Lab 2 - Linguistics Data, Stat 215A, Fall 2021

Your Name

October 07, 2021

1 Introduction

Dialectometry as introduced by Nerbonne and Kretzschmar [1, 2] is the measurement of dialect differences whose distribution is determined primarily by geography. This geographical variation of language can be reflecting regional history, demography, economics, etc. However, collecting data for dialectometry research across geographical regions and the dialects can be a daunting task. The linguist Bert Vaux has there created a online questionnaire with 121 well-designed question to survey regional habits of pronunciation and wording in the United States, known as the Harvard Dialect Survey [3]. Such data can be very helpful for understanding dialectometry in the US but it comes with very large size. Therefore, compuational and statistical approach is needed to analyze such data. In this report, we will start by briefly describe the dataset, including data collection, data format and size. After data cleaning, we try to perform dimension reduction to the dataset and visually identify groups associated with geography. Then we explored two clustering methods trying to find clusters in the data. We studied the geographical distribution of the clusters and the which answers to the questions are the most characteristic features of the clusters. Finally, we evaluated the stability of the clustering results to different initializations and subsampling.

2 The Data

This dataset is from a web-based survey in which each respondent will answer 121 questions about dialectal habits. Here we focus on questions reflecting lexical (relating to vocabulary) differences rather than phonetic (relating to pronunciation) differences, which is question 50 to 121, excluding 108, 112, 113, 114 and 116. For example, question 61 asks about the word used for "area of grass that occurs in the middle of some streets". And the answers include "boulevard", "midway", "island", etc. These answers presumably reflect the different pools of vocabulary that people use in different geographic regions. For each respondent, the city, state and zip code are also recorded. In the given dataset, there is also the latitude and longitude of each sample based on the zip code, which are fetched by previous students in this course. The dataset has in total 47,471 samples, namely respondents. Each sample has 67 categorical variables representing the answers to the questions, where zero represents no response, and the 5 location variables as mentioned above.

2.1 Data Cleaning

We first examined if there is any missing values in the data and found that NA's only exist in the `ZIP`, `lat` and `long` columns. There are three samples with missing zip codes and 1020 samples with no coordinate information (which means the zip codes have no corresponding location information in the database). So we still retained all the samples for the following analysis and only exclude these samples when visualizing data points on the US map.

Next we calculated the response rate for each question and for each respondent (sample). We found that all questions have response rates over 90%. While there are 1040 respondent has no response to any of questions. We therefore excluded these samples from the analysis. Then we have 46,431 samples with each having 67 question variables for the following analysis.

2.2 Exploratory Data Analysis

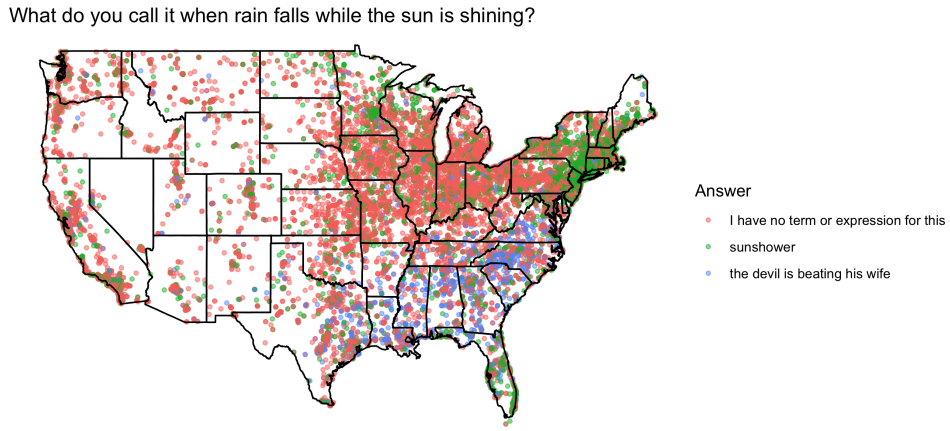

Figure 1: Geographical distribution of answers to question 80

Here we studied the question 80 and question 105. Question 80 asks about the word to describe the phenomenon of raining when sunny. In Fig. 1, we pick three major answers to this question and show their distribution on the US map. Here we found a clear geographic pattern. Respondents from the northeast tend to use the word "sunshower" to describe this phenomenon. Those from the southeast mostly use the expression "the devil is beating his wife". And those from other parts of the US do not have a specific word to use. In Fig. 2, we did the same visualization for question 105. It asks about the generic term used to describe sweetened carbonated beverage. A clear spatial pattern can also be observed. Respondent from the northeast and the west coast tend to use "soda", and those from the south tend to use "coke". And people from other regions uses "pop" mostly.

