Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

SID: 3033013609

October 5, 2017

1 Introduction

This report is divided into two main sections. The first section dicusses kernel density estimation and locally weighted scatter plot smoothing (LOESS) for data collected over forty-four days on two redwood trees in California. The study looked at measurements of temperature, humidity, and solar radiation, at different heights on the trees. Here I present analysis of the effects of bandwidth choice, as well as kernel type for kernel density estimation and polynomial degree for LOESS. The second section focuses on data collected from a survey where individuals answered questions concerning pronounciation and word choice. Participants across the United States participated in the survey. In this section I investigate the connection between answers and geography as well as how answers relate to each other using principal component analysis (PCA) and spectral clustering. Finally, I look at the stability of these algorithms by perturbing the data to determine if my findings should be trusted.

2 Kernel Estimation and LOESS for Redwood Data

In this section I discuss kernel density estimation and LOESS with redwood data.

2.1 Kernel Densisty Estimation

To look at how bandwidth and kernel choice effect the estimate of the density I look at the values for temperature measurements. It should be noted that this data was extensively cleaned, which is described in detail in a previous report. In this context we are considering the kernel density estimate as a way to visualize trends in the dsitribution of temperature, not as a functional estimate of the desnity itself. The estimate is found by computing

$$\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)$$

for data x and kernel K_h with tuning parameter h. Figure 1 plots the kernel density estimates using gaussian and triangular kernels for varying bandwidths overlayed on the histogram to illustrate how closely the estimate follows the histogram. From these plots we can see that for smaller bandwidths the triangular and gaussian kernels produce similar results. However, as we increase bandwidth they start to move apart from one another slightly, as can be seen when using a bandwidth of five. Additionally, it can be seen that as the bandwidth increases the density estimates get smoother. If too small of a bandwidth is used we risk overfitting, but too large creates the potential to lose information. This data appears to have three distinct modes, so it would be good to choose a bandwidth that reflects that trend. When the bandwidth is increased to 1.5 we start to see only two modes, so a bandwidth of 1 seems to be an appropriate choice here because it captures the peaks from the hotter days as well as night time temperatures and those from the more normal days.

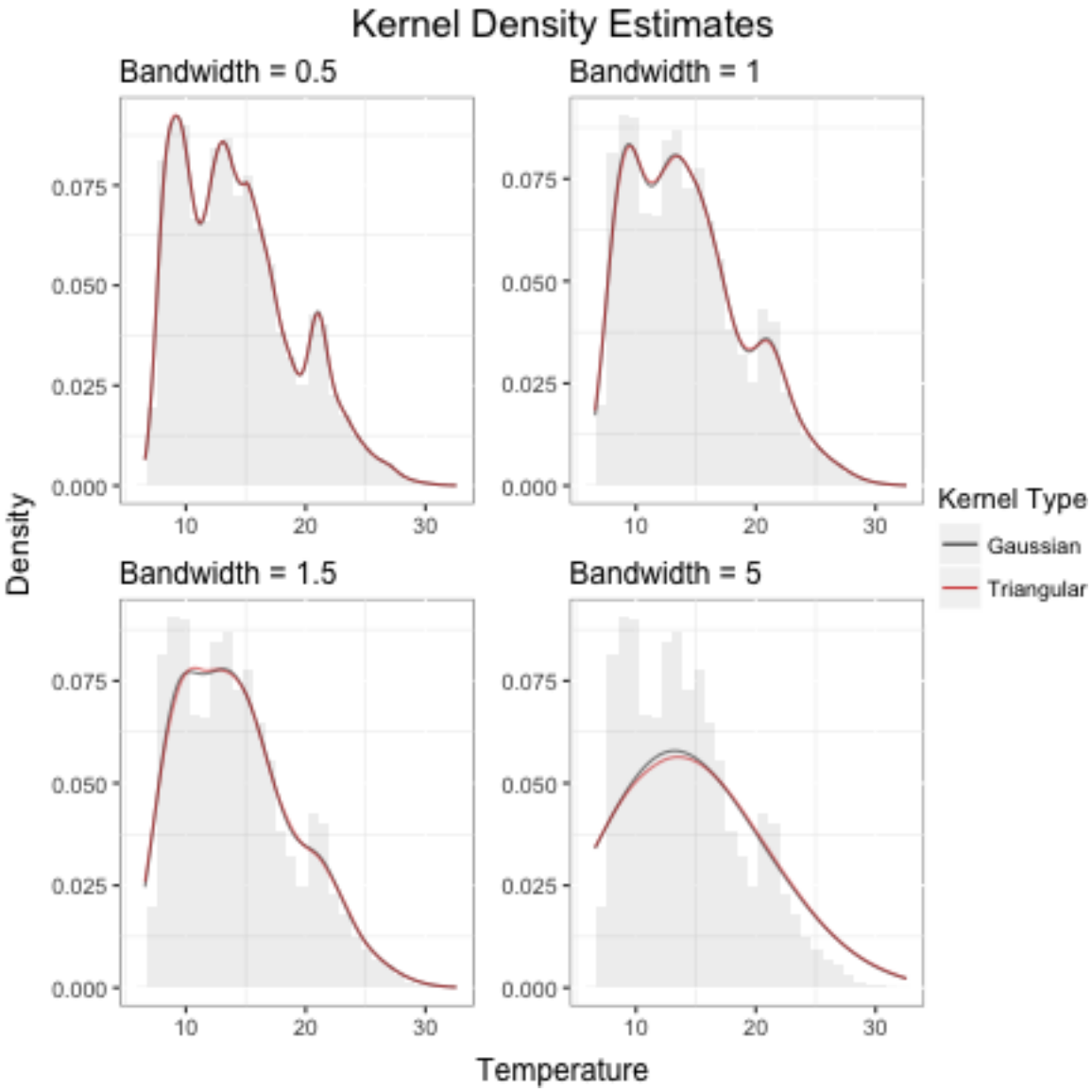

Figure 1: Plot showing the effects of bandwidth and kernel choice for kernel density estimation.

