Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

SID:3032130362

October 5, 2017

1 Introduction

Dialectology is the scientific study of linguistic dialect. It studies variations in language based primarily on geographic distribution and their associated features. My study focus on the relationship between dialect with geography distribution, which based on the assumption that the language use of a person is influenced by his/her surroundings.

In Nerbonne and Kretzschmar's two papers, they both reviews that past study of dialect: featured questions and canonical studies were conducted by the dialiectologists systematically. However, the deliberate assumptions, different weight considerations, insightful formulas, together with the modern computational methods are among the most advanced techniques to help analysis dialects. My study applies the common dimension-reduction and clustering methods to analysis the geographic effect on dialect. The main tool for programming is R, but Python, which is more effective is also used to get the Silhouette scores and to develop the interactive web application.

2 The Data

Overall, the data is of good quality. *lingData.txt* contains the geographic information and the answers from the respondents. 1020 lines of missing values are found and some the names of states/cities are totally wrong, such as Flatbush(Brooklyn)andWyanda, fhhjdhj and etc. Interestingly, some the name of the cities in the table do not match the zip codes, which is compared with the *zipcode* package. For longitude and latitude, some of the data were collected in Alaska or Hawaii, the number of samples from there are too sparse to make a difference on the whole data set(only 1.06%).

The *lingLocation.txt* file is a complete dataset: no missing values and no obvious outliers. Ideally, every participant should answer exactly 67 questions, but it is not the case. Figure one shows that in some places participants answered less than 40 questions on average, while most places answered more than 60 questions. Finally, the *question_data.RData* shows the answers corresponding to the questions.Ummm....nothing special.

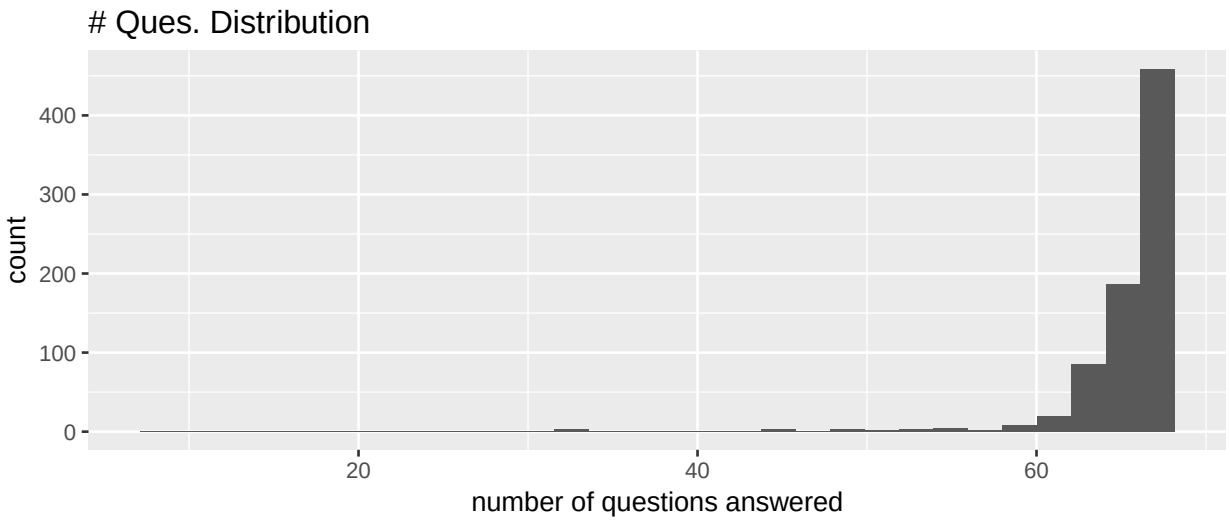

Figure 1: Histogram for the average number of questions answered in each location: more than 95% of the places with the participants answering more than 60 questions.

2.1 Data quality and cleaning

This dataset isn't as bad as the redwood data, but there are still some issues. You should discuss them here and describe your strategies for dealing with them.

2.2 Exploratory Data Analysis

First, I selected one of the questions and its answers to get an intuitive impression. Some of the questions do not have large differences with respect to the locations. Question 80: *What do you call it when rain falls while the sun is shining?* has totally eleven different answers. The most common answers are: *liquid sun, money's wedding, sunshower and the devil is beating his wife*. One of the most significant pattern here is a north verse south trend. The answers *sunshower* were mostly found in the northeast part, while *the devil is beating his wife* mainly found in the southeast part of U.S.

