lab2

3034261351

9/22/2018

1 Introduction

In the two-part paper, we will first tune parameters of kernel density plots and loess smoothing with the temperature and humidity of redwood dataset, and then dig into a project of linguistic data.

In tuning paremeter part, data measures the temperature and humidity of redwood trees during a given time period. We experiment with different parameters while plotting the kernel density estimates of distribution of temperature, and while plotting a regression smoother to fit the relationship between temperature and humidity. After that, we illustrated the effect of parameters in our estimation as well as smoothing.

In the linguistic project, the data analyzed in the linguistic project was collected through a Dialect Survey online conducted by Bert Vaux. All the survey questions and each participants' answers are recorded in the survey, together with participants' geographical information including city, state, zipcode, latitude, and longitude. The purpose of the project is to explore potential association between linguistic features and geography. We start by analyzing whether the answers for individual survey question have connection with geography, and then we try to gain insight into all the survey questions simultaneously. Last but not least, we show some interesting finding and illustrate the stability of the finding.

2 Kernel Density Plots and Smoothing

2.1 Kernel Density Plots

To compare the effects of different kernels and bandwidths, we plot the density estimates of temperature of redwood dataset with a series bandwidths and different kernel functions (uniform kernel, Gaussian kernel, and triangular kernel).

1 is a plot of uniform kernel density estimates. When bandwidth is smaller, the plot of estimated density is more jagged. And the plot is smoother when bandwidth gets larger. However, if the bandwidth is too large, we can't portray the data very well because of excessive smoothing.

2 is a plot of Gaussian kernel density estimates. We can still observe that the estimate is more smooth when the bandwidth increases, and the estimate is more jagged and doesn't show the trend of distribution very well when the bandwidth is smaller. The kernel density estimate is almost the same when bandwidth is 5 and 10, which indicates that there is no use in keeping increasing bandwidth when the density estimate is smooth enough.

3 is a plot of triangular kernel density estimates. The plot of kernel density estimate is rougher when the bandwith is smaller, and it is more smooth when the bandwidth is larger. It's almost impossible to extract distribution when bandwidth is 0.1 because the plot is too zigzagged.

To compare among the three kernel functions, their estimated densities in general match with each other. A nice example to compare the three kernel functions would be when bandwidth is 1. The three estimated densities are almost the same, while the plot of Gaussian kernel is slightly more smooth, and the plot of triangular kernel is more jagged.

To sum up, the choice of smoothing parameter is always more important than the choice of kernel functions. If the bandwidth or smoothing parameter is too small, our estimated density might have multiple false modes

