Lab 2 - Linguistics Data, Stat 215A, Fall 2020

October 08, 2020

1 Introduction

DIALECTOLOGY is the study of dialects and the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography. We want to learn the geographical characteristics of dialectology and make clustering to find some interesting regularities in dialects in US. We will focus on the questions that look at lexical differences as opposed to phonetic differences. First I did EDA to have a general impression of data and some potential geographical patterns. Then I used PCA to reduce dimensions and utilized K-means and NMF to cluster the data. This report concluded that the clusters of answers are related to geography and the segmentati_ _ of Census Regions defined by U.S. Bureau of Census.

2 The Data

This report uses linguistic data from a Dialect Survey conducted by Bert VauX (link). The questions and answers are found in the question_data.Rdata (this information was found and processed from dialect.redlog.net

We use two datasets: lingData and lingLocation data. lingData contains the answers to the questions for 47,471 respondents acroos the United States.The dataset contains ID of repondents, CITY, ZIP CODE, STATE, lat (latitude), long (longitude) as well as their answers for Q50 - Q121. CITY and STATE were self-reported by respondents. lat and long are the center of Zip Code, based on the reported city and state. question_data.RData contains information of questions.

2.1 Data Cleaning

I deleted 1020 missing data in latitude and longitude, and 3 missing data in state. The datapoints from AK and HI were also removed because of the scale limit of map. I also removed rows with wrong STATE abbreviations as well as the repondents who didn't answer any of 67 questions that we are interested in.

2.2 Exploratory Data Analysis

This section shows how the respondents and their answers are distributed across the country. It will helps us have a preliminary impression about the geographical charateristics behind the linguistic data. First I drew scatter plot of respondents on the map, colored by the state of the responents. We can see that most respondents are from the west coast and the east part of the country.

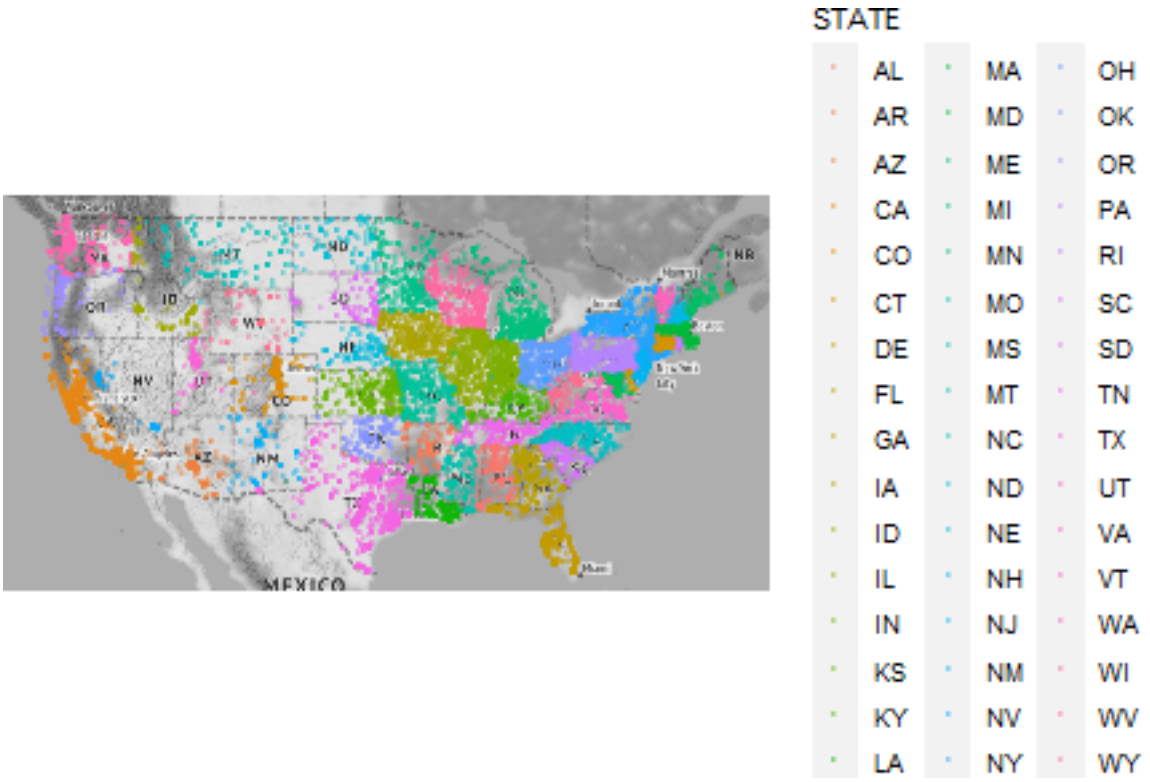


Figure 1: Location of Respondents

My first question to explore is Question 065: "What do you call the insect that flies around in the summer and has a rear section that glows in the dark?"