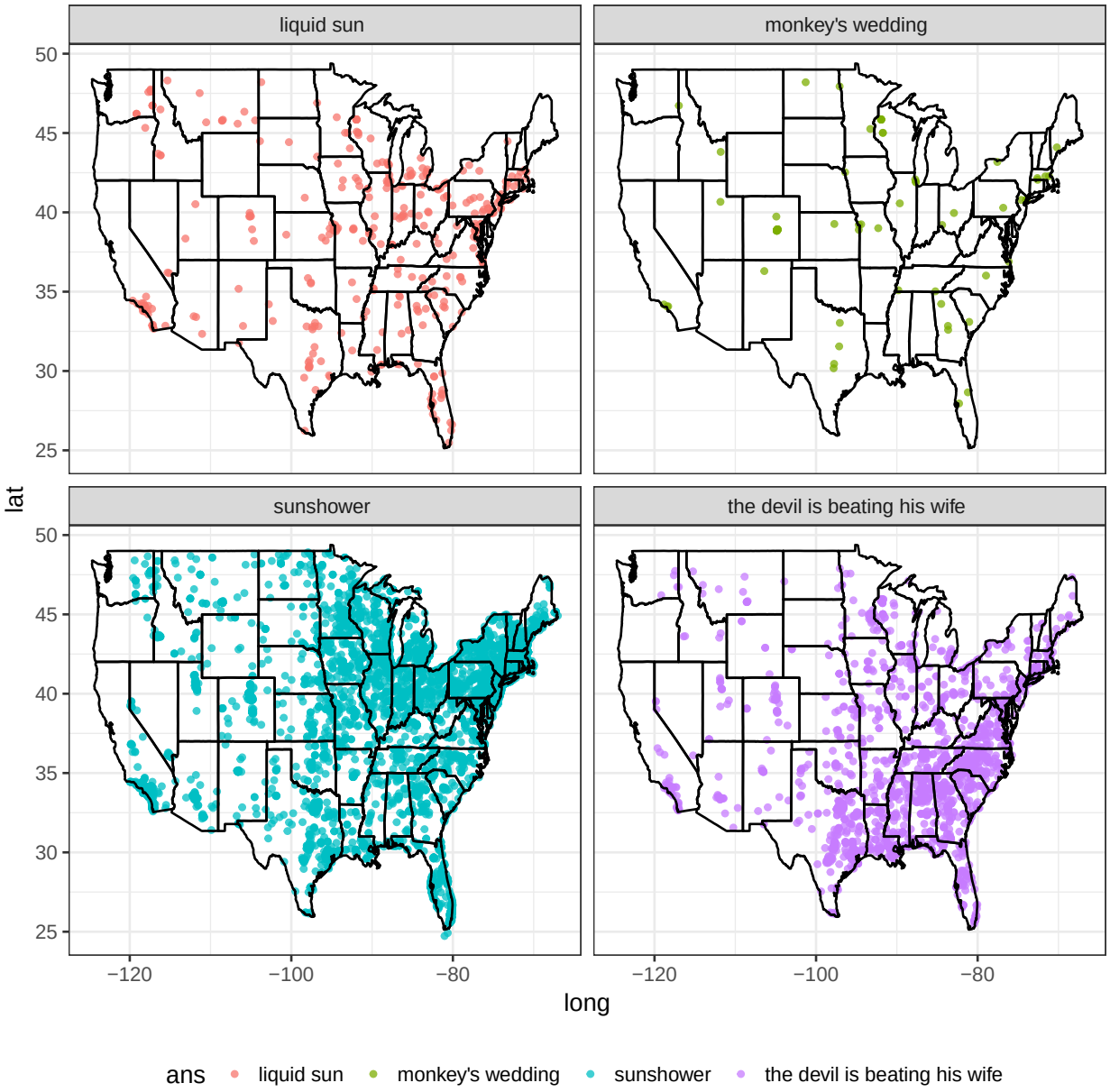

Figure 2: Scatter plot of the answers of Question 80.