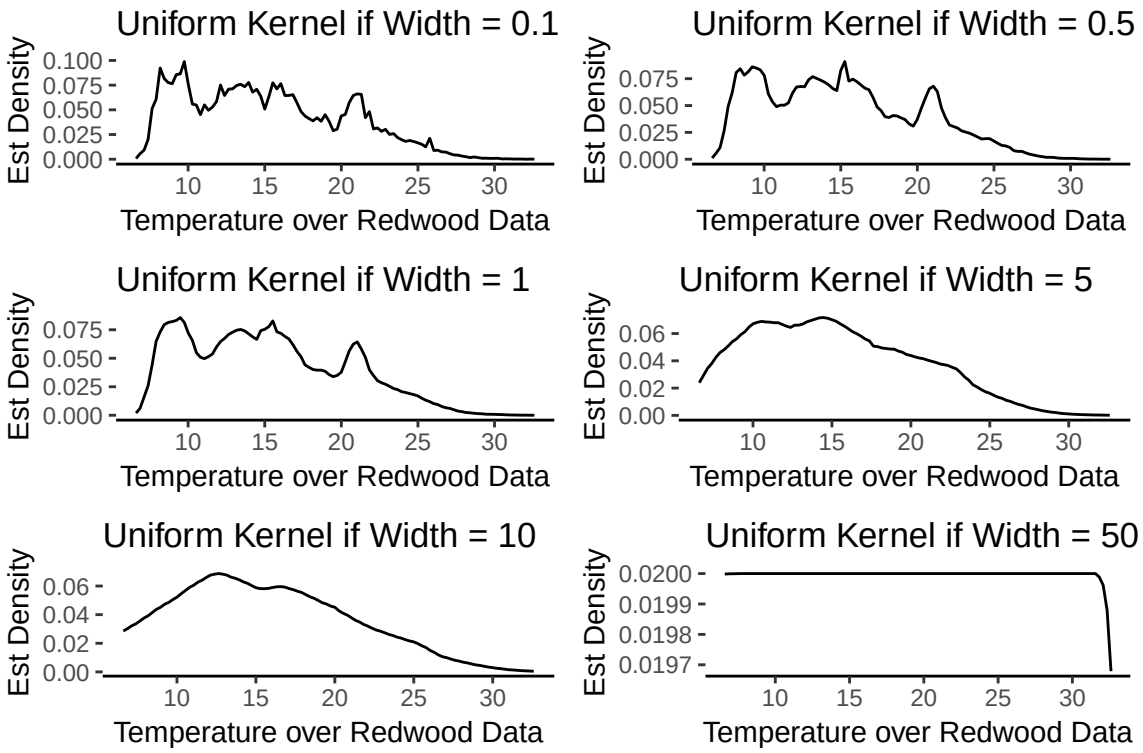

Figure 1: Uniform Kernel Density Estimate Of Temperature vs Bandwidth

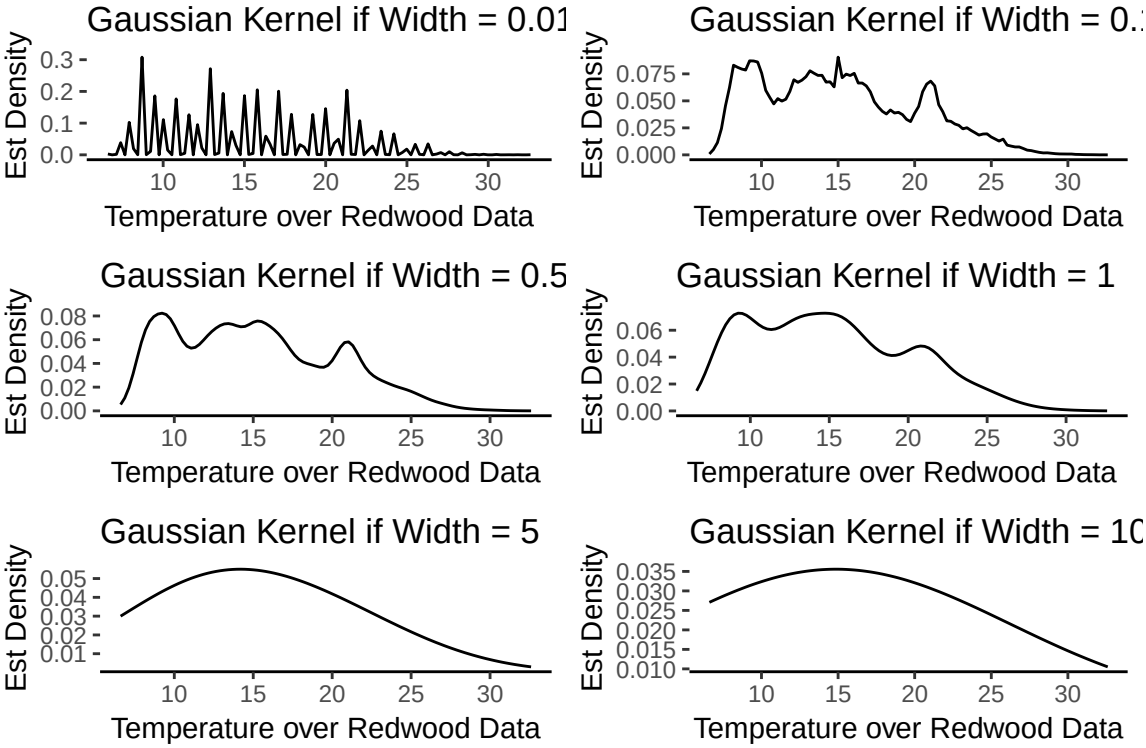

Figure 2: Gaussian Kernel Density Estimate Of Temperature vs Bandwidth

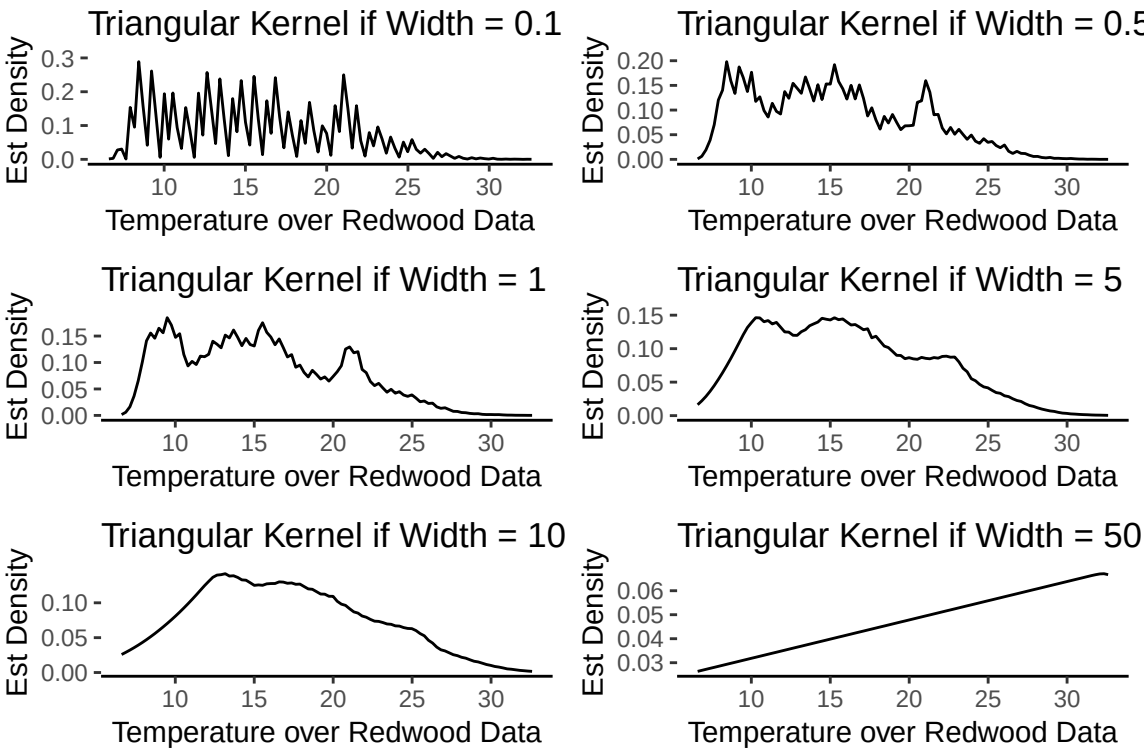

Figure 3: Triangular Kernel Density Estimate Of Temperature vs Bandwidth

with a noiser plot. On the other hand, if the bandwidth is too large, we might throw away many important features of the observed distribution by smoothing.

2.2 Plot Smoother

We choose to plot the temperature versus humidity at noon (12:00:00) every day, and then add a loess smoother for regression. 4 shows the change of loess smoothers as the degree of polynomial varies, and 5 illustrates the change of loess smoothers when the bandwidth changes. From 4, we know as degree goes larger, the model is more likely to overfit data. And the loess smoother fits better when n is 1 or 2. From 5, we can see that as smoothing `span` or `bandwidth` changes, the accuracy of the fitted model also changes. As bandwidth is smaller, the degree of smoothing decreases, which leads to more noise and jagged lines. But if the bandwidth is above some thresholding, the loess smoother wouldn't vary much.

3 Linguistic Data

3.1 Data Cleaning

All survey questions are contained in `quest.mat` in the file `question_data.Rdata`, and we only focus on 67 questions about lexical differences rather than phonetic differences, which are contained in `quest.use`. All answer choices to the questions can be found in `all.ans`. And we have answers of 47,471 respondents across the United States in `ling.data`, together with their self-reported city, state, zipcode, corresponding latitude and longitude information. And there is another pre-processed dataset `lingLocation`, which turned categorical responses into binary data, and then clustered all answers from the same latitude and longtitude together.