29% of respondents use "lightning bug", 30% of them use "firefly", and 40% of them use those two terms interchangeably. The rest 1% used other words or had not word for this. To explore whether there is a geographical characteristic in the answers, I drew a scatter plot map of the first three answers using "ggmap" method. We can see that respondents in the west of US use "firefly" or use "firefly" and "lightning bug" interchangebaly, and"firefly" dominates other choices in the westcoast. The pattern flips for the answers from the middle US to the east areas, where respondents use "lightning bug" or use it with"firefly" interchangebaly. While "lightning bug" doesn't dominate other answers in the east coast as it does in the west coast, people living near Great Lakes Region, northeastern and southeastern borders are more likely to use them interchangeably rather than only use "lightning bug".

Table 1: What do you call the insect that flies around in the summer and has a rear section that glows in the dark?

qnum	ans.let	per	ans
65	a	29.07	lightning bug
65	b	30.43	firefly
65	c	39.81	I use lightning bug and firefly interchangeably
65	d	0.02	peenie wallie
65	e	0.35	I have no word for this
65	f	0.32	other

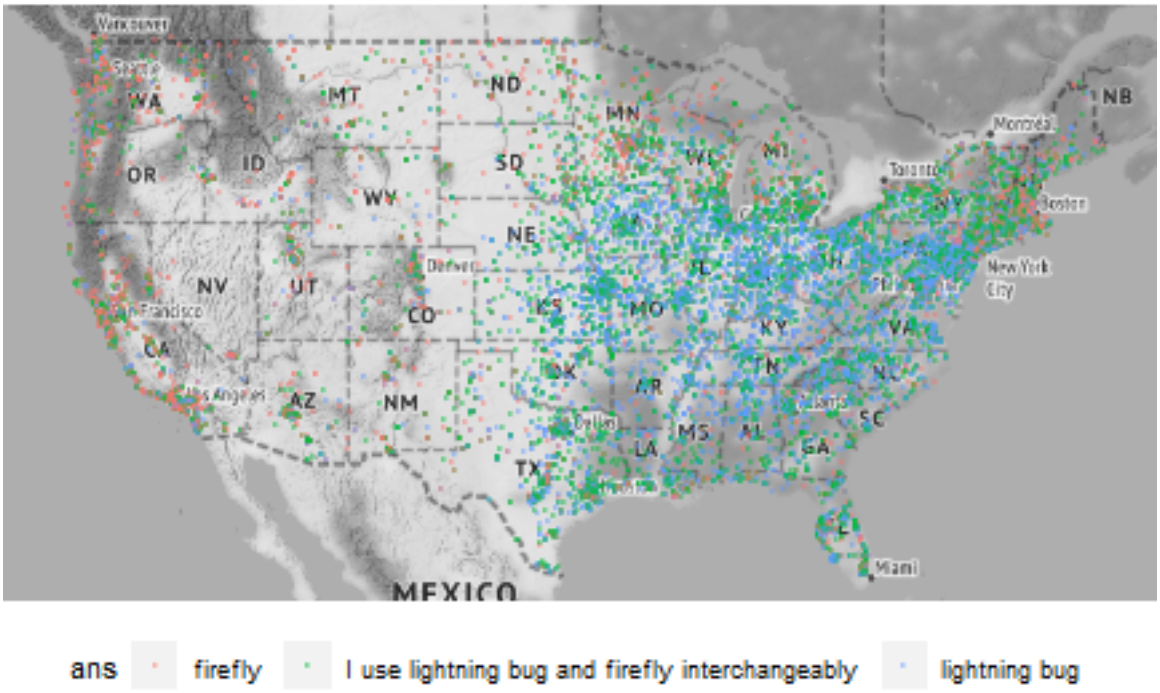


Figure 2: Map of Answers for Q65

My second question to analyze is Q097. The question asks the preference to describe trash cans. 35.5% of respondents prefer "trash can", 27.4% of them prefer "garbage can". 0.41% of them prefer "rubbish bin" and 1.06% of them prefer "waste (paper) basket". The rest 33.26% of respondens think these words refer to different things. From the scatter plot, we can see that respondents in the west of US think those words are different. "trash can" become much more popular in South Atlantic areas and around the border of West North Central, West South Central (those terms are census divisions defined by U.S. Bureau of Census; see Figure 5). And the answers from North Central, and East South Central mix up together, showing no significant clustering patterns.

