) the Northeast, 2) the Southeast, and 3) a super region including the Midwest, the Southwest, and the West. The geographical clusters are not clean-cut, however. There is a fair amount of overlap between clusters on border regions; we also see many outliers, such as the members of the "Northeast cluster" residing in Florida.

To explore some of the survey questions that separate the groups, I looked at questions 73 and 50, which had the highest loadings for the first and second principal components, respectively. As we can see in Fig 9A, question 73 separates the Northeast from other regions. More specifically, those who responded "sneakers" are concentrated in the Northeast whereas those who responded "tennis shoes" are spread across the rest of the country. Question 50, on the other hand, separates the northern regions from the southern regions. The term "y'all" is predominantly used in the Southeast and Texas, while the other terms such as "you" and "you guys" are more commonly used in other regions of the country. Interestingly, we can see that the separations defined by these two questions are also present in the geographical clusters we found before in Fig 8, which agrees with our expectation that these two questions contribute to the separation of the groups.

As mentioned before in Fig 7, 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.

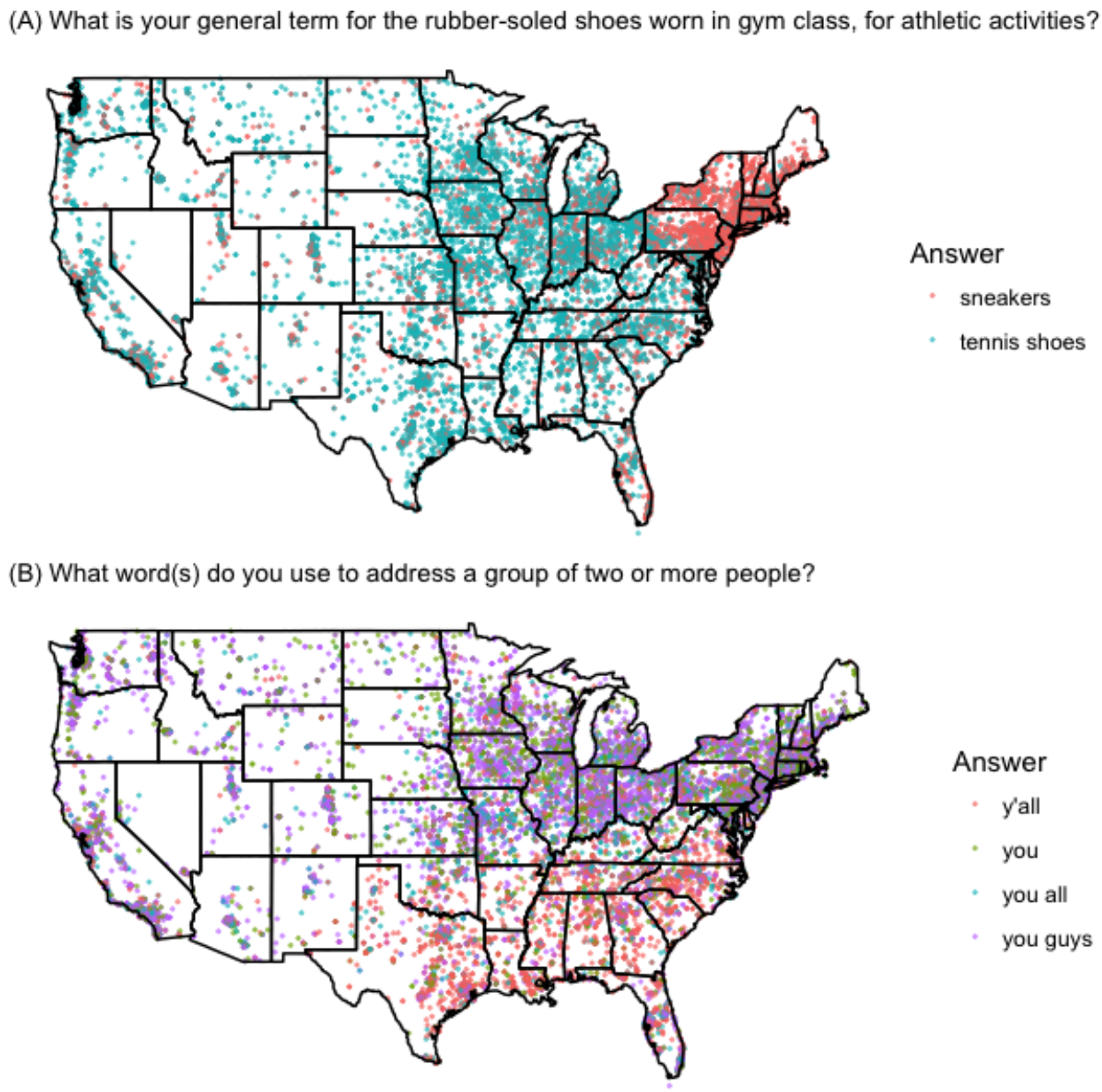