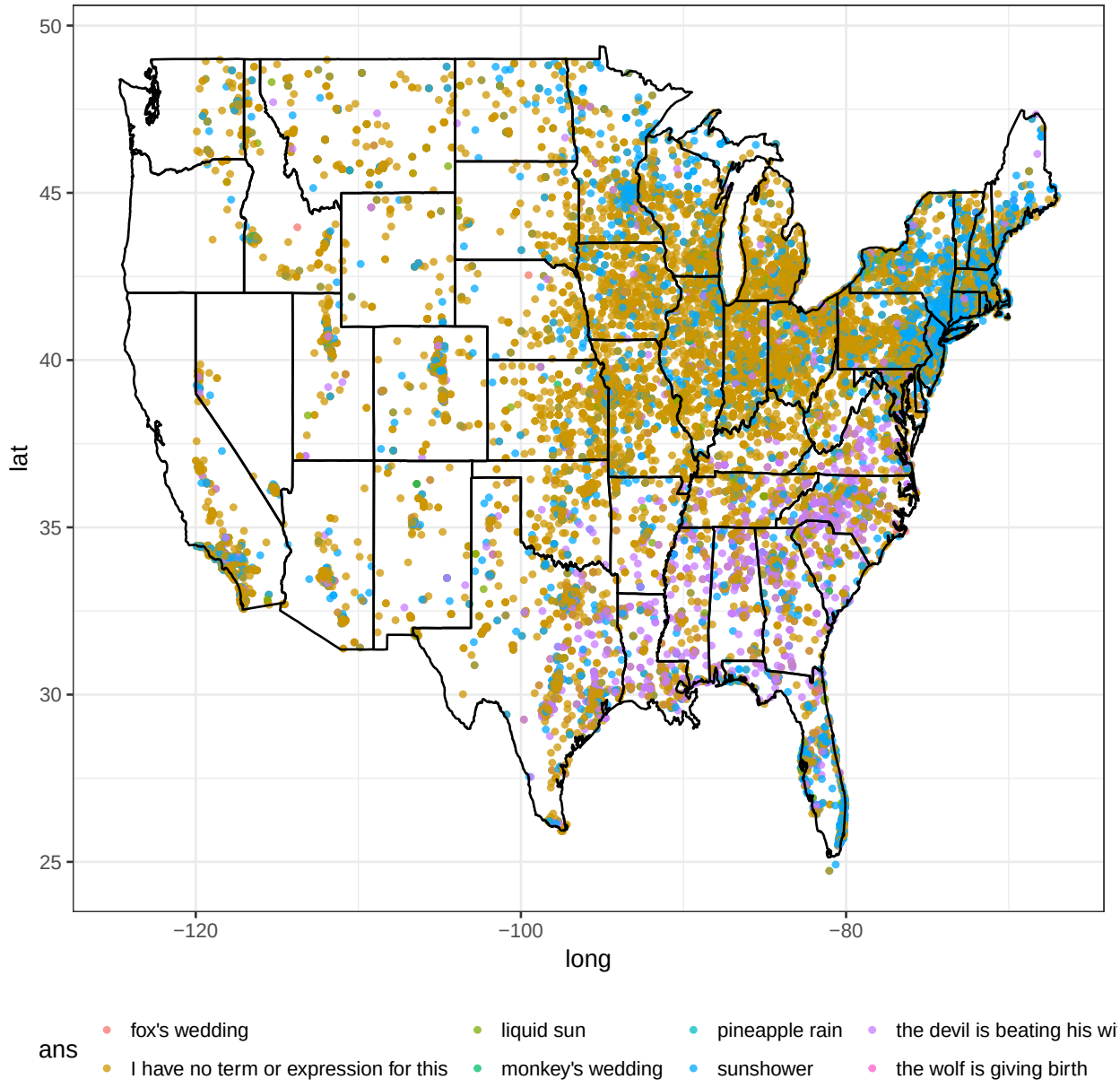

Figure 3: the aggregated plot of question 80. It is more clear than there is north verse south trend. Also the answer *I have no term or expression for this* is actually a large group, which is not analyzed in figure two.

Figure 3 shows a more clear separation: northeast, west and south. The locations of the highest population(number of participants in the study) density are shown in figure 4: New York, Chicago, Houston, Los Angeles and etc. Although the east coast shares a small fraction of areas of U.S., the population there is so large to be representative to form a dialect group.

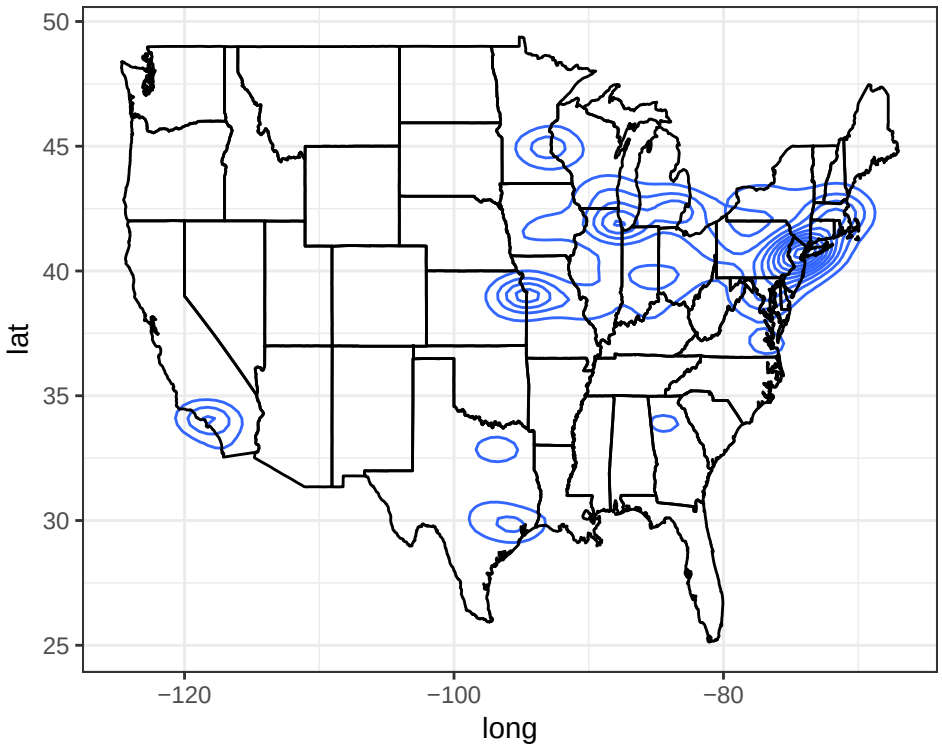

Figure 4: Population density contour. The blue circles point out the locations with the highest population density.

3 Dimension reduction methods

In this part, I will discuss and show plots about the results of the dimension reduction techniques I tried: main PCA, hierarchical,K-means, random projections, etc.