Table 2: Which of these terms do you prefer?

qnum	ans.let	per	ans
97	a	35.53	trash can
97	b	27.38	garbage can
97	c	0.41	rubbish bin
97	d	1.06	waste(paper) basket
97	e	33.26	These words refer to different things
97	f	2.36	other

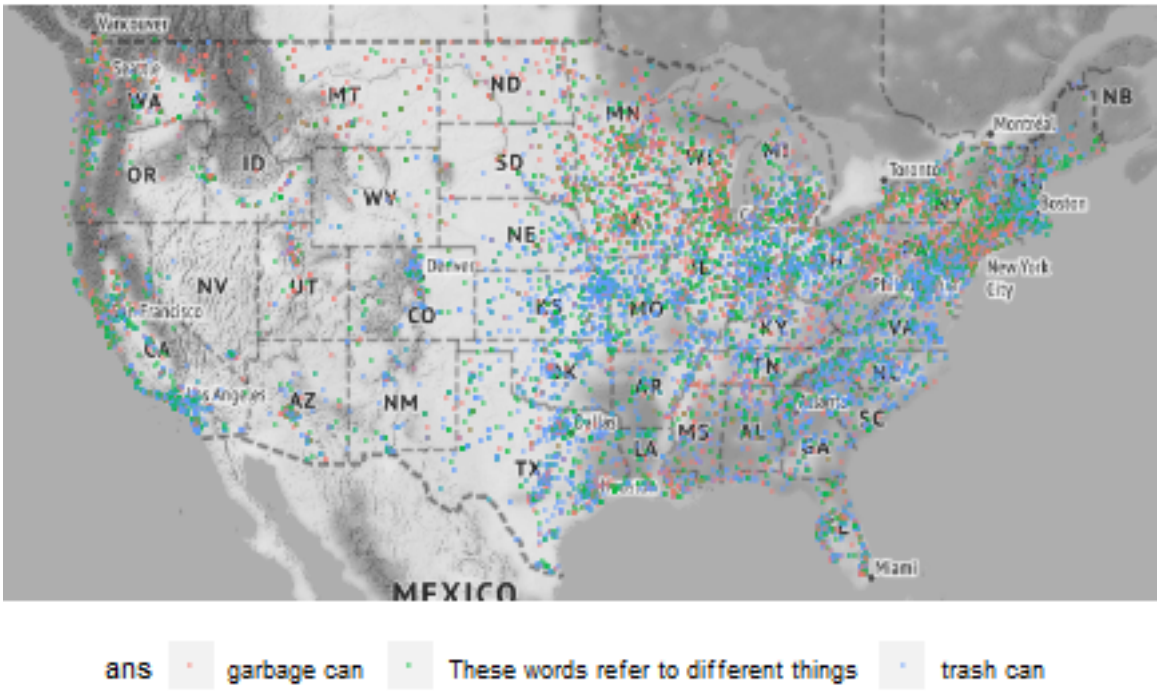


Figure 3: Map of answers for Q097

3 Dimension reduction methods

3.1 PCA

To reduce the dimensions of the data, I converted the dataset into binary format. To reduce the computation workload and leave some data for stability check, I sampled 70% of data as "training set" and the rest as "testing set". Then I ran Principal Component Analysis on training set. I centered the data but did not scale it. I don't want to scale it because there is no significantly large difference in variance of columns because they are binary. Even if PCA will prefer variable with large variance, I think that is beneficial for us because it helps us to find more important and representative questions and answers to catch lexical difference in different regions.

Now let's read the plot of the cumulative proportion of variance explained by PCs. We can see that the first 91 PCs explain 75% of variance.