Figure 9: Distribution of the responses to the questions with the highest PCA loadings. (A) Responses to question 73 (highest loading for PC1); (B) Responses to question 50 (highest loading for PC2). Only the answers with at least 10 percent response rate are shown.

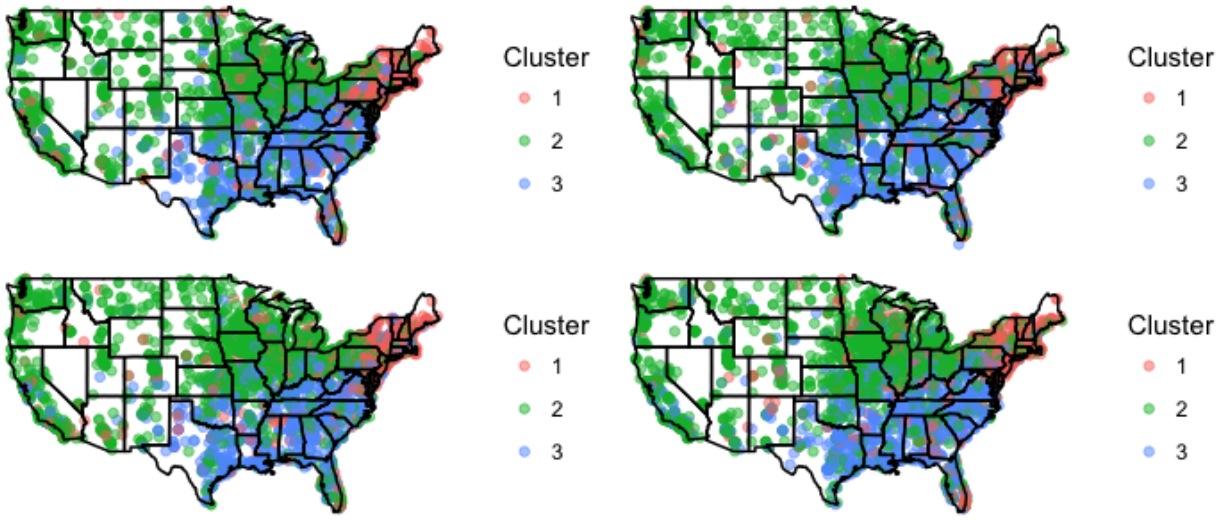

Figure 10: The results of four runs of k-means with different starting points. All four runs show similar geographical clusters we found before. The mapping from color to cluster is arbitrary.

5 Stability of Findings to Perturbation

Here we check whether the three geographical clusters we found in the previous section are stable under various perturbations to the data. First, I ran the k-means clustering algorithm four times with different starting points. Fig 10 shows that each run of k-means produced very similar clusters and our finding, therefore, is stable under different starting points.

Next, I downsampled the data by half and re-ran PCA and k-means on the first two principal components. As we can see in Fig 11A, the three geographical clusters are still present on the downsampled data. To investigate the effects of removing questions from the dataset, I excluded from analyses questions 73 and 50, which had the highest loadings for the first two principal components, respectively. Fig 11B shows that we still find similar geographical clusters, implying that there must be other survey questions in the perturbed dataset that separate the groups in a similar way.