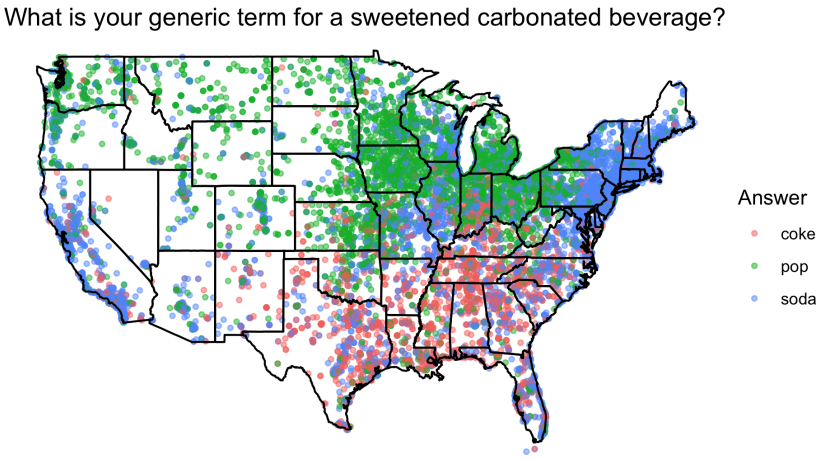

Figure 2: Geographical distribution of answers to question 105

We further studied the association of this two questions. In line with what we observed on the map, we can observe in Fig. 3 that the most of the respondents who use "sunshower" to describe the rain when sunny also use "soda" to describe carbonated drink, which corresponds to the northeast region of the US. And the majority of those who use "the devil is beating his wife" to describe the rain when sunny also use "coke" as a generic word for carbonated drink. And this corresponds to the southeast region of the US. This exploration indicate that at least some question questions do reflect regional difference of the vocabulary. Hence, we further study this variation with more features using dimension reduction and clustering methods.

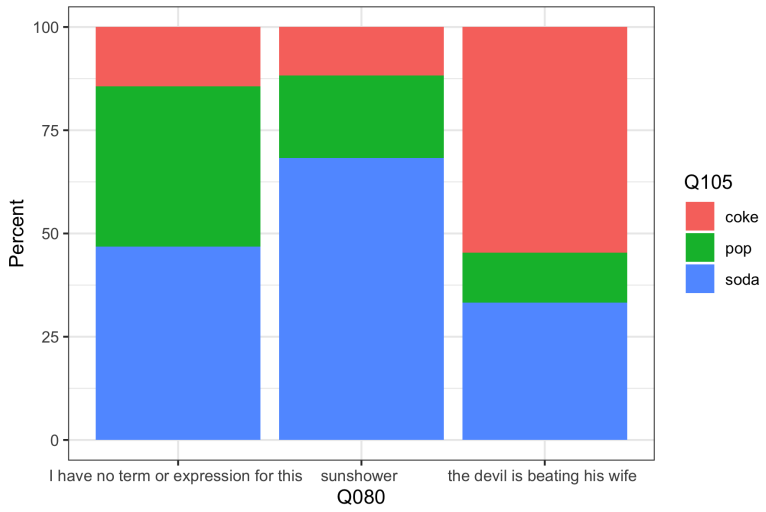

Figure 3: Relationship between question 80 and 105

3 Dimension reduction methods

We first one-hot-encoded the data as suggested by the instruction. This is a commonly used encoding method for categorical data. Why don't we just use the original `lingData`? The answers for each question do not have a natural order. But the ordinal encoding in `lingData` by `0, 1, 2, ...` creates an non-existing ordinal relationship between the answers (i.e. answers will not equivalently treated in the model we put it into). While in one-hot encoding, each answer is one binary vector with no ordinal relationship. Therefore we do not introduce this artificial bias. After encoding, we have a 46,431 by 468 data matrix.

Here we mainly explored principal component analysis (PCA) as the dimensional reduction method. Our assumption is that there is dialectometry (i.e. geographical differences in the dialect) in the dataset. So we would like to check if there are any visual groups by geography on the reduced space. We grouped the samples by four geographical regions according to the `state` dataset in base R: Northeast, South, North Central and West, based on the state from which each sample was collected. There are a small fraction of samples (327) with erroneous state name and we just disregard them (shown as `NULL` in the legend).