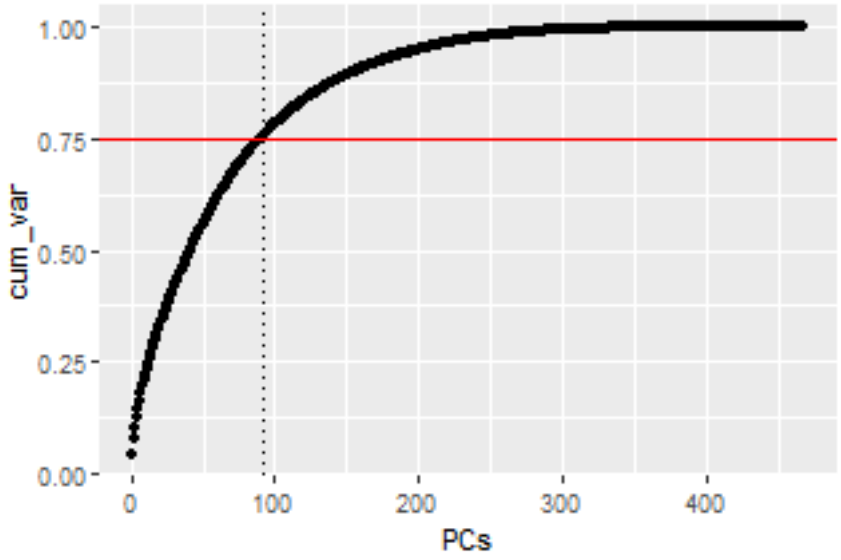


Figure 4: Cumulative Variance Explained by PCs

Then, in order to check whether the PCs are showing some geographic charateristics, I utilized the census divisions and regions in the U.S. (provided by U.S. Bureau of the Census) as a reference The dataset is downloaded from this github. As shown in the map below, the country is composed of four regions: West, South, Midwest and Northeast. And each region comprises several census divisions.

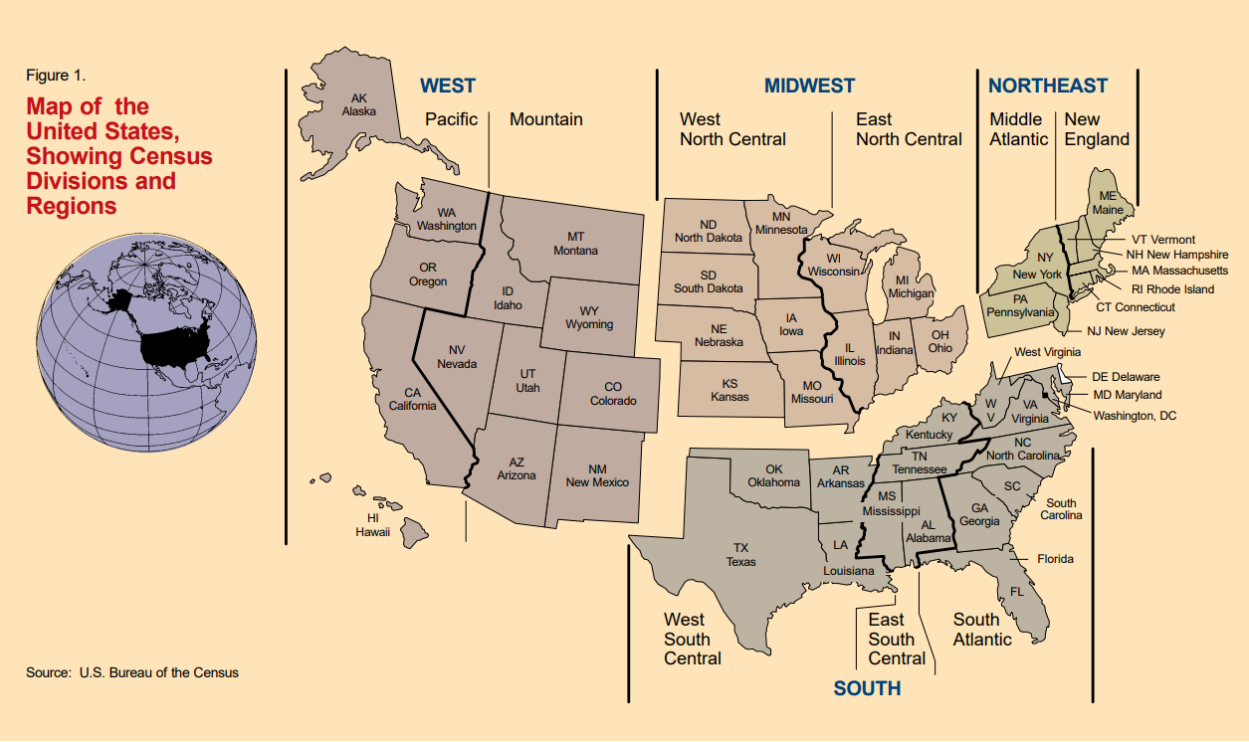


Figure 5: Map of the United States, Showing Census Divisions and Regions

The matrix of 5x5 pair plots is showing relationship between the first 5 PCs. Each point is colored by the regions. We can see that the pair plot of PC1 and PC2 is showing a good clutering pattern of four regions.