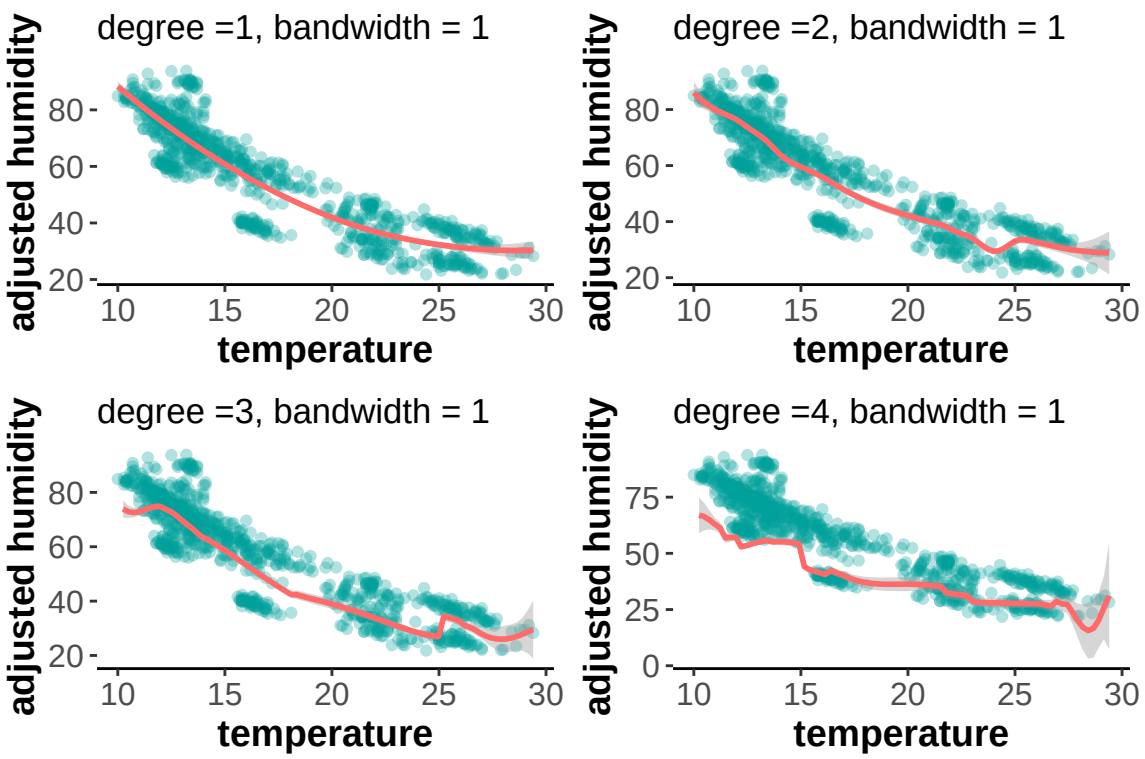

Figure 4: Humidity vs Temperature Loess Smoother: Changing Degree of Polynomials

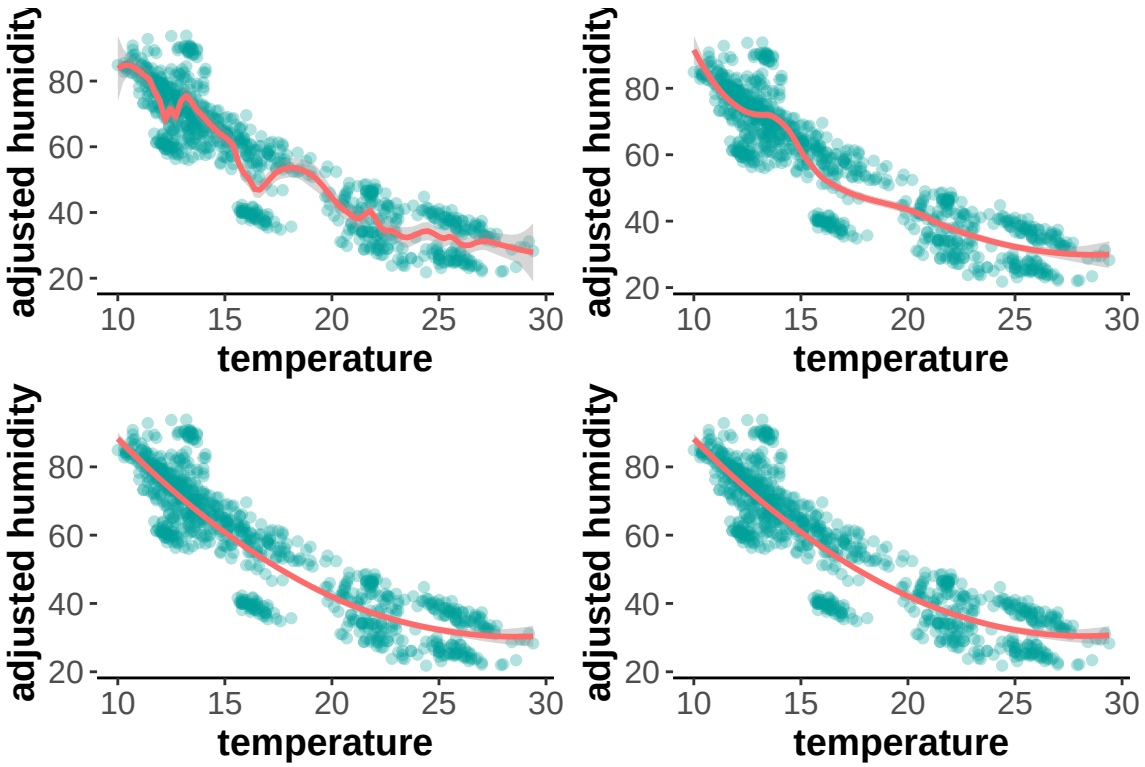

Figure 5: Humidity vs Temperature Loess Smoother: Changing Bandwidth

However, there are missing values in `ling.data`, and some participant-reported location is not in the United States. Below is a detailed description of data cleaning of `ling.data`.

1. Fill up missing values

The geographical information of 47,471 respondents across the United States is contained in `lingData`. To deal with missing values in longitude and latitude information, we geocode with dsk (Data Science Toolkit) to find longitude and latitude information based on the reported zipcode, city, and state. Another problem with missing data is that some respondents didn't answer all questions. There are NA% pariticipants didn't finish the whole survey. Among those who didn't answer at least one question, NA% individuals didn't answer over 60 questions. These people couldn't offer enough information concerning their linguistic features because they have missed the majority of the questions. And they only take up NA% out of all survey participants. Thus, we decide to remove all information of the tiny portion of people because they can't offer any sufficent useful information for our linguistic analysis. At the same time, NA% of individuals didn't answer at most 10 questions. Although they didn't answer all questions, but we still want to keep their submitted answers. In doing so, we denote those missing response as `0` in `ling.data`.

2. Subset of a US map

There are 91 levels of state in our data, which exceeds the actual number of states in the United States. After looking through each state level, we find there are some reported states/location which is out of the United States, and some reported state information doesn't make sense, i.e. 94, N., M, etc. the abovementioned state information are either irrelevant or not trustworthy. Therefore, we decide to only include 48 states of the United States in our downstream analysis, excluding Alaska and Hawaii. Because there are only 1.1955443 % of respondents who have state information other than the 48 states in the United States, which is relatively small.

3.2 Investigation of Individual Survey Question

```
##
## ------------------------------------
##  you all   y'all    you    you guys
## --------- ------- ------- ----------
##   5617     7855    10463    18752
## ------------------------------------
##
## Table: \label{tab:tab1}Top Responses of Q50
```

```
##
## ------------------------------------------------------------------
##  other   I have no word for this   crawdad   crayfish   crawfish
##                  critter
## ------- ------------------------ --------- ---------- ----------
##  1830              2083              10554     12281      17735
## ------------------------------------------------------------------
##
## Table: \label{tab:tab2}Top Responses of Q66
```

```
##
## ---------------------------------------------
##  cocola   tonic   other   soft drink   coke
## -------- ------- ------- ------------ ------
##   113      296    1176       2550      6351
## ---------------------------------------------
##
## Table: \label{tab:tab3}Top Responses of Q105
```

We pick some survey questions and plot the major answers of each question in a US map. Individuals with no response for the question will be excluded from the map. For the sake of clear visualization and less variability, we decide to only adopt top responses in the linguistic map so that we'll be more likely to get distinct geographical groups. The accuracy and objectivity of the data points will not be affected, as only a merely neglectable portion of data is excluded. **??** shows the sorted responses of question 50, and the top four responses take up 95.2409639% of all the data, which represents the data pretty well. **??** shows the sorted responses of question 66, and the top five responses take up 99.2481035% of all the data. **??** is the sorted responses of question 105, and 99.2481035% of the data is represented by the top five responses. Therefore, data quality is guaranteed as we reserve 93-96% of the data information.