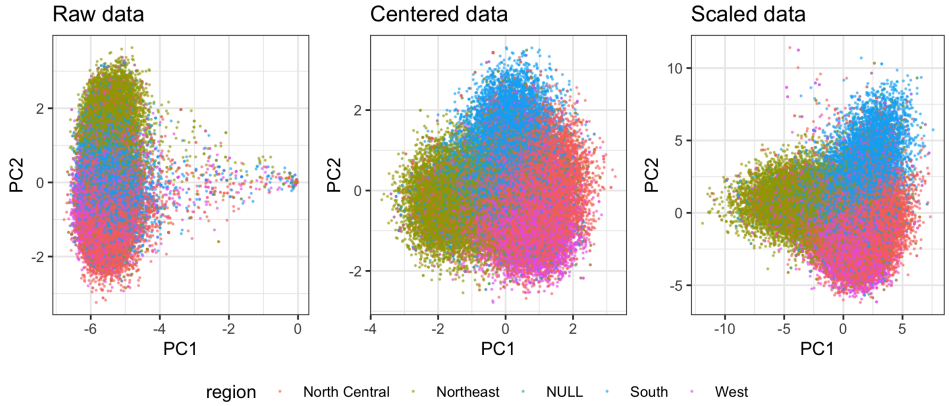

Figure 4: Principal component analysis with raw, centered and scaled data

First we want to see if centering and scaling of the data would help with PCA. We embedded the data points into the space spanned by the first two principal components (PC) and colored them by region. As shown in the left plot of Fig. 4, first we can see that the first PC is dominated by a group of outliers and the majority

of the data points have negative values in PC1. This is expected because PCA can be regarded as a regression model without an intercept. If the data is not centered, the first PC will still be forced to go through the origin, thus its direction will not capture the largest variation in the data. Therefore in the R `prcomp` function, the data will first be centered by default. And although the separation are not entirely clear, we can already see the data points are roughly grouped in terms of the regions. Then we can see that centering greatly improves the visualization (middle plot). We can see visual clusters of data points in approximately round shape and all the points are centered around the origin. The regional groups become quite clear on the PCA plot. Finally on the PCA plot with the scaled data on the right side, we can further see an slightly clearer separation between the regional groups of samples. Scaling makes the standard deviations of the features uniform and therefore PCA will more accurately capture the variation of each feature rather than the scale. Thus, we used the scaled data for our analysis.

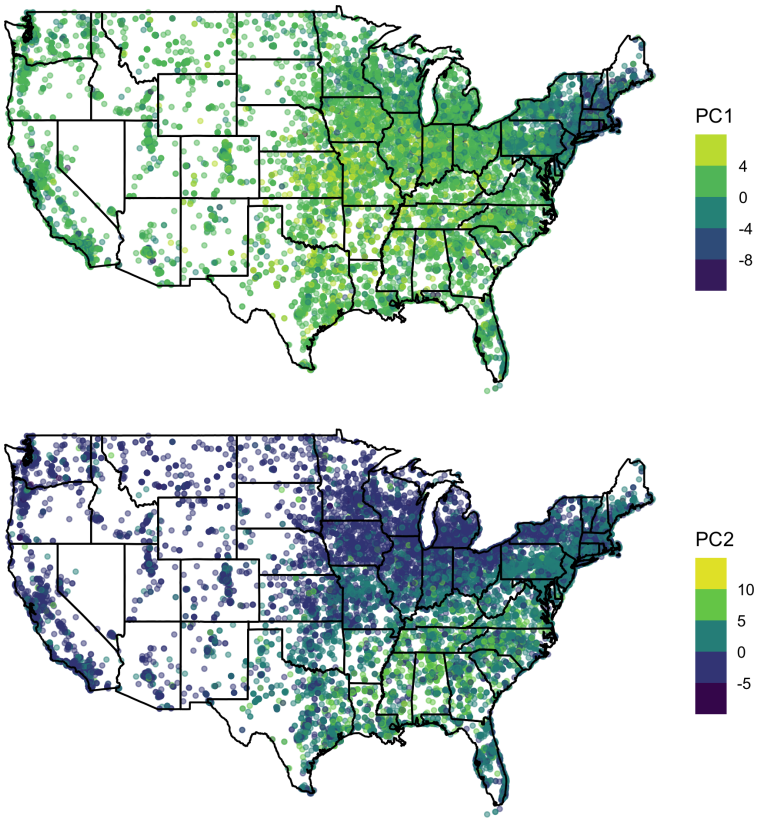

Figure 5: Projecting PC1 and PC2 onto the US map

To further interpret the PCA results, we investigated how the PCs related to geography by projecting them onto the US map. In Fig. 5, we see that PC1 mainly separates samples from the northeast to the other regions, showing only low values on the northeast part of the US and generally high values on the other parts. For PC2, we see a clear variation from the northwest to southeast. Samples in the northwest half of the US has relative high PC2 values and the other half has lower values.