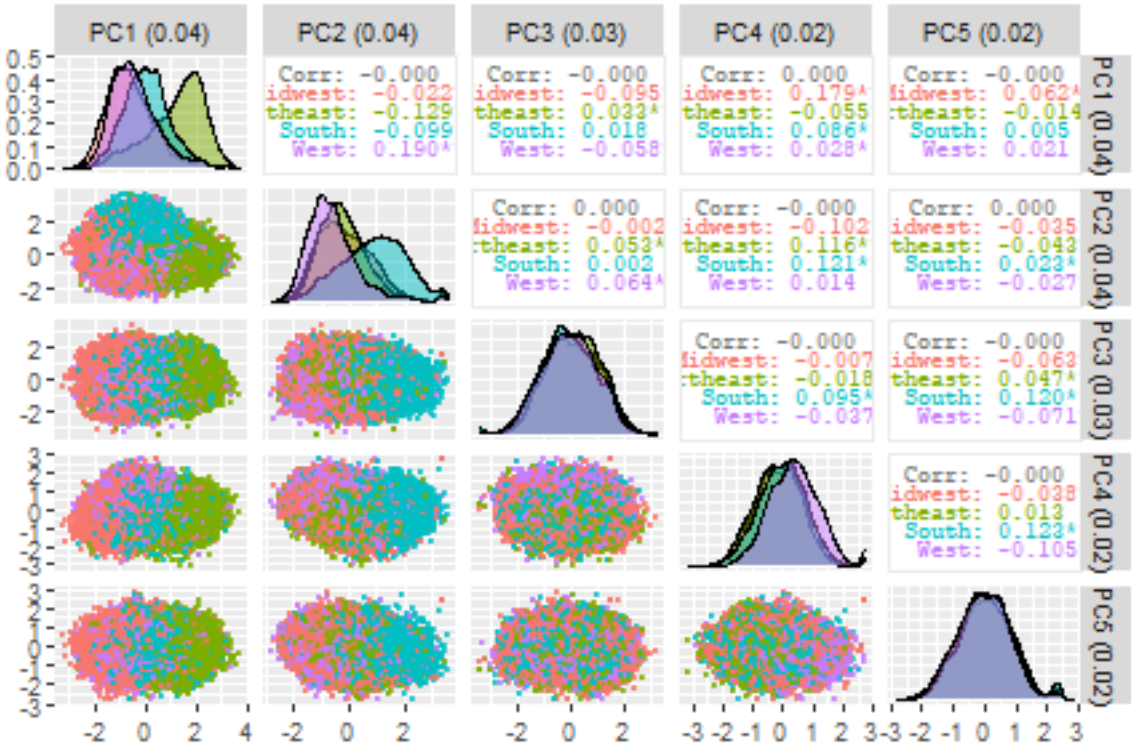


Figure 6: Paired Plot of Top 5 PCs

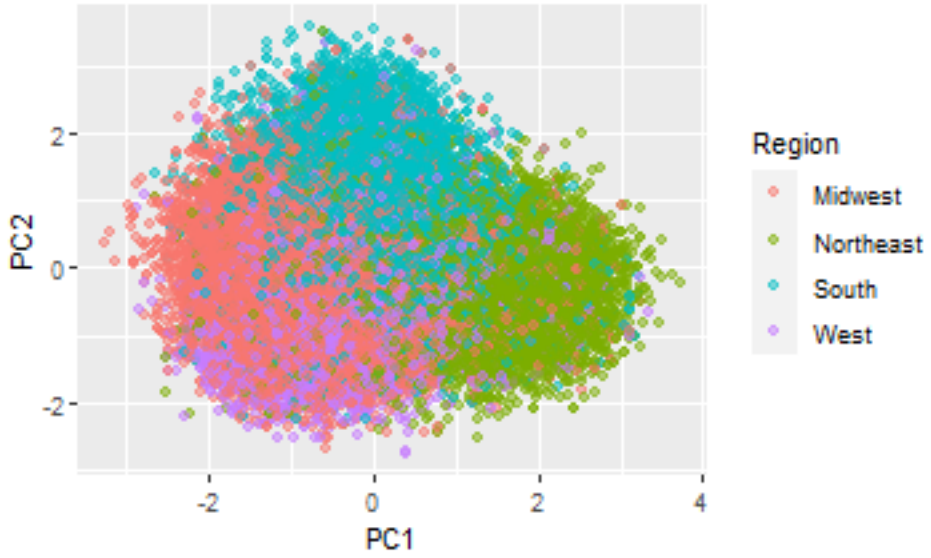


Figure 7: Scatter Plot of PC1 and PC2 (colored by Region)

Although PC1 and PC2 imply clustering about region, in each region there is no obvious clustering pattern

for different divisions, neither in other pair plots. This can imply that the linguistic difference in adjacent divisions in the same regions are not as huge as that in adjacent regions.