- PCA: original lingData has 47471 rows of observations and after being converted to binary data, it has 468 columns as the answers to the questions. I tried the PCA with scaling and PCA without scaling. To scale the data in this binary case has little reason. And the results of those two PCAs further justified that the non-scaling one is much better, which required about 160 principal components to explain 90% of the variance, while the scaling one needs about 330.
- Kmeans: the methods for kmeans I tried first in the default *kmeans* function in R. Several numbers of clusters were tried to get different results. The *pamk or clusGap* function in the cluster or fpc package, which is used to determine the best number of cluster, is unfeasible here because of the little RAM of my laptop(even after random sampling to 10000).
- Determine the number of clusters and number of Principal components: the Silhouette scores for clustering are calculated in Python. And I also tried to use different number of principal components for clustering. Interestingly, the clustering results from two principal components have sometimes(kmeans is not stable)little differences from the results of 155 components, even though the first two components explain little of the variance.

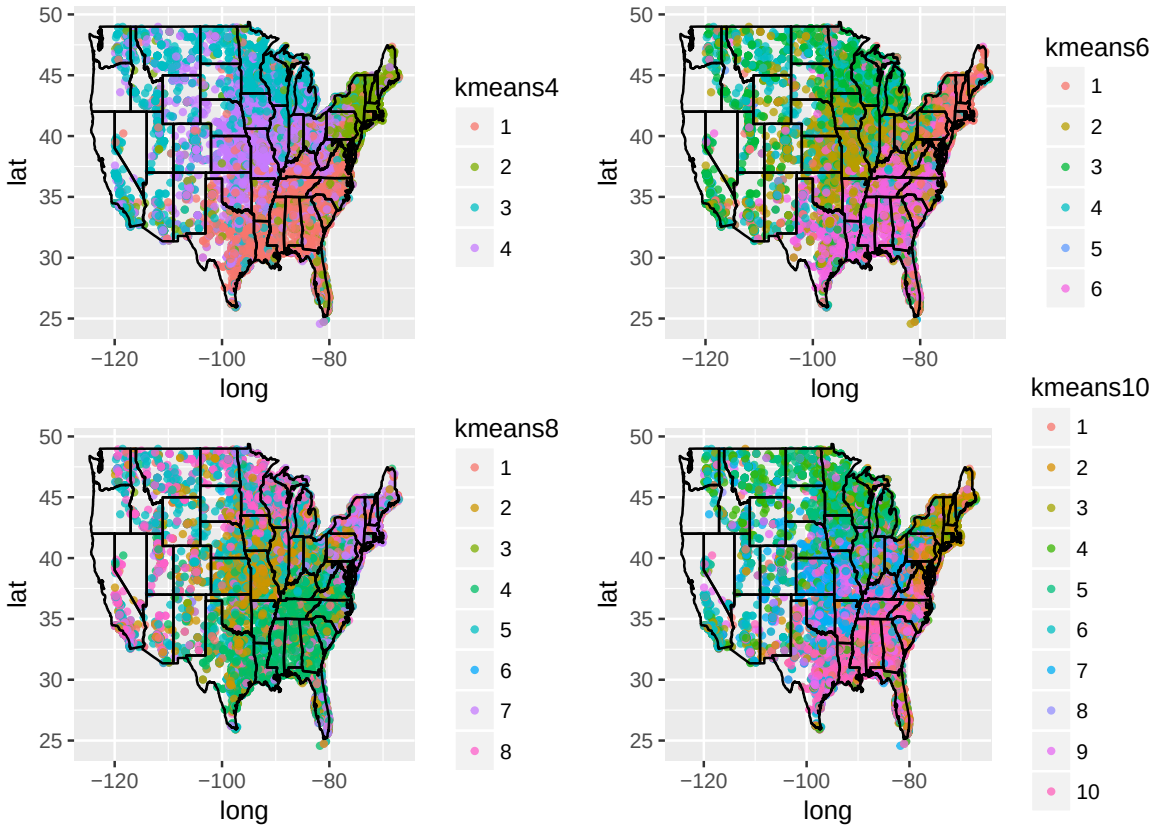

Figure 5: Kmeans with different groups.

I used PCA for dimension reduction on the linguistic data set and it performed very well as expected. The burglarizing step projects the original data set into a higher dimension, which leads to sparsity. Of the 468 dimensions, I picked up the first 155 dimensions for analysis, and then performed kmeans of 4, 6, 8, 10 groups for clustering. Figure 5 shows that different groups have the similar general pattern: we can find at least three different areas(northeast, southeast and the rest.) Some of the clusters show a clear difference between the central United States and other places. After that, I did the same kind of clustering base only on the first two components, which account for only 7.5% of the total variance. Figure 7 compares the clustering results from the two ways. Similar geographical distributions were found: northeast, south, middle and west, which confirms the findings in the previous study.