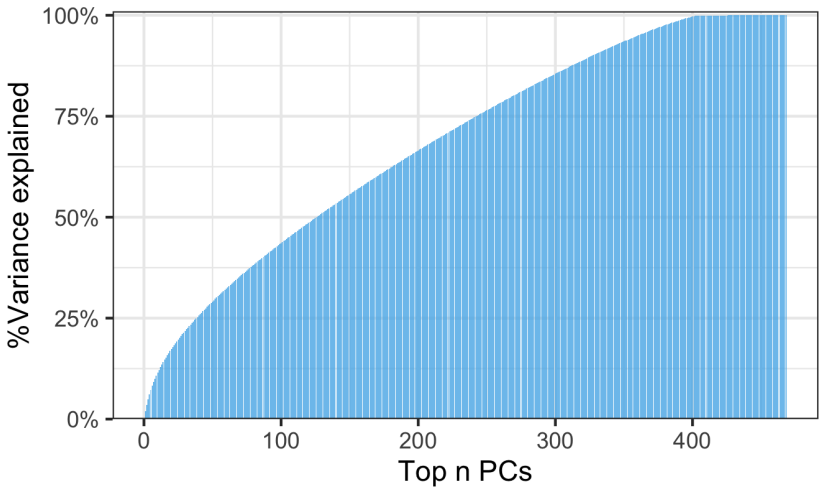

Figure 6: Cumulative percent variance explained by PCs

Perform clustering on the PCA reduced data can reduce computing time, deals with colinearity in the data and remove noise (if we consider small PCs to be predominantly noise). We thus finally investigated the proportion of variance explained by the PCs. As shown in Fig.6, we can see that the top PCs do not explain a high proportion of the total variance in the data. The first two PCs only account for 1.9% and 1.6% of the variance, respectively. And it takes the top 126 PCs to cover 50% of the total variance. Hence, there is not a clear indication of how many PCs we should use for clustering. We therefore further experiment with the cut-offs in the following section.

4 Clustering

Our goal in this section is to identify clusters of the samples and hopefully find relationship between the clusters to geographical regions. We explored K-means and non-negative matrix factorization (NMF) as clustering methods.