2.2 LOESS

Next we look at the trends in the relationship between temperature and humidity using LOESS. LOESS is a non-parametric way of visualizing non linear trend-lines, which fits local polynomials to small subsets of the data and aggregates that into a smooth curve. When performing LOESS there are two basic choices, first, which polynomial is chosen to fit locally, and second, the bandwidth. Smaller bandwidths lead to smaller windows for each local polynomial that is fit to a subset of the data. I choose to look at the relationship between temperature and humidity of the redwood trees at two o'clock in the afternoon. From Figure 2 we see that as bandwidth increases the fit lines become smoother and less sensitive to noise in the data, while the smaller bandwidths lead to overfitting. Additionally in fitting the first and second degree polynomials we see that the second degree polynomials cause a lot more wiggliness in the fit than the first. If we choose to small of a bandwidth for the second degree polynomial, the areas between the clusters in the temperatures cause sharp dips or peaks in the curve, which is not ideal. Inspecting the plots shows that for a linear local polynomial a bandwidth of 0.75 captures the trend without overfitting, while larger bandwidths seems lead to overly smoothed curves. Using looks second degree polynomial it looks as if a bandwidth of 1.5 works best, however overall the linear fitting works better in this instance.

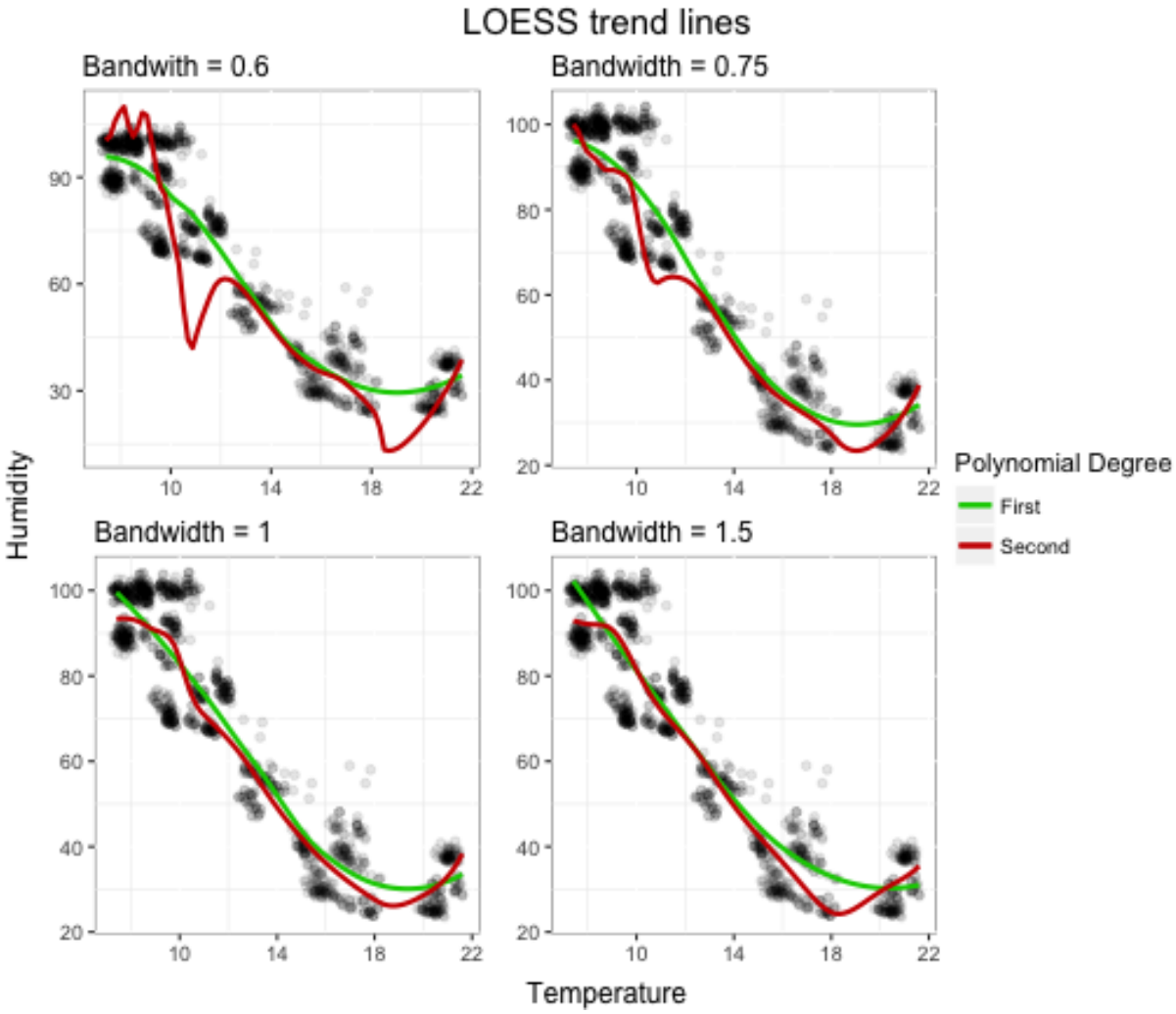

Figure 2: Plot showing the effects of bandwidth and polynomial choice for local polynomial smoothing.

3 Linguistic Data

Now that I have looked into kernel density estimates and LOESS I move to discussion of the linguistic data, including data cleaning, exploratorary data analysis, dimension reduction, clustering, and stability.