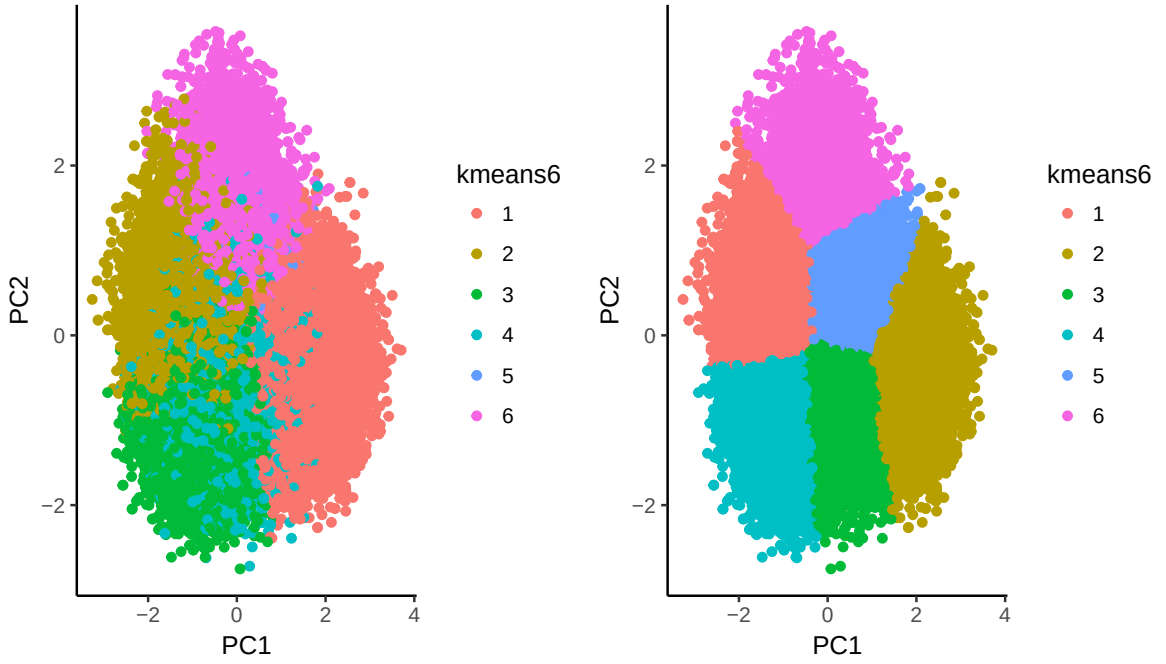

Figure 6: clustering comparison between 155 components and 2 components. As the kmeans method is unstable. The plot may vary each time.

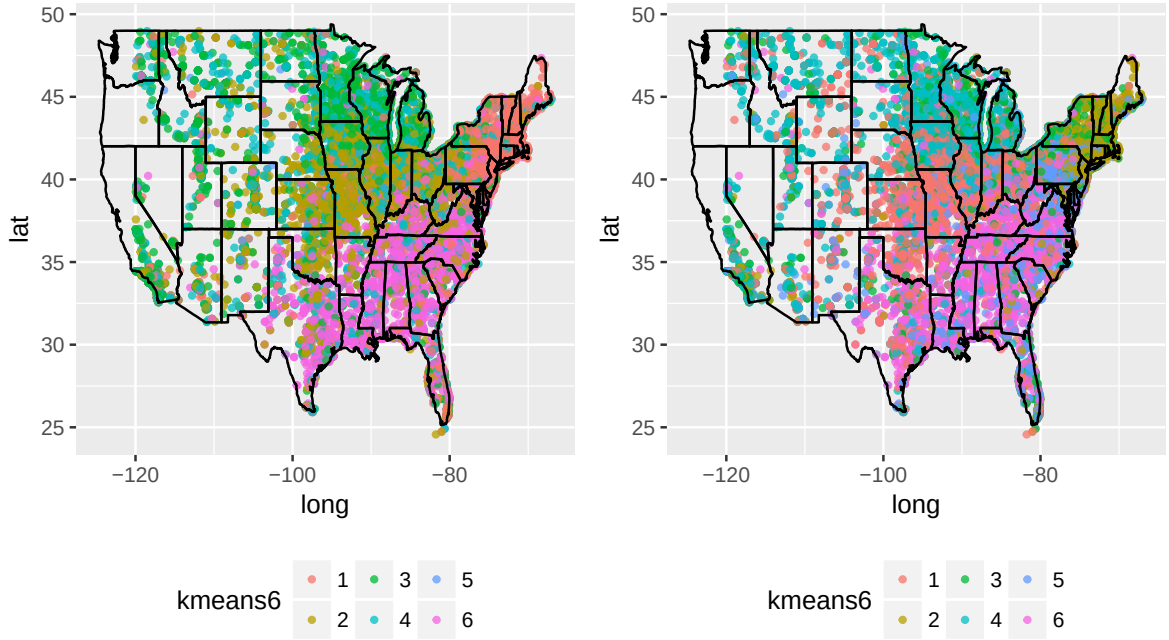

Figure 7: clustering comparison on map between 155 components and 2 components. Even though the kmeans method is unstable, those two both catch the geographic differences for dialects.

Another interesting finding came from clustering the lingLocation data set. I implemented the *pam* with manhattan distance, compared with the Euclid distance, and both of them worked well.