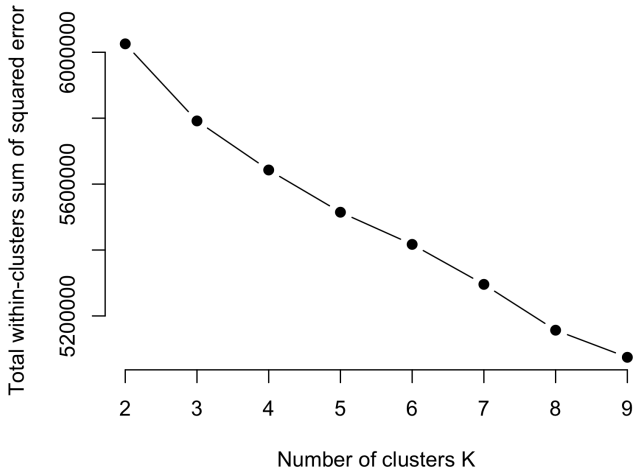

Figure 7: Elbow plot for K-means clustering

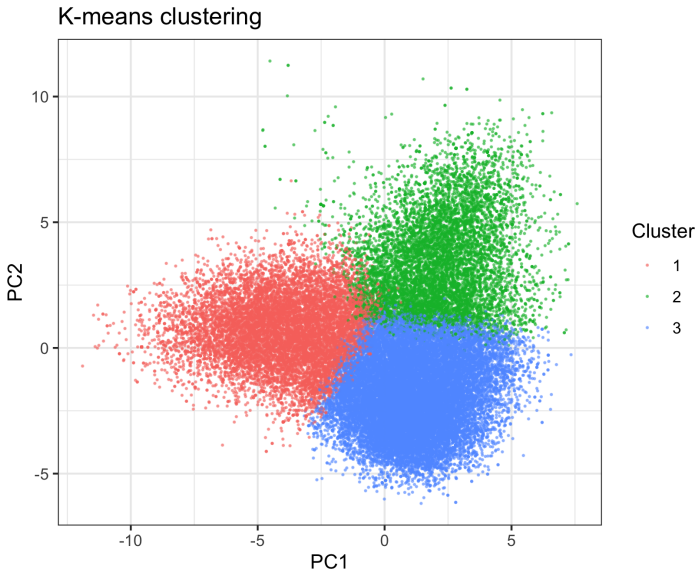

Figure 8: K-means clustering results on PCA plot

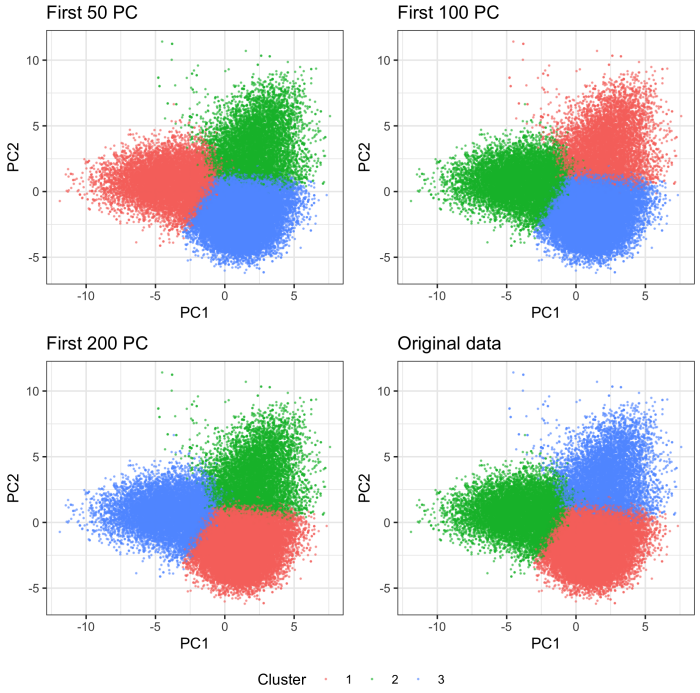

Figure 9: K-means clustering result with different PCs used

For k-means clustering, we need to specify the hyperparameter k which corresponds to the number of clusters to find. Here we search for the optimal k by plotting the total within cluster loss for each k, namely the elbow plot. The United States can be roughly divided into 4 regions (mentioned before) and 9 divisions (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain and Pacific) according to the `state` dataset. We therefore choose to search k from 2 to 9. We used the `kmeans` function in base R to perform clustering using the first 50 PCs. Noted that in our analysis, we always set `nstart` argument to be 10 to reduce the instability of results caused by bad initialization. This essentially runs the algorithm each time with 10 random initial sets of centroids and pick

the one result with the minimal loss. As shown in Fig. 7, we can see that the first elbow appears at k=3 on the plot. In Fig. 8, we see that the three clusters are clearly separated on the space spanned by the first to PC. So we choose k to be 3 for the following analysis.