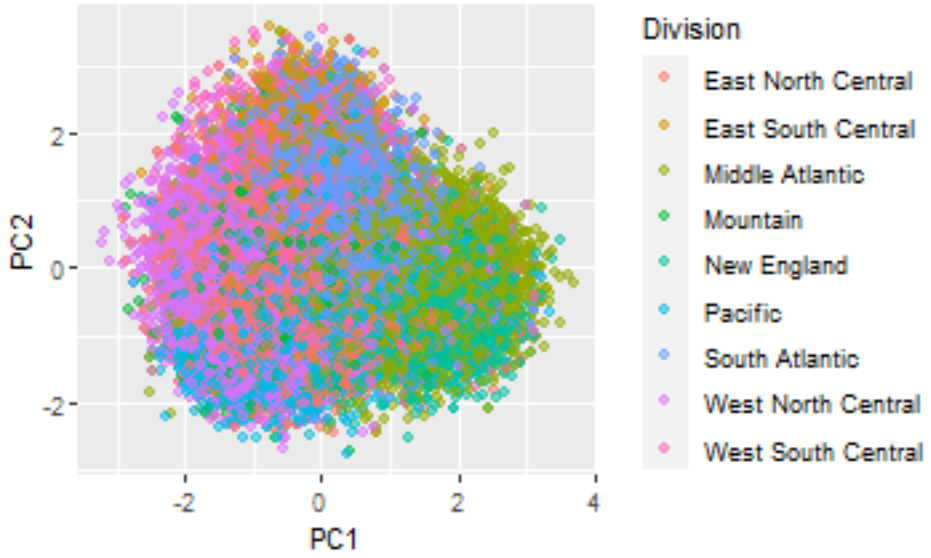


Figure 8: Scatter Plot of PC1 and PC2 (colored by Division)

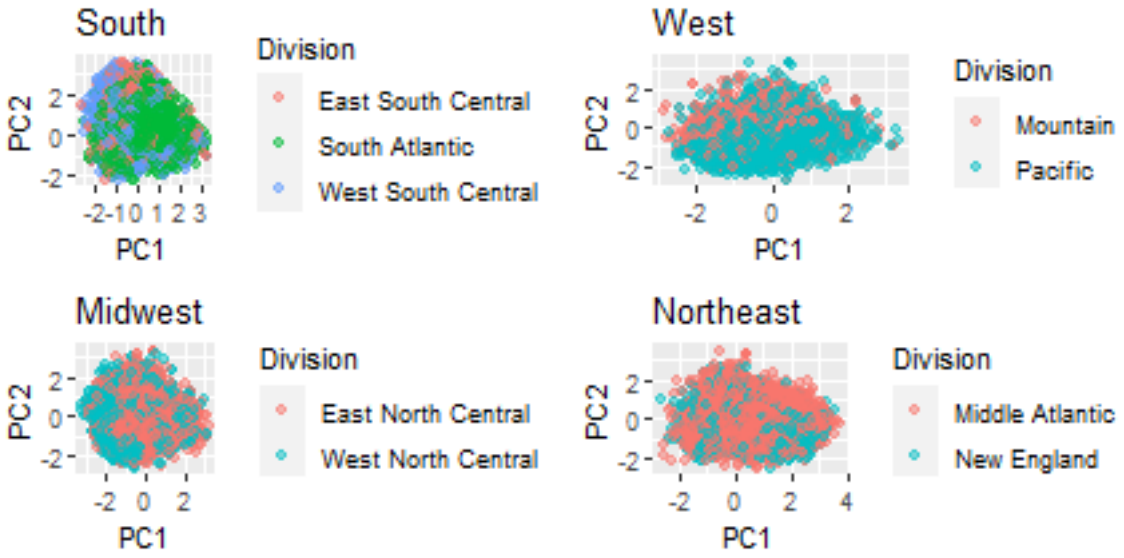


Figure 9: Scatter Plot of PC1 and PC2 in each Region (colored by Division)