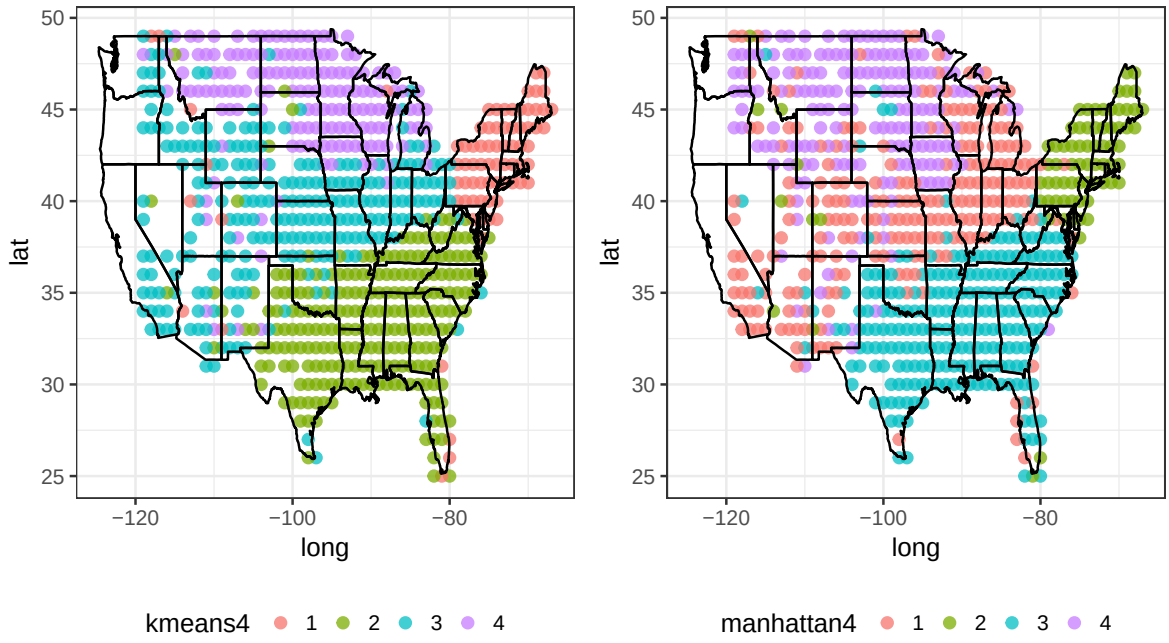

Figure 8: comparison between Manhattan and Euclid metric from the pam function. The four-cluster case is presented here and the clustering results are similar to each other.

In this part, I will answer the following question: what questions separate the groups? I tried several ways to answer it and I finally come to borrow the idea of the information theory(or decision tree) I assume that the four-group kmeans clustering results as the label of the data, and the 67 questions are the features. For each question, I looked the empirical mutual information of between the label and such question, that is

$$I(Question, Label) = \sum_x \sum_y p(x,y) log(\frac{p(x,y)}{p(x)p(y)}) = H(Label) - H(Label|Question)$$

Question	Highest Mutual Information	Question	Lowest Mutual Information
Q073	0.36	Q092	0.07
Q105	0.33	Q057	0.10
Q076	0.28	Q121	0.10
Q106	0.26	Q055	0.11

Table 1: Rank of questions by mutual information

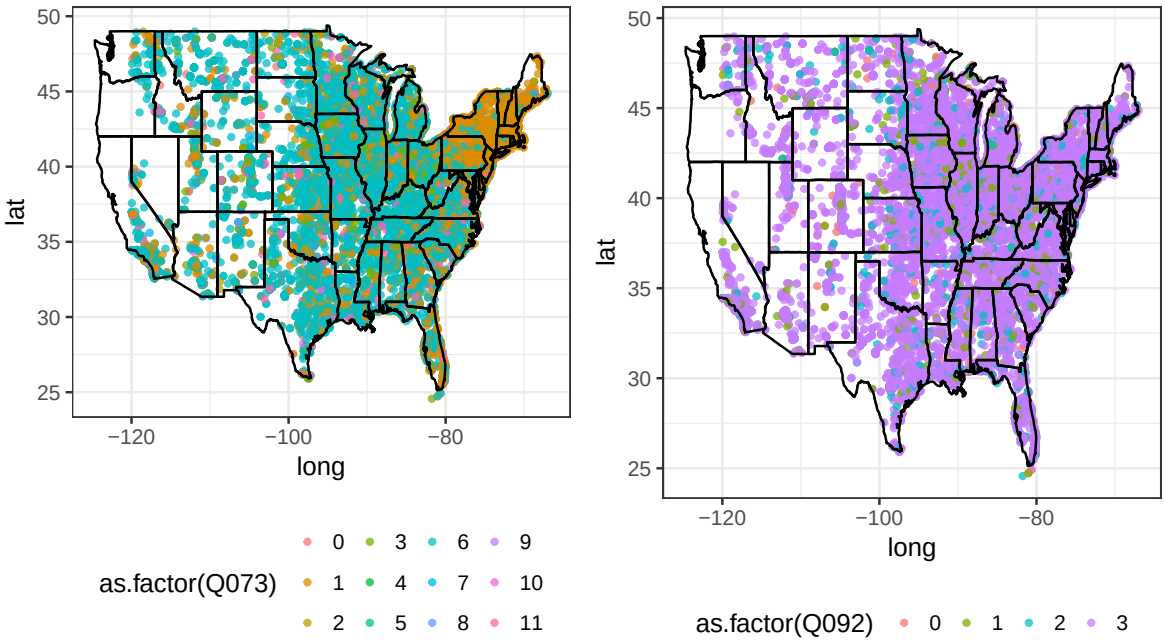

Figure 9: Answers from question 073 and 092. Left panel, Q073 gives a clear separation of dialect groups, while the right one, Q092 gives little information.