Below are linguistic map plots of question 50, 60, and 105. These plots are user-interactive bubble maps (plz refer to the html webpage). Each data point represents a response from one respondent. There are some interactive functions user might be interested in exploring:

1. The response and city of the respondent will pop up if user puts the cursor over some data point.
2. Users may double click the legend to filter answers in the map. Specifically, the corresponding group of answers will disappear from the map if some legend name turns dim. In that way, users can better distinguish between different answers and explore their own interested answers for each question.
3. Users may zoom in or zoom out to observe the distribution of answers for one specific geographical district.

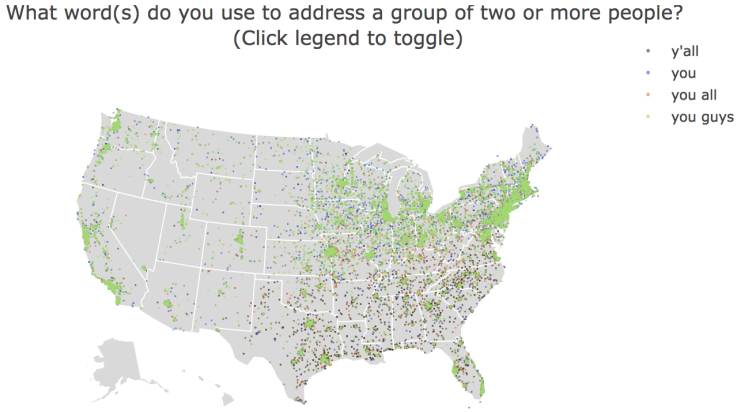

Figure 6: Grographical Distribution of Survey Question No. 50

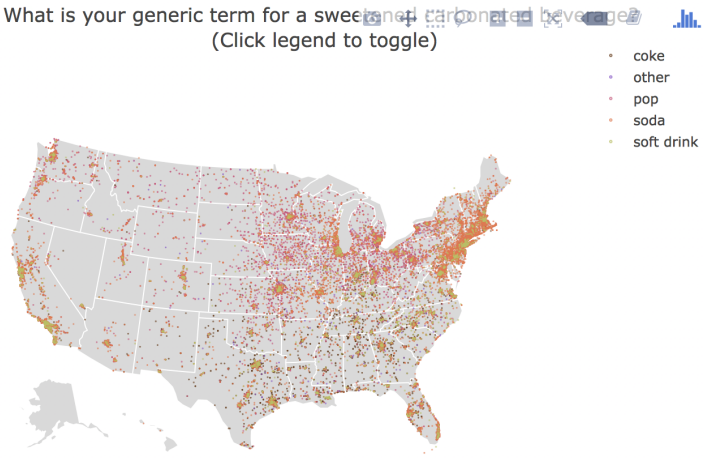

Figure 7: Grographical Distribution of Survey Question No. 105

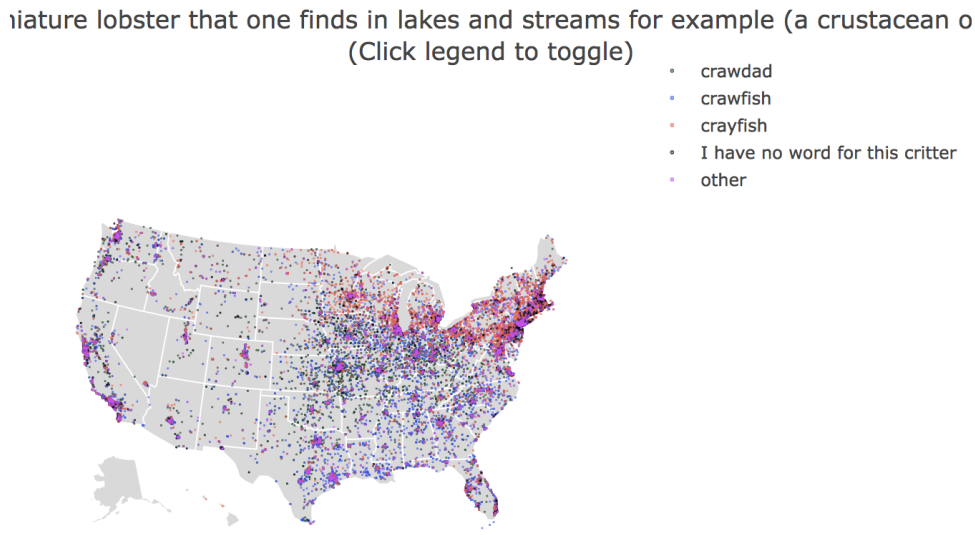

Figure 8: Grographical Distribution of Survey Question No. 66

We first notice data points in all linguistic maps are not evenly distributed among the United States. Most individuals taken the survey live in the northeast part, and New York in particular. On the contrary, data points in west side are quite sparse, and there are only some clusters near the coast of California. We need to bear the unbiasness in mind, because the absolute number of data points can"t decide geographical groups. We can only declare some answer choices have a geographical group when the relative amount of individuals choosing the answer is still large compared with the overall number of individuals within that region.

6 explores the distribution of answers in question 50. The answer of 'y'all' is mostly in the southeast (coastal southern and mountain southern) area, whereas 'you guys' takes up the northern and western part. Comparing it with 7 in question 105, the answer of 'coke' is mainly around the southest part, and the southeast part is roughly dominated by 'coke' too. In that way, we may define the southeast side as one distinct geographical group defined by question 50 and 105. If someone answers 'y'all', we can reasonably state that this person is very likely to be into brand quality, and refer a sweetened carbonated beverage as `coke`. And vice versa. From 7, we can see answers with 'soda' are most clustered in New York area, and answers with `pop` are mainly in the north side (great lakes, and upper midwestern area). Interestingly, the answers of 'crayfish' in question 66 from 8 are also around the northeast side (great lakes, upper midwestern, and New York area). Thus, we may define another distinct geographical group in the northeast side. And we may seperate New York area as one distinct region too. If someone picks 'crayfish' as the answer which is in the northeast side of US, we predict there is higher probability that this person would refer `pop` as a sweetened carbonated beverage. Besides that, we don't have further concrete evidence to identify geographical groups or predict. Next, we will try to gain some insight from the whole dataset with all questions, and see if we could have stronger evidence for clustering.

3.3 2. Dimension Reduction with Binary Data

We first convert categorical data of the full dataset into binary data, i.e. each column represents an answer choice for some question. If one respondent chooses some answer, then the corresponding number for the person and choice is 1, otherwise 0. And then we perform principle component analysis (PCA) on both `ling.bin` (binary data converted from `ling.data`) and `ling.location`, to compare which does a better job in explaining the variance, or give better clustering result. We decide not to rescale the binary data because we are more interested in the difference between two categories 0 and 1, and it is improper to treat discrete 0/1 in any way as continuous. If someone doesn't respond to one question, the vector encoded in `ling.bin`

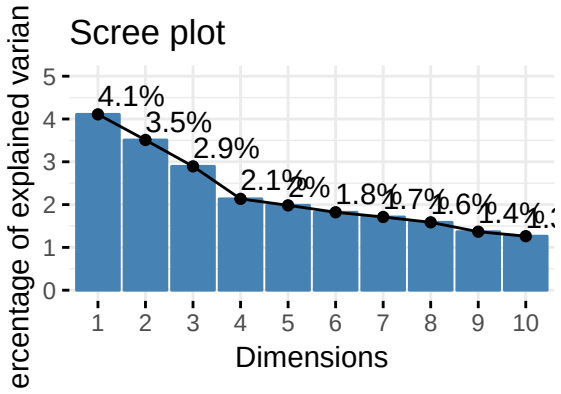

Figure 9: Plots of Ordered Eigenvalues of ling.data

will be all 0 and we will not take no response as one separated level. For example, suppose John answered all questions except for the first one, and the first question has 4 choices, then ling.bin would encode (0, 0, 0, 0, . . .).