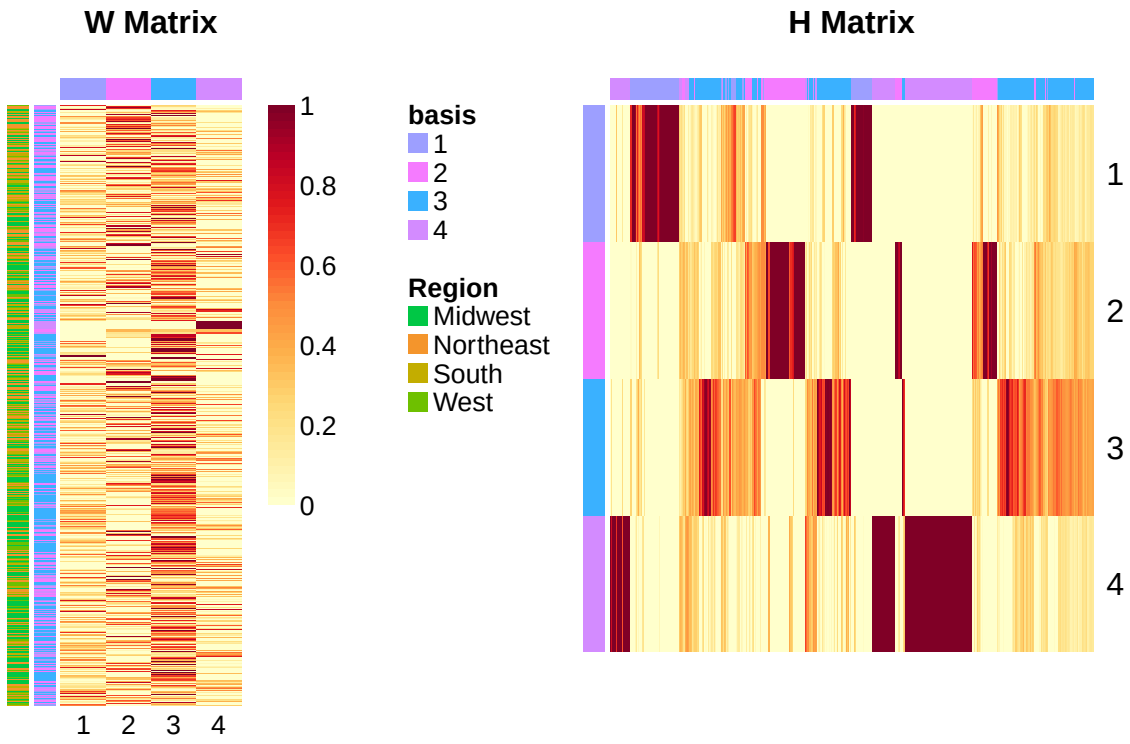
4 Clustering

Clustering on the whole binary dataset is not feasible because of the high dimension problem. Instead I clustered the data of with PCs. It is also a good chance for us to verify how well the PCs reflect geographical characteristics by comparing the result of the segmentation of census regions. I first used K-means method (with 4 centers) to cluster the dataset of the first 42 PCs, which explain 50% variance of data. Although

the clusterings are not completely separate, they work well and the scatter plot of clusters is very similar to those colored by the official census regions.

Figure 10: K-means clustering on PC1 and PC2

Then I clustered the data by NMF with 4 basis. When I plotted the NMF clusterings on axes of longtitude and latitude, despite there is still some overlapping in different groups, generally the clusterings are well separated from each other. It fits our previous observation from PCA analysis and K-means clustering that the answers to dialect questions show separated geographical patterns.