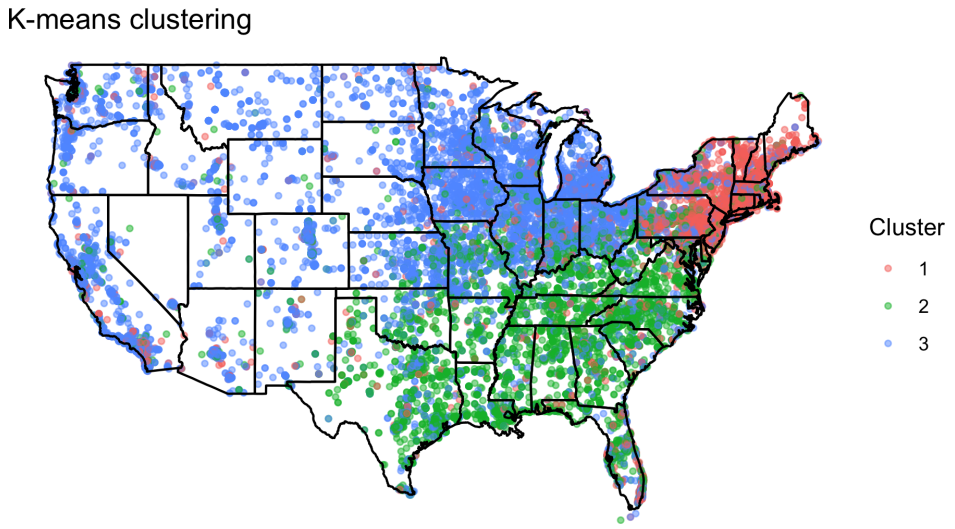

Figure 10: Geographical distribution of K-means clusters

We then examined how the number of PCs used affects the clustering result. We used the first 100 PCs, first 200 PCs and the scaled original data to perform the clustering. As shown in Fig. 9, we can see the cluster we obtained are all very similar. We use the Adjusted Rand Index (ARI) to measure quantitatively the similarity between the clustering results. ARI ranges from -0.5 to 1, where 0 implies random permutation and 1 implies identical clusters. Comparing the results with first 100 PCs, first 200 PCs and full dataset to the original result, we get ARI 0.988, 0.983 and 0.979, respectively. This shows that the number of PCs to use does not have a noticeable influence on the clustering result. We therefore continue to use the first 50 PCs in order to save computing time.

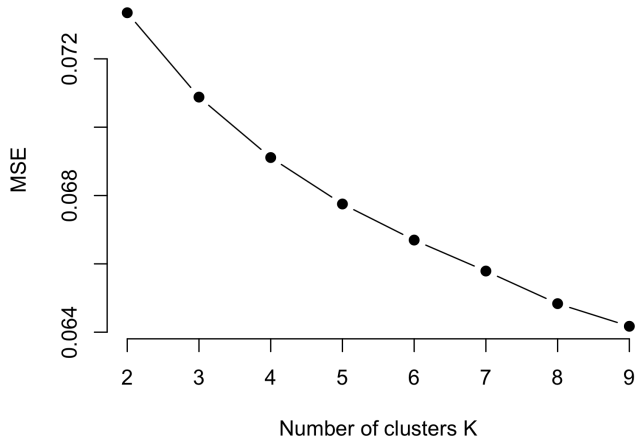

Figure 11: Elbow plot for NMF

Next we visualize the clusters on the US map. We can see that three clusters clearly correspond to the northeast, southeast and the west/central regions, respectively. If we group the samples into these three geographical regions, we can compare this grouping to the clusters we obtained. This results in an ARI of 0.36. This shows that we can indeed cluster the samples that reflects geographical dialect differences.

We next tried non-negative matrix factorization. We used the `nmf` function in `RcppML` package, which is optimized for sparse matrices (our data has ~85% zeros). Similar to k-means, we made an elbow plot using the mean square error (MSE) of the matrix factorization. We again found the first elbow to appear at k=3. By assigning each sample to the factor with the highest value, we obtain the cluster memberships. As shown on the PCA plot and the US map (Fig. 12, 13), we found that the clusters are visually very similar to those obtained by k-means. Calculating the Adjusted Rand Index (ARI), we found that the two cluster separations are indeed similar with ARI = 0.55.