3.1 The Data

The data in this section is composed of responses to a questionnaire where 47471 individuals responded to questions concerning how they pronounce words or the different terms people use for the same entity, and we focus on the latter of which there are 73 questions asked. As an example, one question on the survey asked "What do you call the long narrow place in the middle of a divided highway?" The potential responses to choose from were: median strip, median, boulevard, mall, traffic island, neutral ground, island, pork strip, I have no word for this, and other. The questions were designed as to not lead the responder into choosing a specific answer. The data was split into multiple datasets for analysis. One dataset contained the responses for each individual with one column for each question, as well as columns with longitude and latitude, city, state, and zipcode. The second primary dataset contained one column for each potential answer with 781 rows corresponding to the data pooled into squares on one degree latitude by one degree longitude.

3.1.1 Data quality and cleaning

When doing an initial investigation of the data I noticed that there were 1023 cases with missing data in the dataset with one row for each individual. The missing data, was corresponding to missing latitude and longitude values. To address this issue, I used the geocode() function in the ggmap package to look up the longitude and latitude for the city, state, and zip code at that observation. I saved what was returned from geocode() to a csv file to enable quick access, as the geocode() function is time intensive. After completing this process I found that there are 12 missing latitude and longitude values so we look into those ID's in the original data.

Looking into those 12 observations I noticed that the corresponding towns or zip codes were not inputted correctly. However, there were of four these observations that I could manually find the latitude and longitude online and I added those values to the data. The next step in the cleaning process was to visualize the data on the map of the United States and remove the observations that were outside the United States. Some responders entered zip codes and towns in Canada or other countries that I choose to remove.

The final step in our cleaning process was to create a data set where we code the answers into binary variables. This was done to enable the use of dimension reduction techniques and to help in understanding how answers relate to one another.

3.1.2 Exploratory Data Analysis

In order to further explore the intricacies of the data I looked at two questions to determine how they relate to each other and how they relate to geography. The questions I choose to look at were (1) "What word(s) do you use to address a group of two or more people?" and (2) "What is your generic term for a sweetened carbonated beverage?" I choose these questions, because when plotting the distributions of the answers of the map of the US they roughly defined geographic regions that from prior knowledge I would assume have different dialects. Many of the other questions exhibited similar patterns when looking at the plots, however these exhibited the behavior more clearly. However, it should be noted there were a number of questions that seemed to give relatively uniform answers across the United States. The questionnaire had ten possible answers for the term for carbonated beverage and nine for the plural term for second person. Looking at the distributions of the answers for carbonated beverages, "soda," "pop," "coke," or "soft drink" made up over 96% of the responses, so I choose to only display those on the map. Additionally for plural second person reference "you all," "you guys," "you," and "y'all" accounted for approximately 94% of the responses.

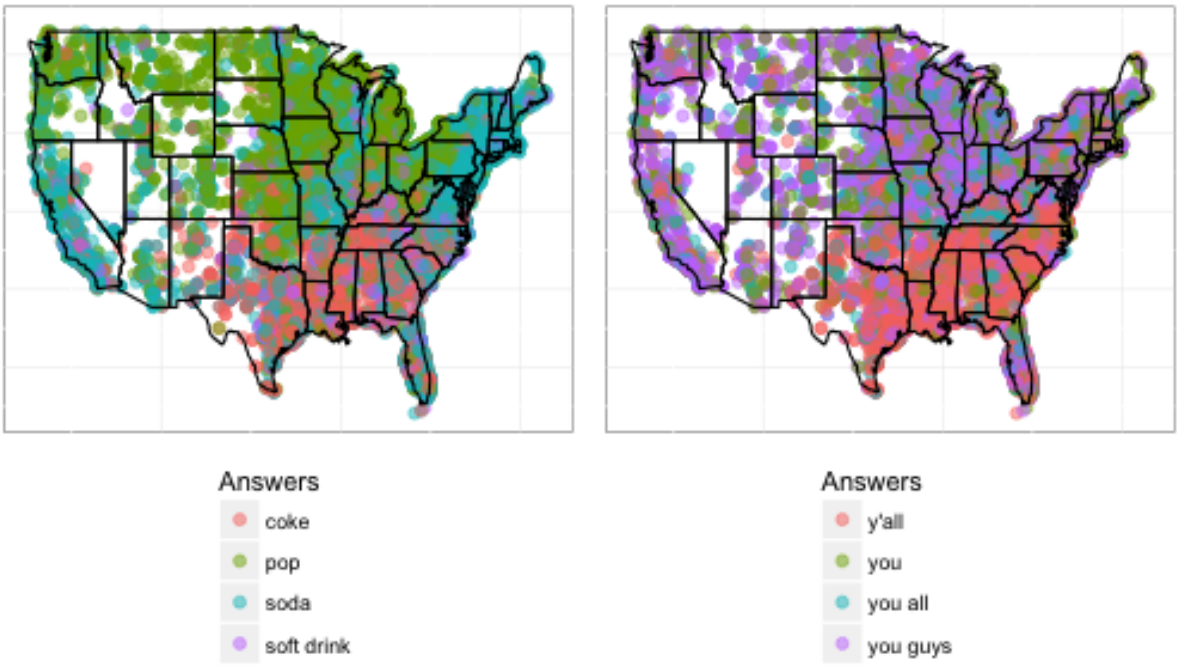

Figure 3: Projection of answers to (1) What is your generic term for a sweetened carbonated beverage? and (2) What word(s) do you use to address a group of two or more people? onto a map of the United States to visiualize the geographical variation in dialect.

Looking at Figure 3 we can see that both questions have answers that are confined relatively to one geographic area. For the carbonated beverage question the large majority of people who say "pop" are in the Midwest and Northwest, those who say "coke" and "soft drink" are primarily in the South, while "soda" seems to be spread throughout the country with higher concentrations on the West coast and the Northeast. The responses to how an individual refers to a group of people define fewer geographic regions as "you," "you all," and "you guys" are scattered around most of the country. However, there is an intense concentration of using the term "y'all" in the South. There are patterns in both of these maps that agree with my intuitive knowledge based on my interactions with people throughout the United States.