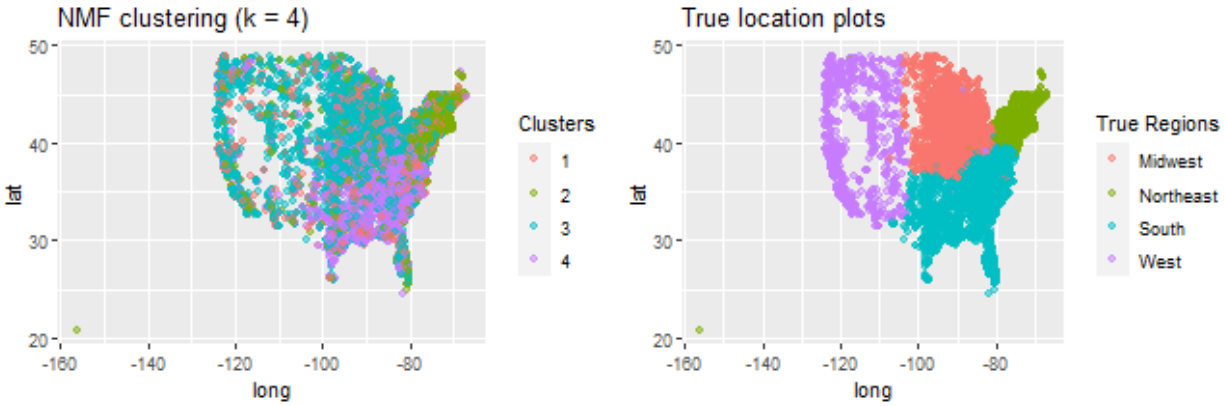


Figure 11: NMF Heatmap on Binary Data Matrix

5 Stability of findings to perturbation

In previous sections, we have observed that answers to dialect questions show separated geographical patterns, which highly fits census region defined by U.S. Bureau of Census. This conclusion is based on PCA analysis, K-means and NMF methods for the 70% randomly sampled data. Now I want to see whether the conclusion still holds on the rest 30% data. As is shown below, 75% of variance are explained by the first 90 PCs. PC1 and PC2 for different census regions are separated well. And the clusterings from K-means also imply the separated patterns of census regions.

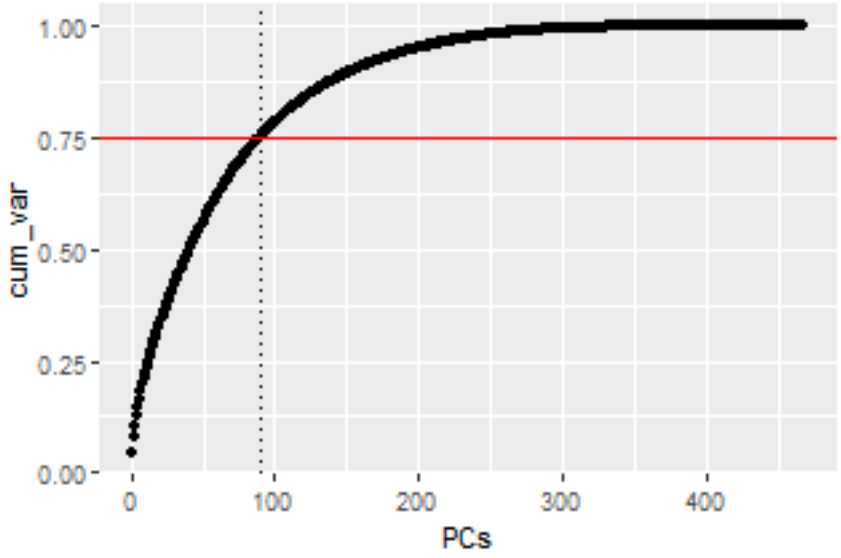


Figure 12: Cumulative Variance Explained by PCs on test dataset

Figure 13: PC Plot and K-Means clustering on test dataset

6 Conclusion

Based on the previous analysis, we can see that the answers to the dialect questions are showing different geographical charateristics. After clustering, the separation of different groups considerably fits the region segmentation defined by U.S. Bureau of Census. However, those groups are not completely separated. There is some continum on the edge of different groups. And the answers in different divisions in the same region are hard to cluster That implies that the lexical difference in neighbouring divisions in the same regions are not as huge as that in different regions.

I think the data and clustering in this project are useful for future decision makings, at least in the next couple of years, because language and dialects are changing slowly. But the data will be less valuable after several decades when words, culture and streams of information are highly likely to transform. And the clustering (related to geography) may be less reliable with the development urbanization and infomation technology, when people will be more and more connected with each other.

7 Academic Integrity Statement

By my honor, I affirm that I have acted with honesty, integrity, and respect for others; I have not been helped by anyone, nor have I helped anyone with this project.

8 Bibliography

[1] Nerbonne, J., & Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. Computers and the Humanities, 37(3), 245-255.

[2] Nerbonne, J., & Kretzschmar, . (2006). Progress in Dialectometry: Toward Explanation. In J. Nerbonne, & W. Kretzschmar, Jr. (Eds.), Progress in Dialectometry: Toward Explanation (pp. 387 - 398). Oxford-New York: Oxford University Press.