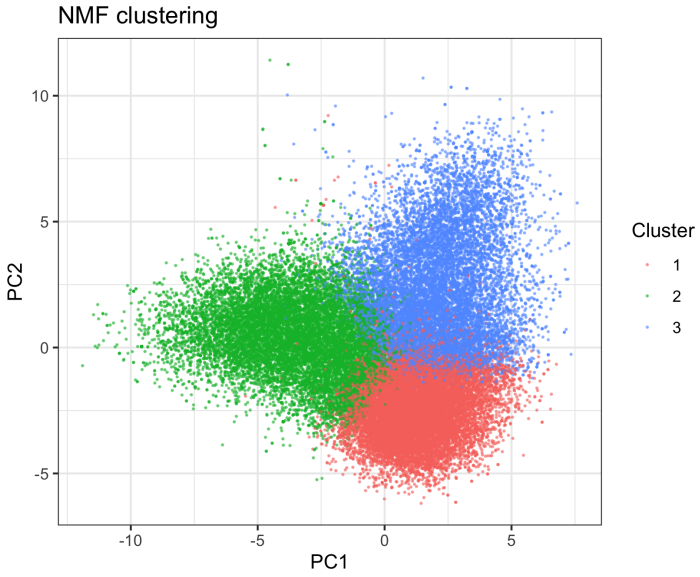

Figure 12: NMF clustering results on PCA plot

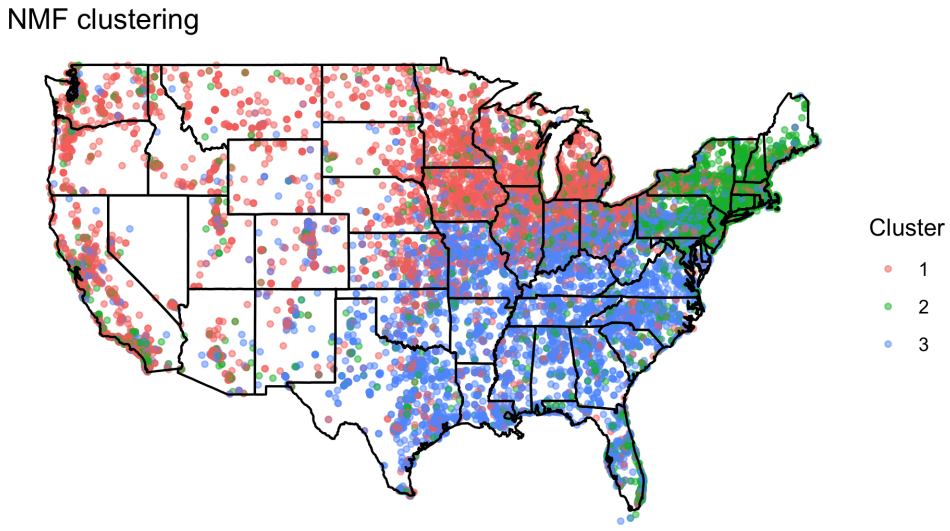

Figure 13: Geographical distribution of NMF clusters

One advantage of NMF is that we can conveniently interpret the clustering result with the fitted model. For example, we can study the dominating features in each factor by looking at the feature matrix W. In the heatmap showing the W matrix (Fig. 14), we can clearly see that there are only a fraction of entries have distinct high values, which suggests these features (i.e. answers) has the most prominent regional differences. We therefore find the 10 answers that have the highest determining effect for each cluster (namely the entries

with the highest value in each factor). The result is shown in Table. 1. For instance, giving the second answer to question 53 is the dominating feature of the respondents in cluster 1, which corresponds to the west and north region. And giving the second answer for question 54 is the dominating feature of the respondents in cluster 2, which corresponds to the northeast region.

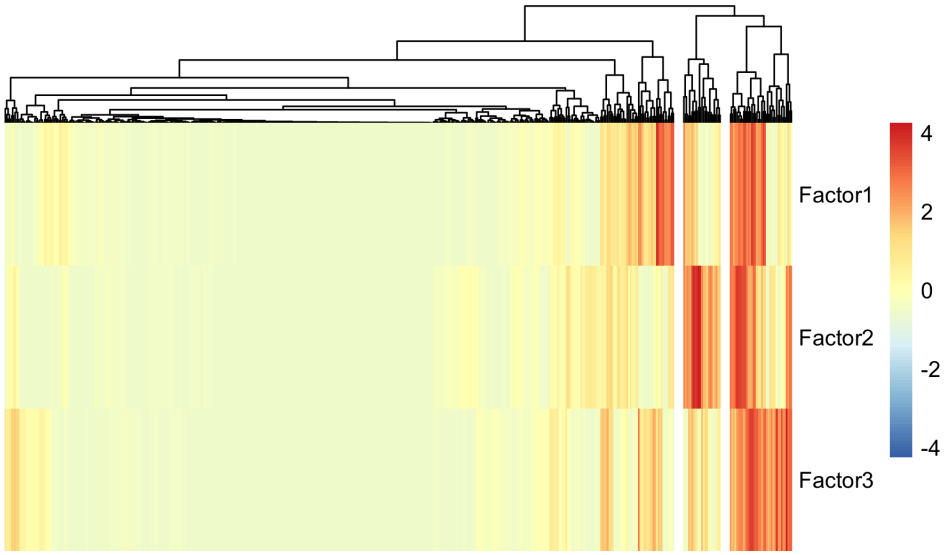

Figure 14: Feature matrix H from NMF

Table 1: Top 10 features of each cluster/factor

Cluster1	Cluster2	Cluster3
Q053_2	Q054_2	Q051_2
Q063_1	Q055_2	Q062_2
Q064_1	Q056_2	Q063_1
Q067_1	Q067_1	Q076_4
Q068_3	Q073_1	Q077_1
Q076_1	Q091_2	Q089_1
Q080_8	Q098_1	Q093_2
Q081_1	Q105_1	Q103_4
Q093_2	Q109_1	Q104_1
Q106_1	Q119_1	Q120_2

5 Stability of findings to perturbation

We checked the stability of our k-means clustering results in two ways. First, for k from 2 to 5, we run the k-means for four times with different initial state and compared whether stable clusters can be obtained. As we can see in Fig. 15, with k = 2 or 3, the clustering results are fairly stable. While with k = 4 or 5, the clusters start to vary across replications. This again supports our choice of k to be 3 in the analysis.