Now that we have considered how the individual questions and answers are distributed throughout the United States, we attempt to determine what relationships exists between the questions themselves. From examining the plots in Figure 3, we can see that there seems to be a connection between answering question 1 with "coke" and question 2 with "y'all," but other connections are not immediately apparent. This led me to consider the correlations between the answers to the questions. I found that the strongest correlation, as expected, is between "coke" and "y'all" with a value of 0.42. So, if we knew a person called a carbonated beverage "coke" we could predict (with more certainty than not having that information) that they would have use the term "y'all," but this would not be correct a large amount of the time as there is some intermingling with other responses. We can also see that "y'all" is negatively correlated with "soda" (-0.13) and "pop" (-0.19). This appears to be because "soda" and "pop" are concentrated in the Northeast or West Coast and Midwest respectively, where the majority of people do not use "y'all". Additionally "pop" and "you guys" are correlated (0.17), which also matches with their appearance on the map, with the response "you guys" covering the large majority of the Midwest were "pop" is primarily used. A final interesting correlation, albeit small (0.01), is between "fizzy drink" and "youse" both of which are terms I would typically associate with British or Australian dialects. Neither of those responses were common, but there seems to be some connection between people who use those terms.

3.2 Dimension reduction methods

After cleaning and exploring the data the next step was to employ dimension reduction and clustering techniques to determine if my speculated findings concerning the two questions in the previous section could be seen throughout the data.

3.2.1 Principal component analysis

To start looking at dimension reduction I used the entire binary dataset and performed PCA on the columns that corresponded to the answers to the questions. However, when looking at the resulting screeplot from that analysis I found that I would need to use the first 126 principal components to explain 50% of the variability of the data.

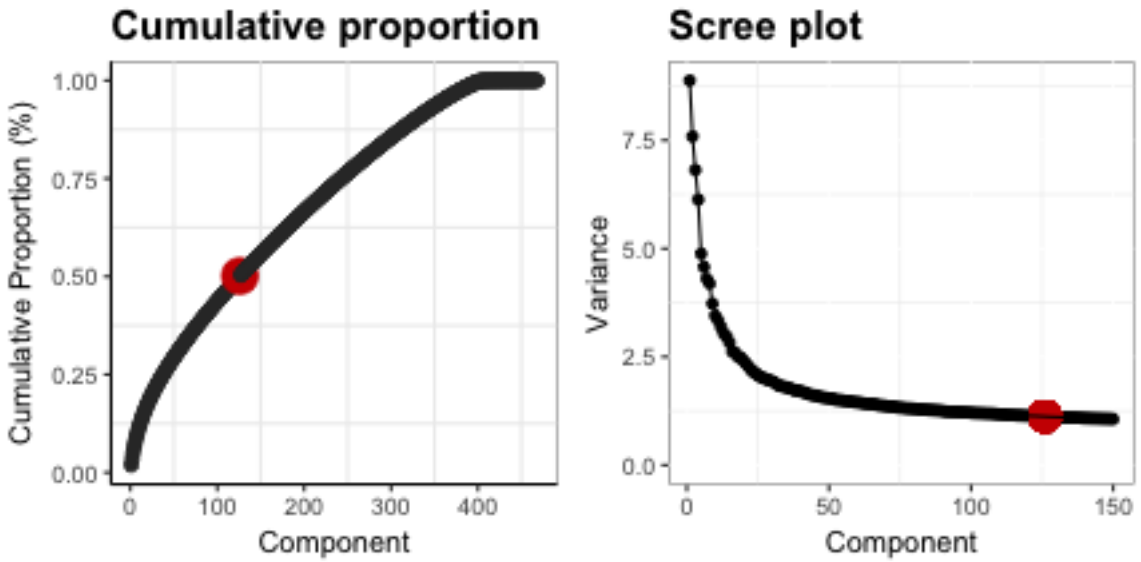

Figure 4: The cumulative proportion of variability explained by the principle components after doing PCA on the binary dataset is on the right. On the left is the scree plot to determine the number of principal components used. The red dot marks the 126 component where approximately half of the variability is explained.

This realization led me to consider the location dataset, which binned the binary answers into groupings that were 1 degree latitude by 1 degree longitude. By grouping the data in this way I do not reduce the dimensions of the covariates, there are still $p = 468$ columns corresponding to each possible answer, but I do reduce the number of rows drastically from $n = 47448$ to $n = 758$. The purpose of taking this step was to potentially find a smaller number of principle components that more effectively summarize the data. In order to perform PCA on the data I first normalized the rows by the number of people in each cell to prevent the cells with large numbers of people from effecting the data. I found that simply scaling prior to doing PCA was did not address this issue. Performing PCA on this new data set and investigating the scree plot below I found determined that after 22 principle components around 50% of the variability is explained and the percent explained by each additional component does not add enough to warrant being included in the analysis. Using the location dataset we can notice that the first 5 principle components appear to explain 25% of the variability, which is a big contrast to using the entire dataset where only 7% was explained.