Finally, I checked whether re-running the analysis on a different encoding of the data would produce similar results. In order to see this, I ran PCA and k-means on LingLocation, which bins the observations in one degree latitude by one degree longitude squares. Fig 12 shows that the three geographical clusters we find are similar to those shown in Fig 8. Thus our finding that linguistic variations relate to three distinct geographical groups is stable under perturbations to the data.

6 Conclusion

In this report, we explored the linguistic variations in the English language by analyzing the data collected by the Harvard Dialect Survey. In particular, dimensionality reduction by PCA showed that the the total variability in the data cannot be explained by a few principal components, implying the complex nature of underlying cause for the linguistic variations. Furthermore, we saw that the individuals form groups that correspond to three fairly distinct geographical regions: 1) the Northeast, 2) the Southeast, and 3) a combined region of the West, the Southwest, and the Midwest. Finally, our finding was stable under several types of perturbations to the data, such as downsampling, removal of several questions, using different starting points for k-means clustering, and using a different encoding of the data, giving us confidence to the robustness of our finding.

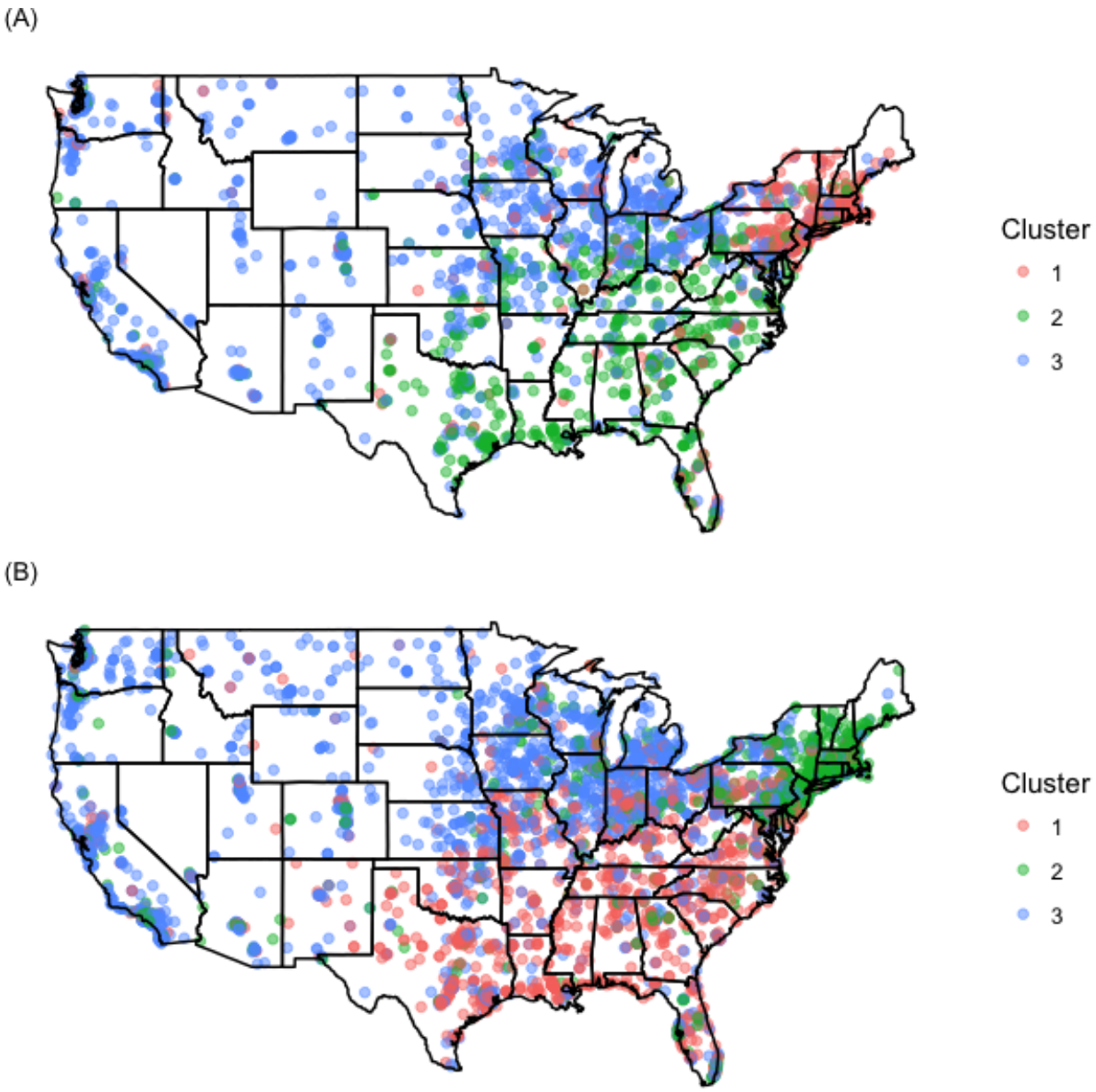

Figure 11: The result of spectral clustering (A) after removing 50 percent of the observations; (B) after removing questions 50 and 73, which had the highest PCA loading for the first and second principal components, respectively. The three geographical clusters found on the original data are still present here.

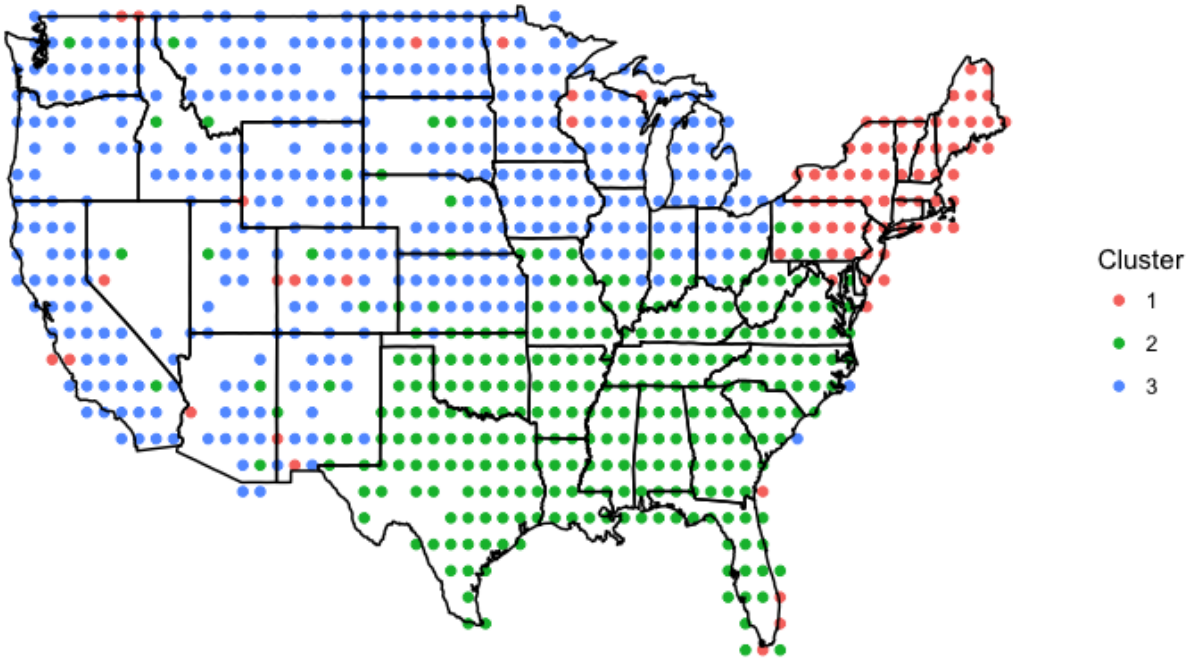

Figure 12: The result of spectral clustering on "LingLocation." The three geographical clusters found on the original data are still present here.

References

[1] Vaux, Bert. "Harvard Dialect Survey." dialect.redlog.net. Harvard University. 2003. Web.

[2] Tolle, Gilman, et al. "A macroscope in the redwoods." Proceedings of the 3rd international conference on Embedded networked sensor systems. ACM, 2005.