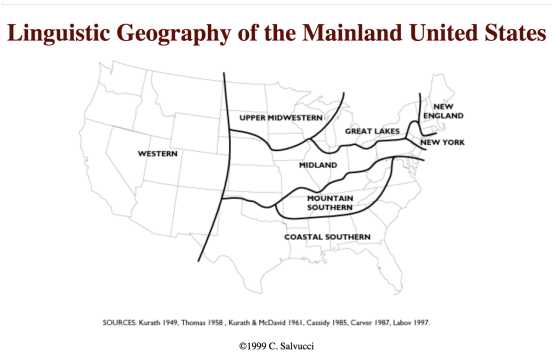

Figure 10: Linguistic Geography of US (Domain Knowledge)

9 is the Screeplot showing the eigenvalues of binary linguistic data for each dimension ordered from largest to the smallest. Top 5 dimensions give us 15.2% of explained variance, and top 10 give 23.2%. Hence we would need too many dimensions to represent the full variability in ling.bin, and top 2 dimensions is not representing data well at all. For instance, in 11 we try to plot the binary linguistic data onto different projections in PCA. Data points in different colors represent individuals in different regions. According to our domain knowledge, Salvucci et al. (10) defined seven linguistic regions of the United States by their knowledge in linguistics. They are Western, Upper Midwestern, Great Lakes, Midland, New York, New England, Mountain Southern, and Coastal Southern regions. We will classify all individuals into those regions by their state information. And people in Alaska and Hawaii will be a unique linguistic region because people there don't share linguitic customs with others. 11 shows the binary data projected onto different pairs of components, labeled by region. We can see that the projection of PC2 vs PC1 give us the relatively nice clusters with respect to regions. However, we can still see severe mixing between different labels/colors, and there are only 3 clusters being separated in the dense tangle of data points. That's most likely because the number of cluster we specify is too large for our dataset, i.e. there are only less than 7 different linguistic clusters in data set, thus we decide to try to find the optimal number of clustering first.

Another potential problem with PCA plots of binary `ling.data` is that all data points are jumbled together. And the top PCs can't explain the variance very well. So we perform PCA on `ling.location` as well, and see if it would make a difference. The dimension of `ling.location` is [781, 468], so it is not as sparse as binary `ling.data`. From the Scree plot 12, we can see the first component explains 63.8% of the variance,