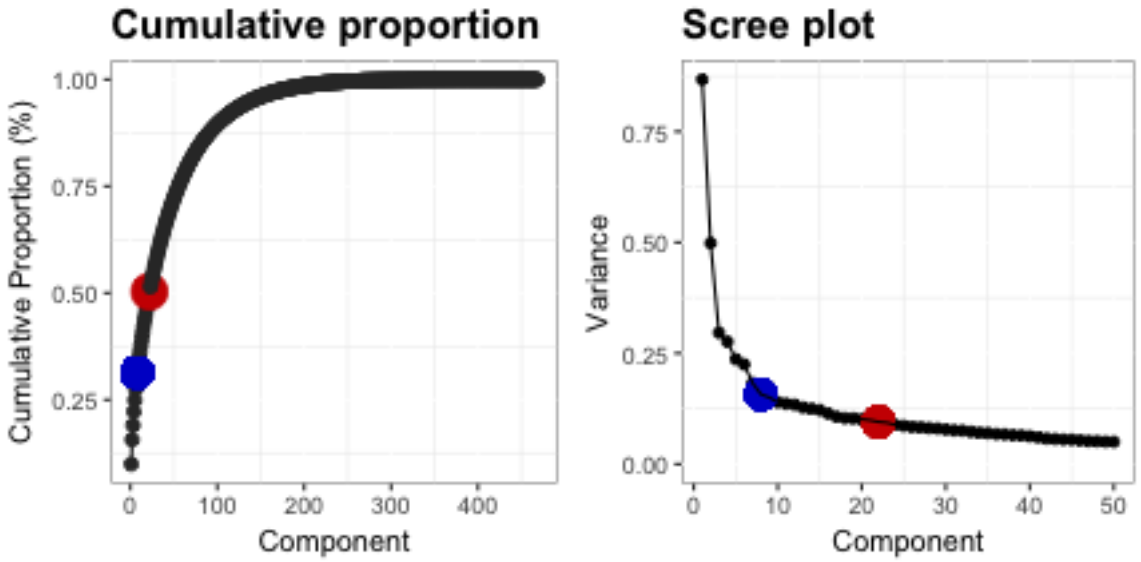

Figure 5: The cumulative proportion of variability explained by the principle components after doing PCA on the binary dataset is on the right. On the left is the scree plot to determine the number of principal components used. The red dot marks the 126 component where approximately half of the variability is explained, and the blue dot is eighth component where the scree plot begins to flatten.

Ultimately after looking at both versions of the screeplot, I determined that using eight clusters was appropriate. Although there is only 0.314 amount of variability explained, the second scree plot appears to have an elbow at eight principle components, where each additional component would not add enough more information to warrant including.

After deciding the number of components to use in my analysis, I look at in my analysis I investigated the loadings from PCA to determine which questions/answers contributed more heavily to the principle components than others.

By looking into the first two principal components I found that answering "y'all" when a person was asked how they refer to a group of two or more people contributed most to the 1st component. Additionally, responding "catty-corner" when asked what a person refers to something diagonally across from them, and calling the object that kids drink from at school a "water fountain" were the next two highest contributors. For the second component the biggest contributor answering that is was acceptable to not wear pantyhose, second was answering "sneakers" for the general term for athletic shoes, and third was responding "soda" when asked what a person refers to a carbonated beverage as. Looking at the projection of these three question onto the map it appears that the question/answer pairs that are most important to the largest principle components are the answers that are most confined to a geographic area and that cover a lot of people. For example, using "y'all" was heavily concentrated in the South as seen in Figure 3, which is a defined area with good amount of people. Additionally, referring to carbonated beverages as "soda" was concentrated both in the Northeast and the West and covered a large proportion of people surveyed. Overall, this gives a good indication that PCA is correctly picking out aspects of the data that give important information.

3.2.2 Spectral clustering (using k-means)

After computing the principal components of the data, I explored clustering the data using k-means on the first eight components of the location datasets. Looking at Figure 6 we see the silhouette plots corresponding to three, four, five, or six clusters using the first eight principal components. I choose to consider these number of clusters from investigating a plot of the within group sum of squares versus the number of clusters, as well as from the intuition that producing more than six groupings would separate the country into more distinct dialect regions.

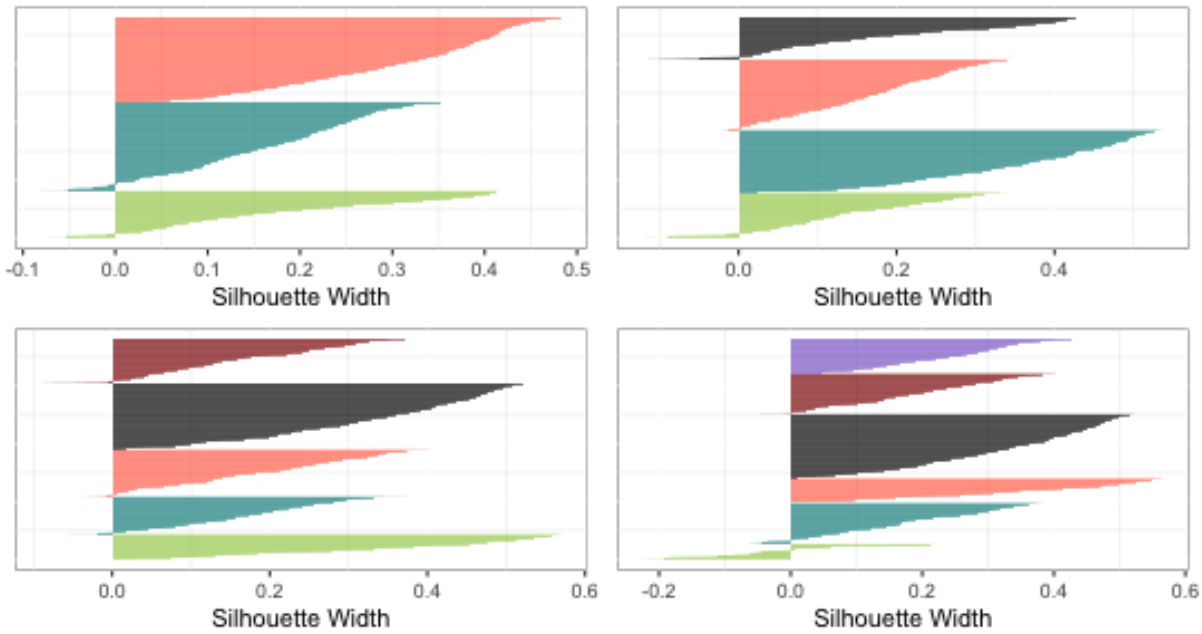

Figure 6: Four silhouette plots corresponding to the using 3, 4, 5, or 6 clusters doing k-means on the first eight principal components on the location dataset.