4 Stability of findings to perturbation

In this part, I added some random noise the scaled lingLocation data to see how sensitive is the pca-kmeans method. I tried the noise level of $0.001, 0.01, 0.1 and 1$ and turned out that the 0.01 case is not strong enough to make large difference to the clustering results. However, 0.1 case is enough to make the clustering totally a mess.

I also found the kmeans function is unstable with the results vary time to time. A better approach to solve this is to use a more stable function, for example, *pam* in the *clusting* package or *kmeanspp* in the *pracma* package.

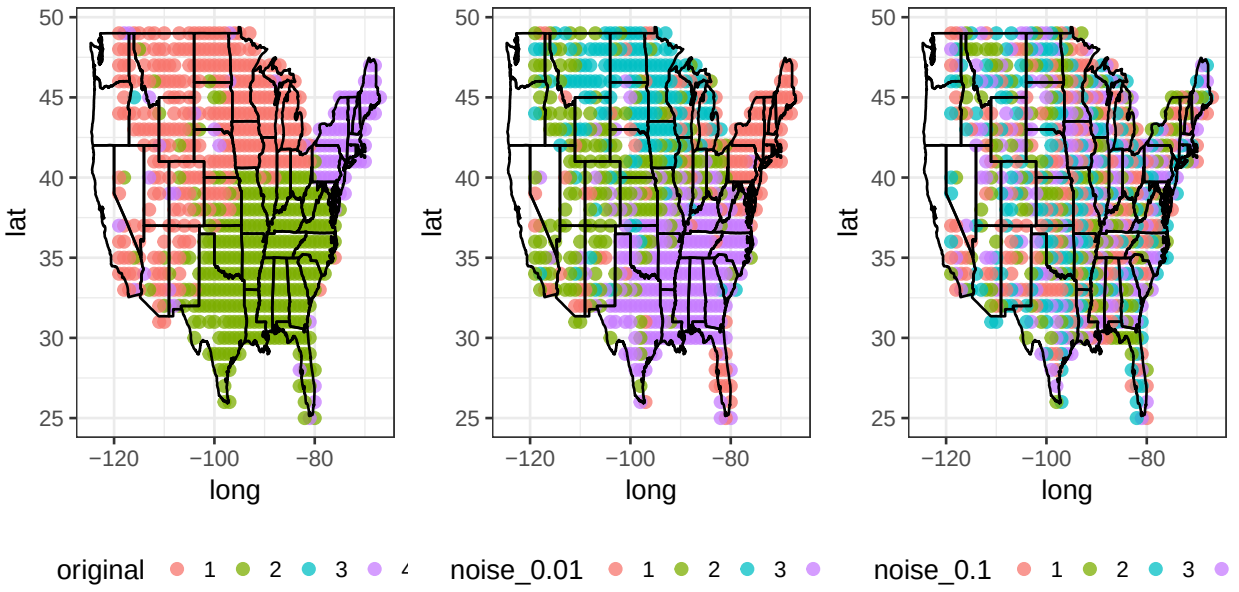

Figure 10: left panel is the original kmeans clustering; middle panel for noise of 0.01 level and right panel for the noise of 0.1 level.

5 Conclusion

Computational methods, which are applied to the linguistics survey dataset, reveals a surprising amount of information about the geographic differences of dialect differences. PCA and kmeans turn out to be very strong and meaningful to conduct dimension reduction and clustering, even though the later is lack of stability and sensitive to random noise to some extent. Other methods may make improvement, such as random projection or hierarchical clustering.