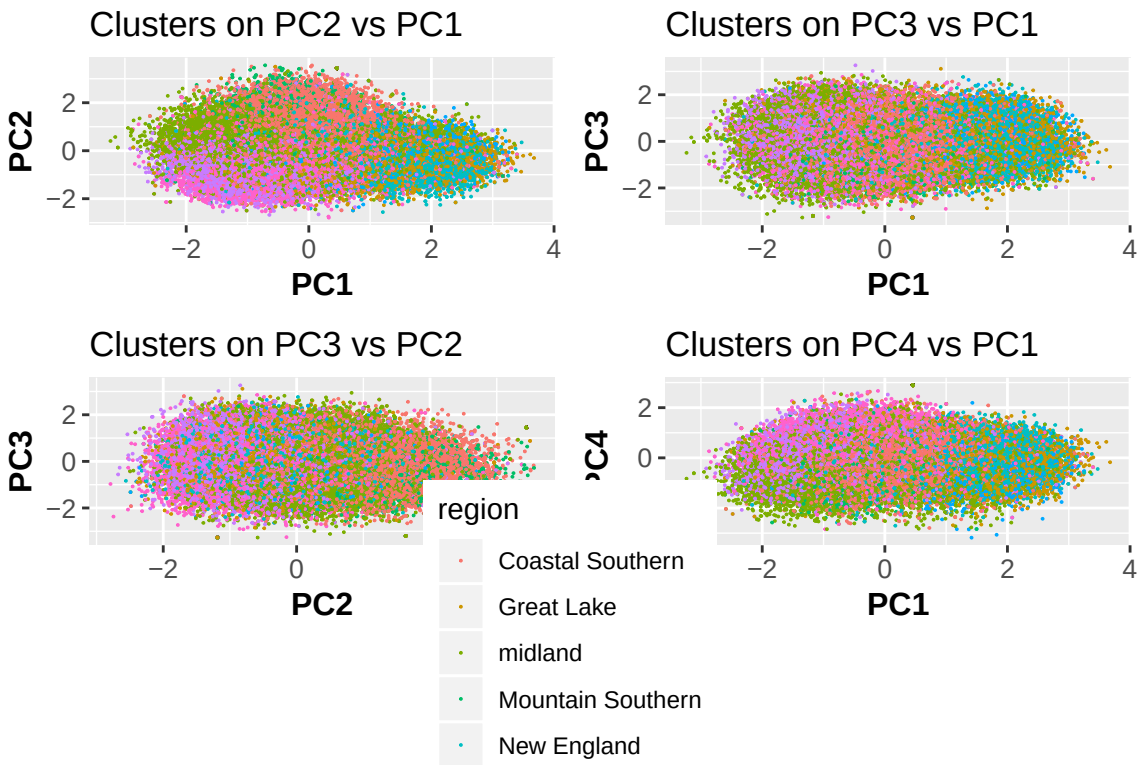

Figure 11: Linguistic Data Projected on Different PC Spaces

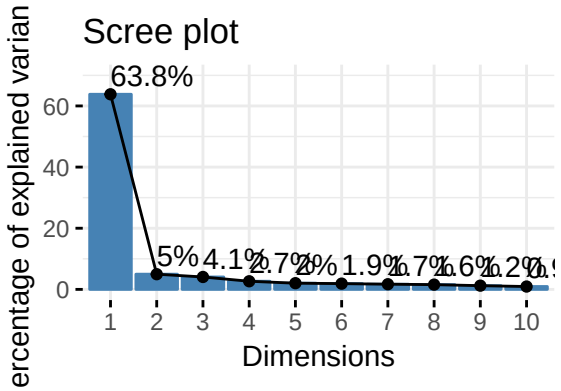

Figure 12: Plots of Ordered Eigenvalues of ling.location

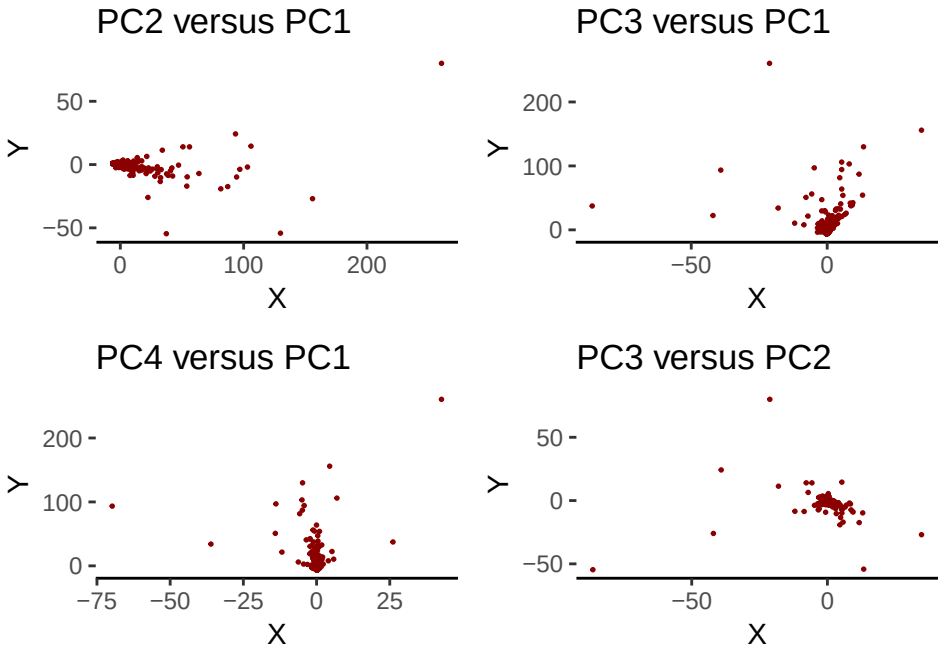

Figure 13: Re-projection of Components for ling.location

and the second component explains 5% of the variance. The top two components can do a much better job in explaining the variance and represent the data compared with binary `ling.data`. 13 shows different projections of `ling.location` data. We can better visualize data from 13 with respect to isolate geographical groups, because data points are much more spread out now. Each data point in 13 represents all individuals of the same longitude and latitude, and we don't care about the within-group difference for all individuals.

3.4 3. Clustering Analysis to Isolate Geographical Groups

We're going to apply three different clustering methods, and compare their results to see if we can find the optimal number of clusters. The first method is to apply k-means on the major PCs, the second method is to perform Partitioning Around Medoid (PAM), and the third method is hierarchical clustering. We're interested in comparing the different effects of these methods, and we hope to isolate distinct geographical groups via the clustering methods.

3.4.1 K-means and Elbow Method

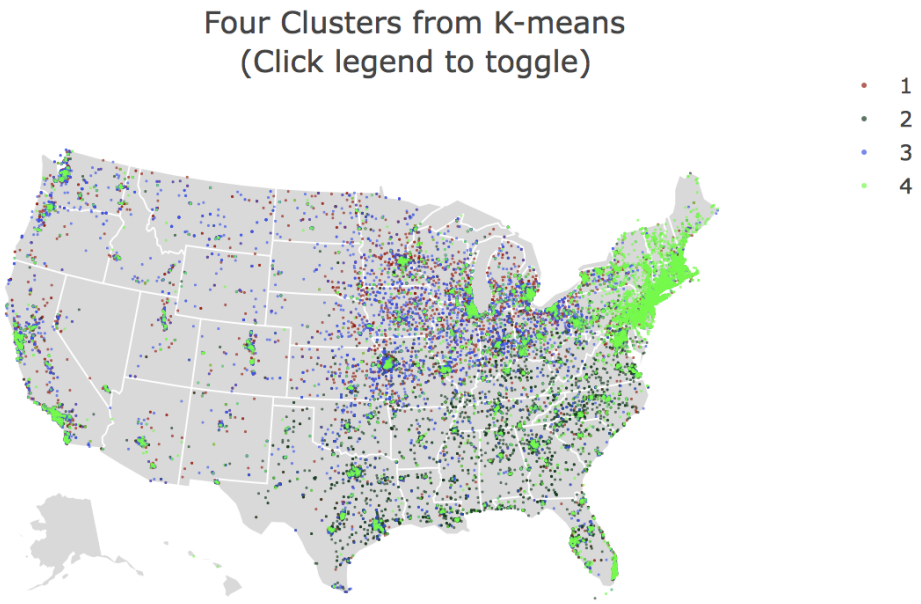

Figure 14: Survey Takers Are Labeled by Their Clusters k = 4

We first try to perform K means on the projection of top PCs. We decide to pick the top 100 PCs, because the top 100 PCs explain 55.78278% of the variance, which is trustworthy to represent the whole set. And then we tried to use the Elbow Method to find the optimal number of clustering K. For computational efficiency, we will look for the optimal K by using `ling.location`. Because the optimal clusters `ling.location` and `ling.data` give us will surely make no difference. The elbow method figure shows that k = 4 would be the optimal number of clusters given Elbow Method. Then we separate `ling.data` into 4 clusters by K-means, and label each individual as one cluster. It can help us to reconstruct the US map with the labeling of everyone's cluster. We can clearly see 3 clusters from 14: cluster 1 on the northeast side including New York, cluster 3 on the south side of US, and cluster 4 on the Upper Midwestern, and western area. It's worth noting that data points which belong to cluster 2 are sparsely spreaded acorss the mainland, thus cluster 2 a redundant cluster in real life.

3.4.2 Multidimensional Scaling and PAM

We first try another dimension reduction technique, called multidimensional scaling (MDS). And then we perform PAM on the reduced dataset. To calculate MDS, we first convert all non-zero counts in `ling.location` into 1, and then calculate the Euclidean matrix of the data. We perform multidimensional scaling based on the distance matrix, and get a dataset structure which is similar to a quadratic line in 15. We then calculate the average silhouette width for different number of clusters to find the optimal k. And we'll get the maximized average silhouette width when the number of clusters is 3, according to 16. We plot the silhouette graph when k = 3 in the right figure of 16, and we can clearly see that the height of each block is almost the same, and every silhouette width is all above the average silhouette width. That's the reason why we choose k = 3 in this case.