We get that the average silhouette width for each clustering is 0.219, 0.217, 0.239, and 0.236 for three, four, five, and six numbers of clusters respectively. The average widths are all very close to each other, and recalculating the values many times can result in any number of clusters giving the highest value. From investigating the plots for many runs it appeared that often the silhouette plot using six clusters produced one cluster that was very small and had largely negative silhouette widths, so I choose to look further into three, four, or five clusters. Additionally, it should be noted I also attempted to use the elbow method to address this issues, but there was similar ambiguity about how many clusters was best.

After projecting many different clusterings onto the map of the United States I decided to present both three and five clusters, because they both give interesting insight into the data and appear to make sense with the geographical groupings we know to be present in the United States. The majority of the time the three clusters correspond roughly to the South, Northeast, and the remainder of the US, while the five clusters incorporate the clusters that roughly outline the Midwest, Northwest, and West Coast in the addition to the South and Northeast. However, it should be noted that multiple runs can produce fairly different clusterings for both three and five clusters. There were instances where one of the three clusters covered the entire East Coast and split the West into two groups, and with five clusters there were instances where the East Coast was merged into one cluster as well. Additionally, other runs with three clusters where the southern part of the West coast, primarily California were grouped with the Northeast or the grouping presented in Figure 7 where the entire northern part of the country had its own cluster and the other two corresponded to the South and Midwest. All these sets of these other groupings have some logical interpretation geographically, and result from commonalities in dialect shared by people along the entire East Coast, or by people in the Northeast with those in California, despite there being certain question/answer pairs, such has calling a carbonated beverage "coke" vs. "soda" that split the East Coast into the South and the Northeast. Overall, it appears that using five clusters is more stable than three, but these issues give rise to the indication that the clustering methods are not very stable, which we will investigate in the next section.

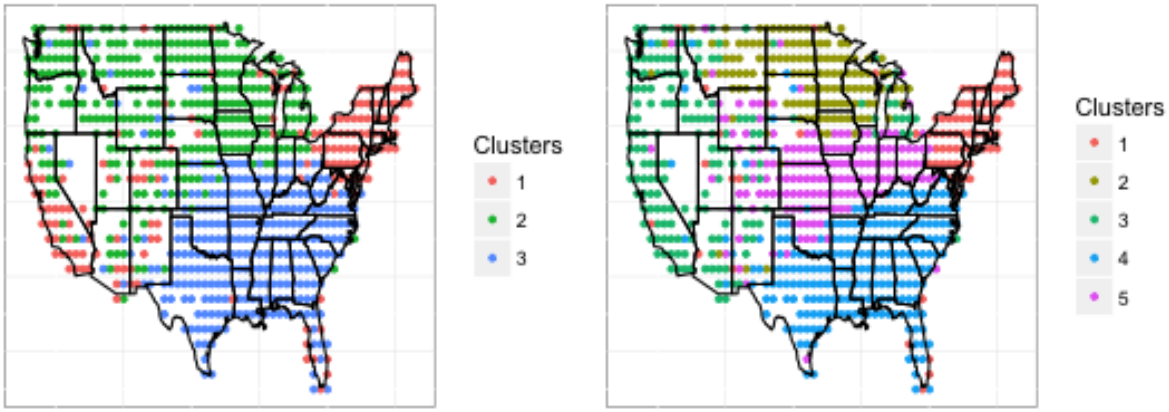

Figure 7: Plot of the survey locations labeled with spectral clustering using the location dataset with three and six clusters respectively

In addition, to considering the location dataset I also looked at how using k-means on the binary dataset, using the first nine principle components, which I determined was appropriate from looking at scree plots. Here we see in Figure 8 that performing k-means with three clusters results in a mapping that corresponds to what we have seen thus far, just with slightly more intermingling of clusters geographically. However, using five clusters results in a clustering that doesn't seem to capture any logical grouping of the data, and typically we only see three primary groups that look similar to the results from using three clusters. Additionally, similar to using the location dataset different runs of the k-means algorithm produce different clusterings for both three and five clusters. This is a good validation that it is looking into the location dataset was the appropriate choice.

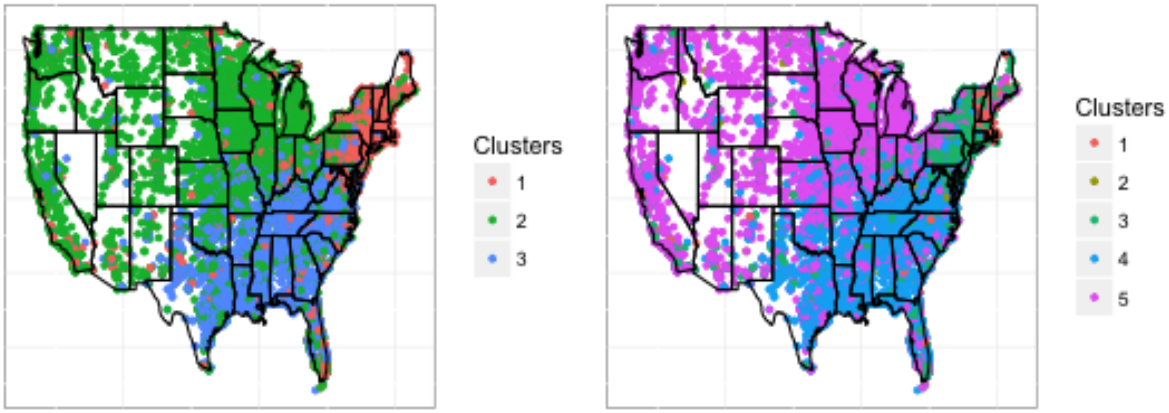

Figure 8: Plot of the survey locations labeled with spectral clustering using the binary dataset with three and six clusters

3.3 Stability of findings to perturbation

As noted in the section discussing clustering different runs of the k-means algorithm produced different results for which locations were clustered together. From Figure 9 and Figure 10 we can see that neither clustering is completely stable, as both figures show plots that have different clusters. However, we can see that the clustering with three groupings is less stable than using five groupings, as each of the four plots in Figure 9 produce different results. Whereas only one plot in differs from the rest in Figure 10, and even in this instance the only substantial difference is the inclusion of the Northwest and West coast into a single cluster, as opposed to two separate ones. This is a trend I saw throughout many runs of the algorithms, not simply the figures presented here.