6 Kernel density plots and smoothing

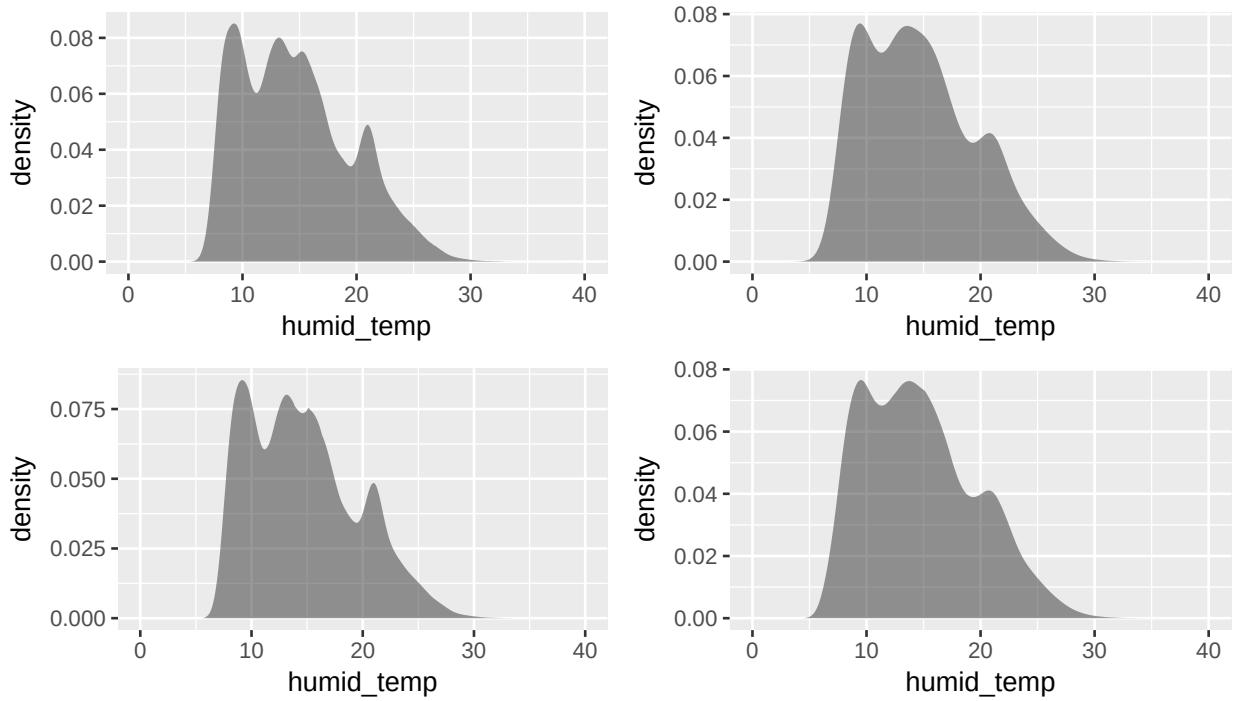

Figure 11: the histogram of temperature smoothed by different kernels and bandwidth. Left upper panel:Gaussian kernel with bw=0.5, upper right: kernel with bw=1; lower left: triangular bw=0.5; lower right: triangular bw=1.

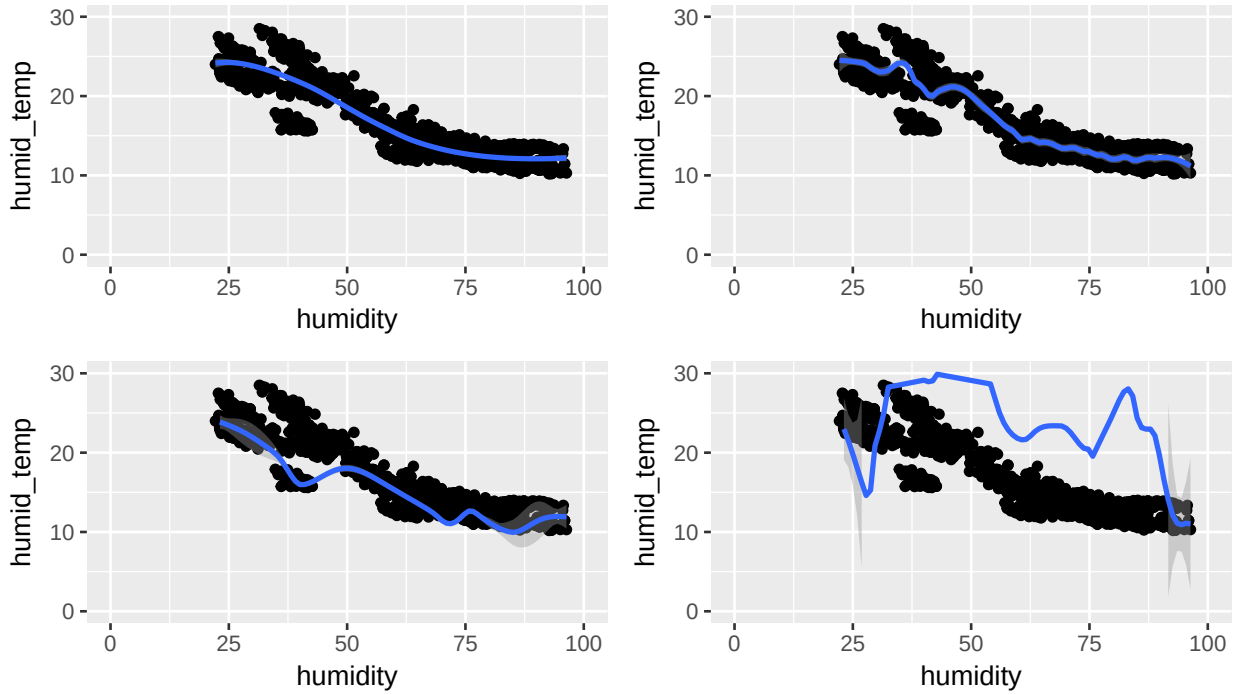

Figure 12: loess smoothing of temperature against humidity. Upper left panel: loess with formula y log(x): loess with formula y x and span 0.1 ; lower left: loess with the polynomial of two degree; lower right: loess with the polynomial of three degree. I found the log transformation in the upper left panel fitted really very well. However, the high-degree polynomials suffer a lot from over-fitting.

7 Appendix: coding by Python

The following figure*(see file .extra/silhouette.png) shows the silhouette plot to determine the best number of clusters. This part is performed by Python. Because of the page limit, I only give the case when $k = 4$.

The html file(see link www.soarya.org/interactive, if the link is lost, see file .extra/interactive.html) is an interactive plot to show the answers for question 73 and question 81. In this interactive plot, you can do:

- move the plot
- zoom the plot by selecting the wheel button on the left toolbar
- zoom the plot by selecting the cut-zoom button on the left toolbar
- hide/show the answers by clicking the legends on the bottom-right corner.