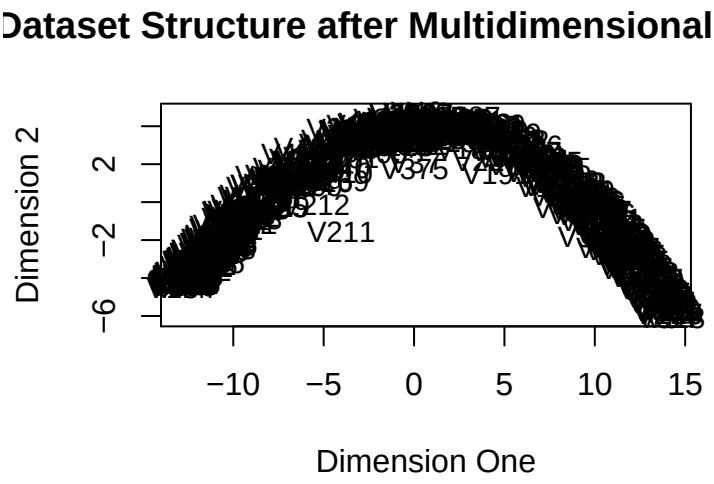

Figure 15: Dimension Reduction Using Multidimensional Scaling

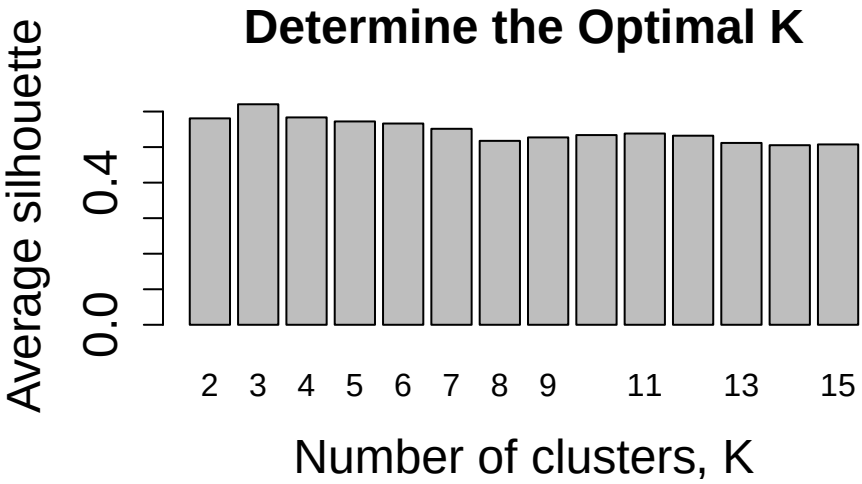

Figure 16: Average Silhouette Length vs Number of Clusters

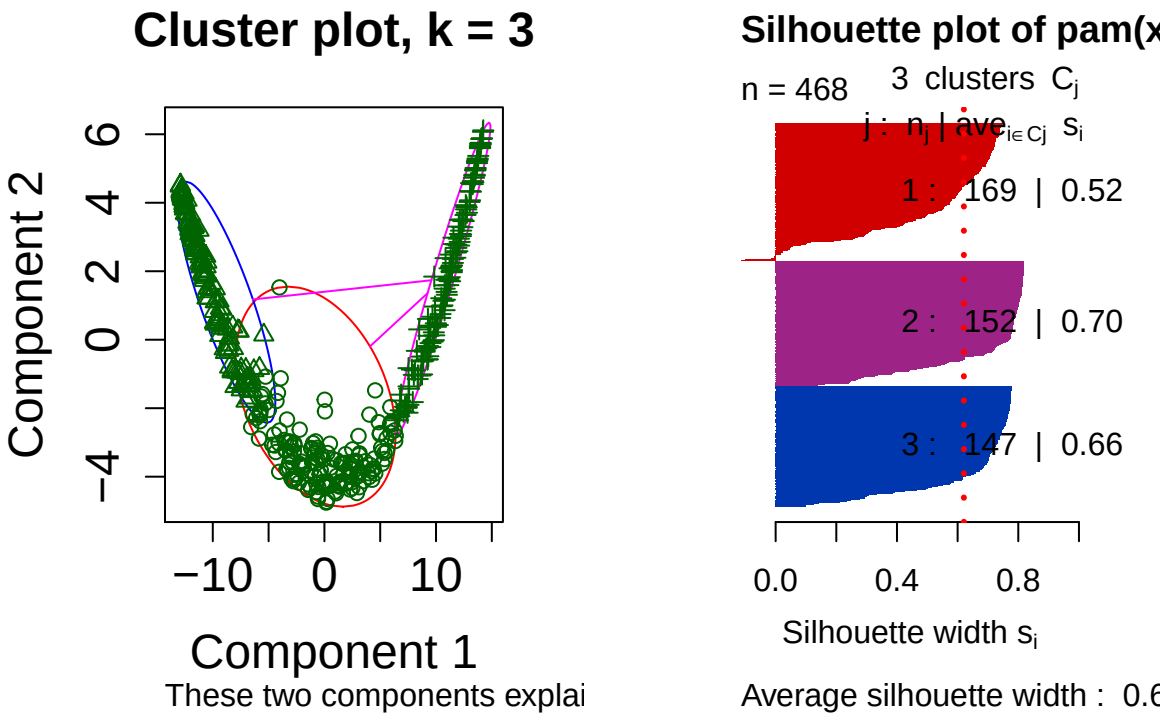

Figure 17: Silhouette Plot with 3 Clusters

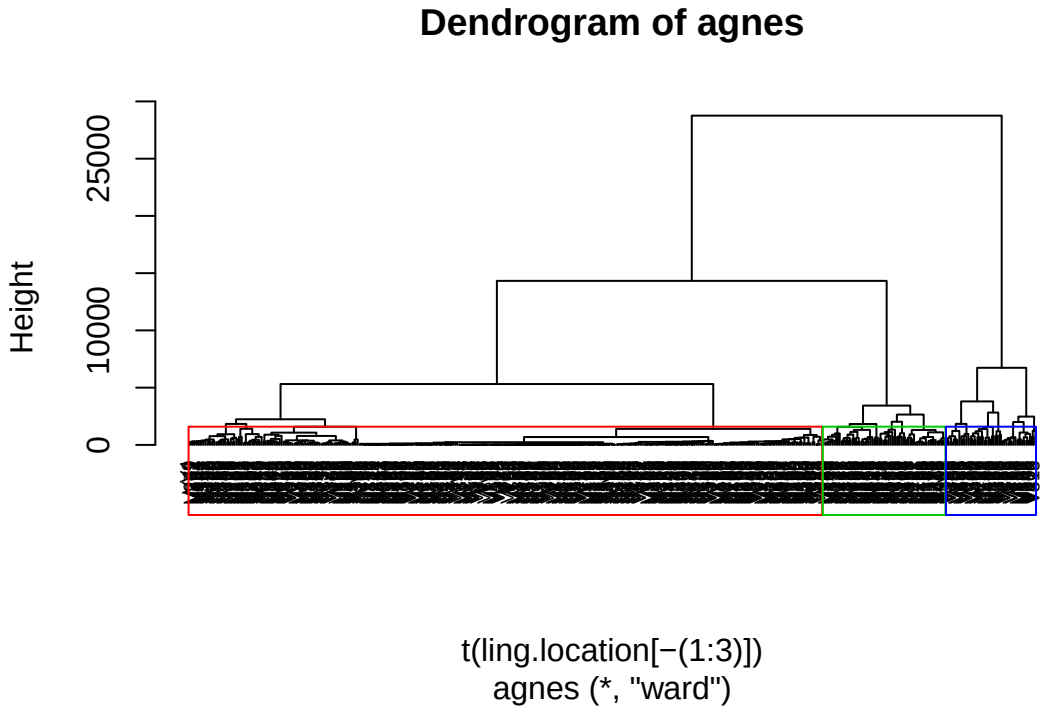

Figure 18: Dendrogram of Agnes

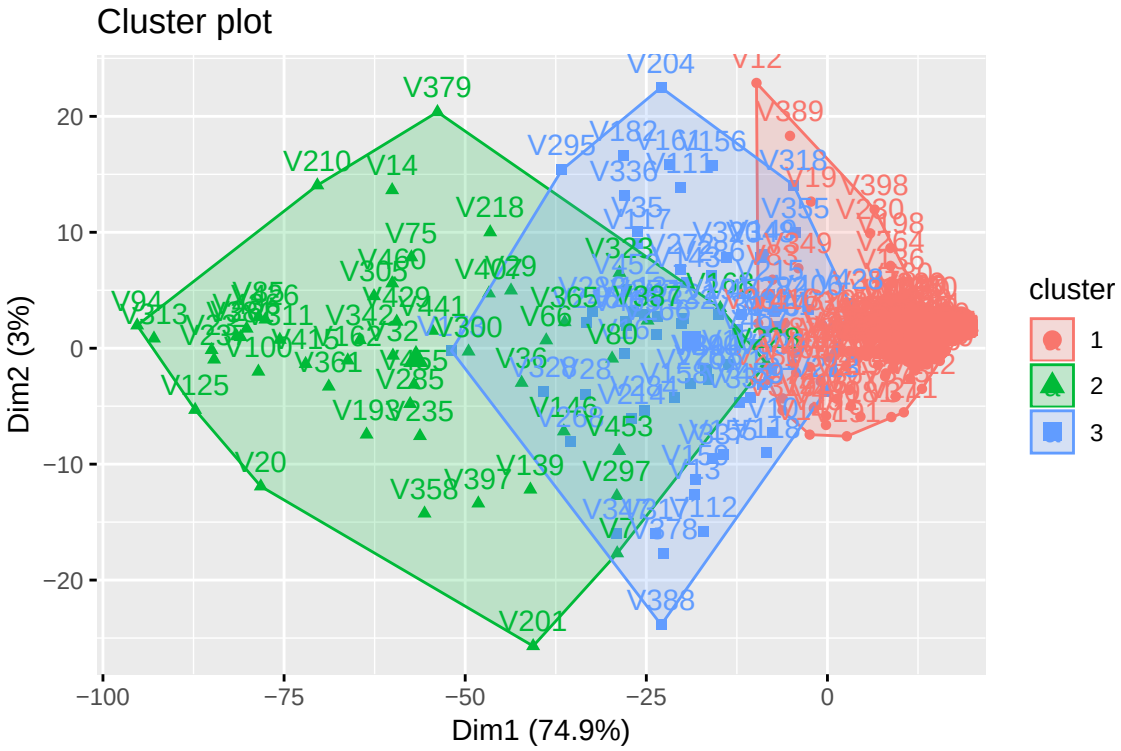

Figure 19: Cluster Plot of Hierarchical Clustering

3.4.3 Hierarchical Clustering

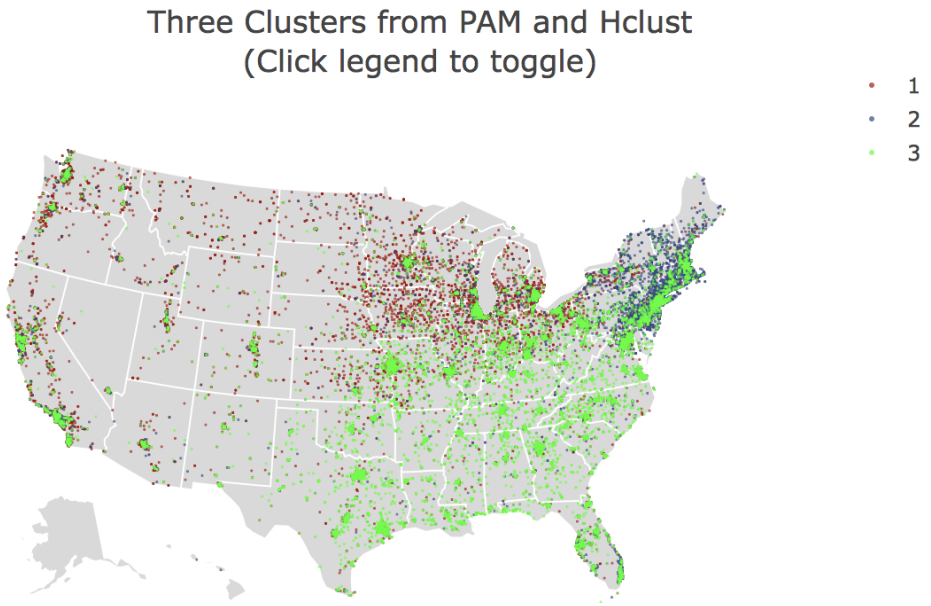

Figure 20: Survey Takers Are Labeled by Their Clusters, k=3

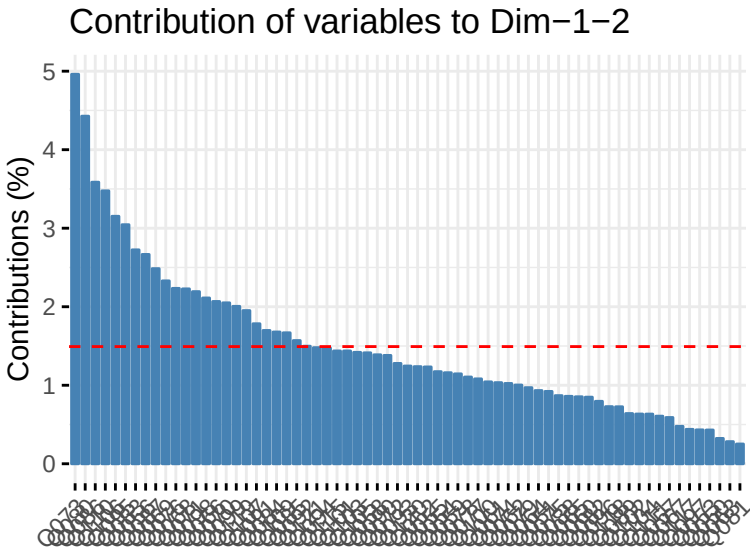

Figure 21: Contribution of Variables to Dim 1 and 2

We all agree that dialects would be differentiated in smaller regions, and the dialects will have more similarity in larger neighborhood. For instance, people might either speak French or English in some country, and those who speak English either speak American English or British English. Those who speak American English will also have their own dialect differences. This kind of hierarchical structure of linguistic data suggest us to explore hierarchical clustering. And we apply Ward's method to draw the dendrogram of agnes first. And we also mark three different clusters in 18 in red, green, and blue. 19 is a cluster plot to show 3 clusters of ling.location data after we perform hierarchical clustering.

Thus, we know the optimal number of cluster is either 3 or 4 according to the abovementioned clustering methods. Comparing between 14 and 20, we know that when the number of clustering is 3, we would have less mixing between clusters. Since the boundaries of three clusters are much more clear and obvious, we decide to adopt the number of clusters as 3, and use 20 as the clustering of the whole dataset. And there is continuum especially at the northeast side of US, specifically New Jersey, Connecticut, Rhode Island, and "the City" New York. By continuum, we refer to the fact that the clusters are highly mixed up there, and it's difficult to distinguish the boundaries of the clusters. By 21, we know the question that explains the most variance to dimension 1 and 2 is question 54: He used to nap on the couch, but he sprawls out in that new lounge chair anymore. And question 54 is the question that produce the continuum because it causes the most variance/ 'noise' to the projection on the first two components.

3.5 Stability Analysis

We can either do algorithm perturbation (i.e. changing the starting points of k-means), or do data perturbation (i.e. subsample or bootstramp). Here we give example of algorithm and data perturbation and show that the clustering is stable when number of cluster is 4. (Note: Given time constraint, I can't output the plot. But the codes are here for your reference.)