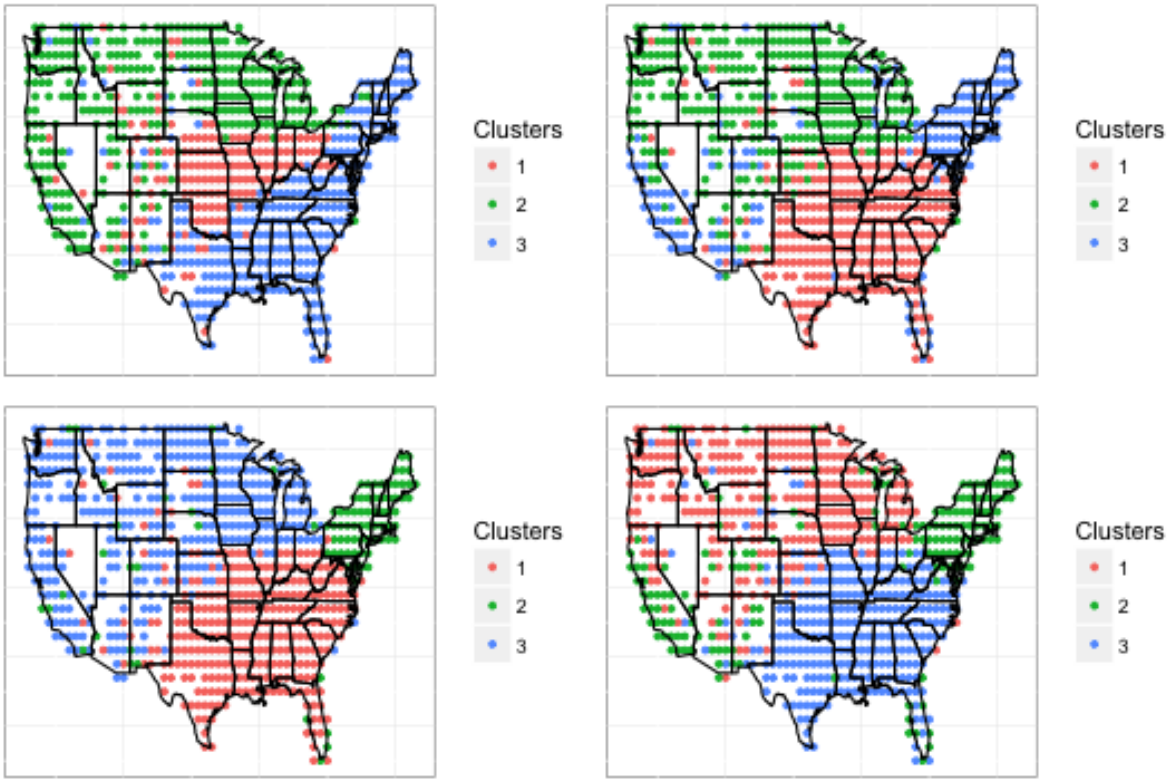

Figure 9: Plots of clustering output for three clusters when running four different k-means on the first eight principal components of the location datasets

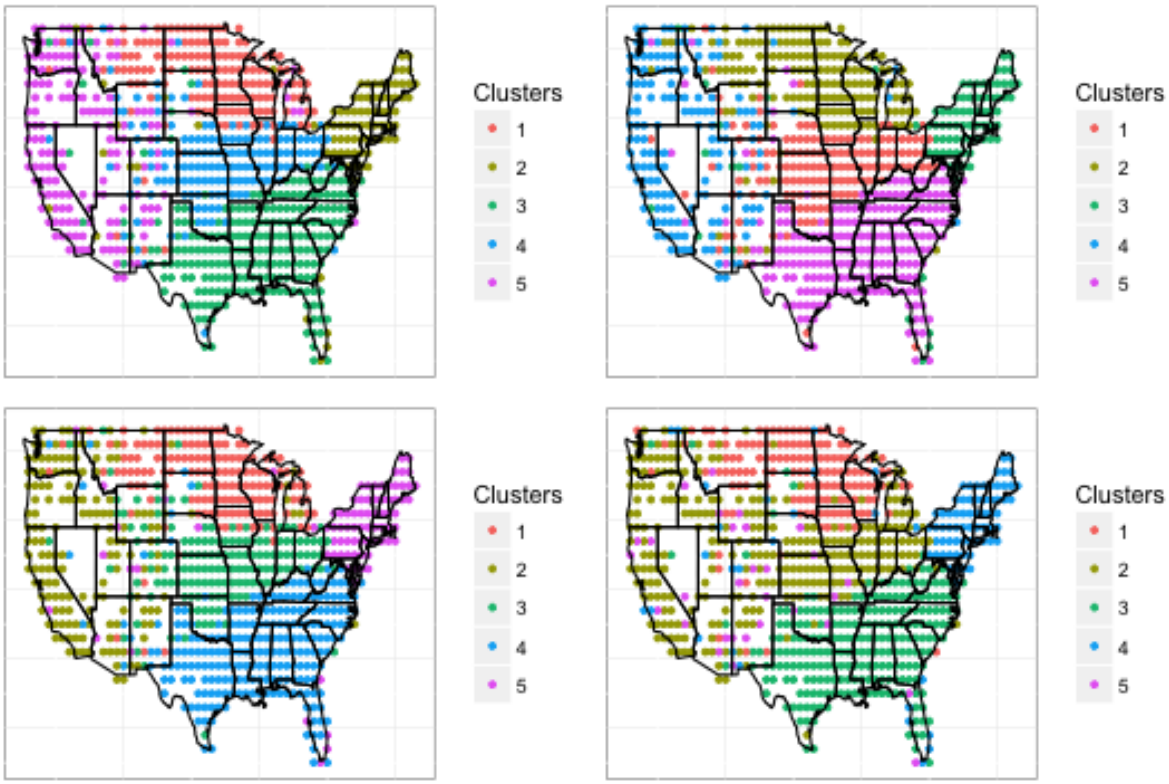

Figure 10: Plots of clustering output for five clusters when running four different k-means on the first eight principal components of the location datasets

Finally, we investigate the stability of the use of five clusters by perturbing the location dataset and determining if we find the same clusters. In order to perturb the data I randomly simulated 10 new pseudo-observations of 758 by 468 matrices filled 0's and 1's, where there was a .14 chance of drawing a 1. I choose this probability because there are 67 questions corresponding to 67s 1's for each individual per 468 columns, and $\frac{67}{468} = .14$. I then incorporated the new "observations" into the binned location dataset and renormalized. I then recalculated PCA and k-means and produced Figure **??**. These observations do not directly correlated to a random answer from a survey, because the 1's could take in place in the vector (i.e. there was no restriction of giving one answer per question).

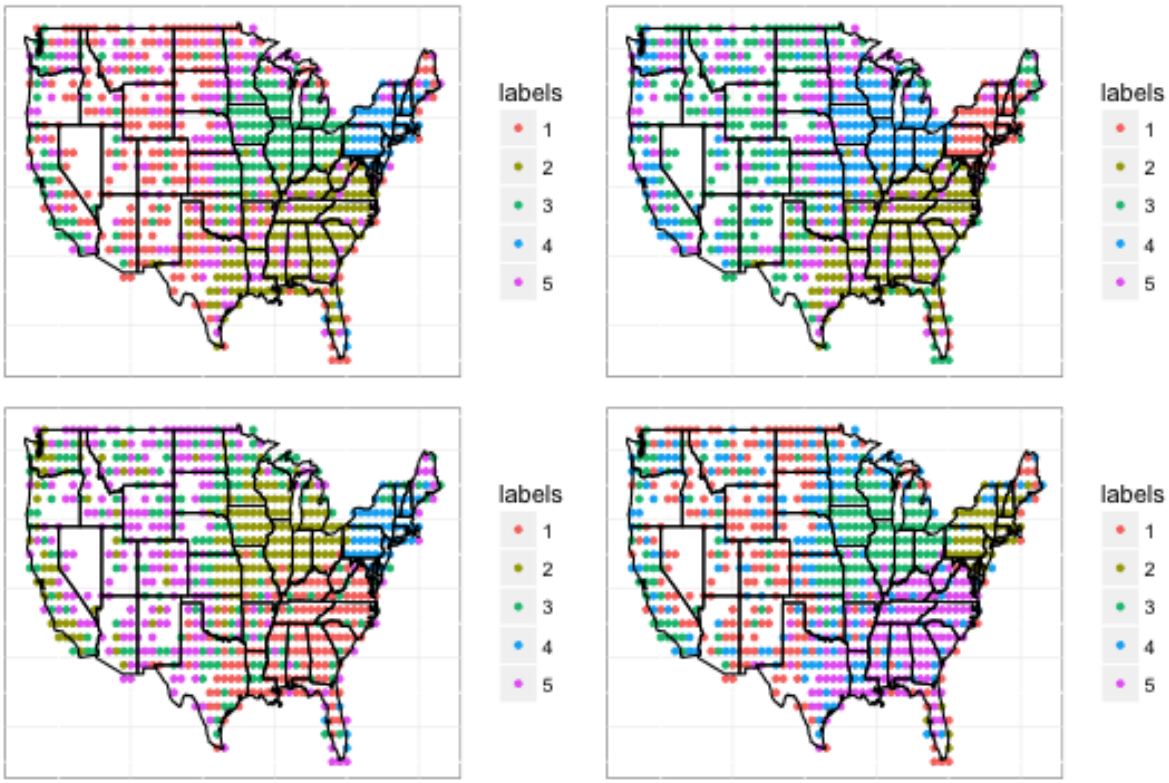

Figure 11: Plots of clustering output for five clusters when running four different k-means on the first eight principal components of the perturbed binned location dataset

Even though, I only added 10 new "observations" the additions changed the the overall structure of the data fairly dramatically, because each "observation" corresponds to adding ten new people into each latitude/longitude square, implying we are adding 7,580 new participants. In the original data many entries remained zero or had a very small number if only a few people choose that answer. The noise that was added had an equal chance of providing any answer as well as an equal chance of being at any location in the United, which we also know not to be true. This perturbation effected the clustering, by making the five groups less defined, but there appears to still be the same basic structure that we saw in the unperturbed data. Additionally, it seems as if the western part of United States was most effected by this perturbation, while the South, Midwest, and Northeastern clusters remained relatively intact. This tells me that those grouping are more stable, than the clusters in the West, which make sense considering there is a larger population on the coasts, making those locations less susceptible to noise. The fact that such adding 7,580 new, completely random observations to the data, did not completely ruin the clustering shows that despite producing different results on occasion the five groupings picked out through spectral clustering have some validity and stability.

4 Conclusion

Overall, the dimension reduction and clustering discussed in this paper appear to show that there the dialectic differences in the United States which can be nicely split into five geographical regions, roughly

corresponding to the common divisions of the Northeast, West coast, Midwest, South, and Northwest. Despite finding some instability in the clustering for different runs of k-means and number of clusters the overall patterns were fairly robust to perturbed the data, particuarily in areas with a large number of observations.