We next examined how stable the clustering result is to subsampling. We randomly subsampled 90%, 60% and 30% of the data 20 times and performed k-means clustering on the 60 subsampled datasets. We then compare the results to the clusters we obtained with all the samples using ARI. It turns out that with 90% subsampled data, the mean ARI of the clustering is 0.99 across the 20 replicates. For the 60% and 30% subsampled data, mean ARIs are 0.98 and 0.96, respectively. Therefore we can conclude that our clustering result is stable across subsamples.

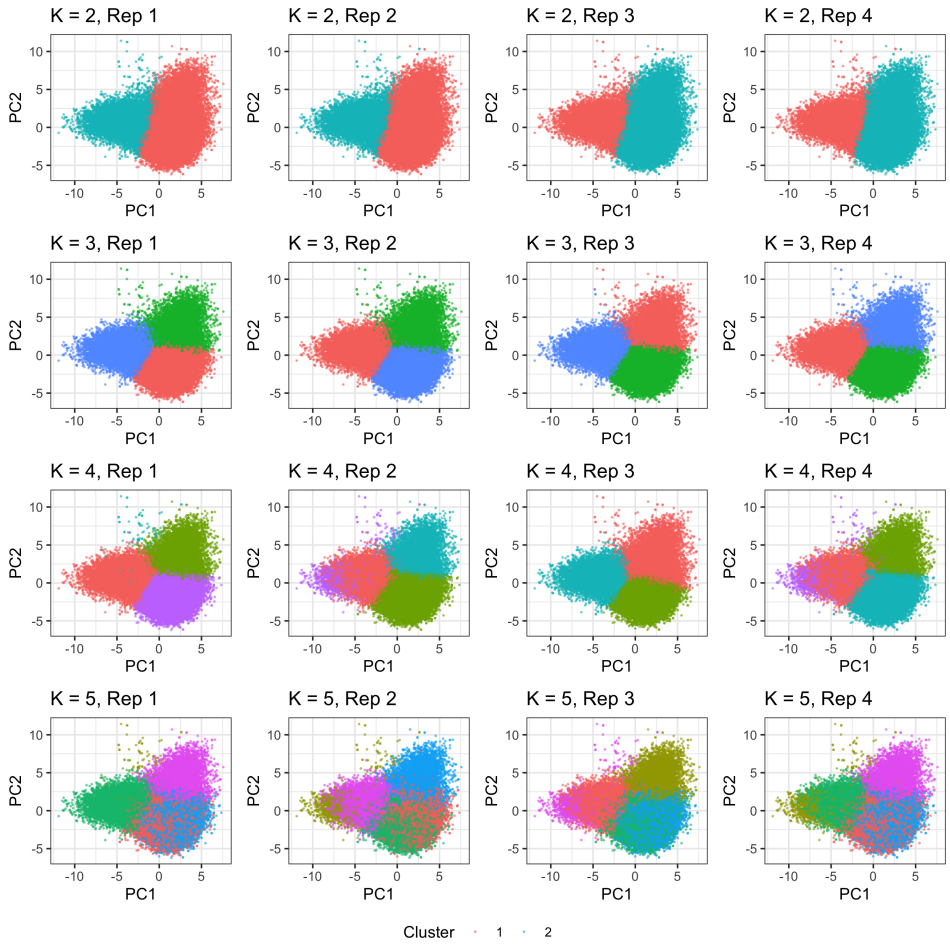

Figure 15: Stability of K-means clustering across different initializations

6 Conclusion

In this data analysis workflow, the reality we want to study is the dialectometry, i.e. linguistic differences whose distribution is determined primarily by geography, of the US. And the data is supposed to reflect this reality, which are the answers to 67 dialectology questions designed by expert linguists from 47,471 people across the US. We consider it to be a good representation of the reality. We used dimension reduction methods and clustering methods to analyze this dataset. We learnt that based on the answers, the samples can roughly divide into three groups. We verified if these groups have geographical correspondence and found each group is clearly associated with different geographical regions of the US.

For reality check, the most reliable way is to perform external validation. We may run another survey on a different set of dialectology questions and see if the answers show clear differences across the three clusters of regions we obtained.

Choosing the number of clusters k is an important yet difficult task in cluster analysis. In our analysis, we found that the stability of clusters could be a very important aspect of choosing the optimal k. The elbow plots in this report did not show very clear elbows. But in the stability analysis, we found the clusters to be very unstable when we tried to cluster with k greater than 3. And this gives a strong evidence of choosing k to be 3. Given more time, I would probably experiment with more dimension reduction methods as PCA may not be the optimal choice for our one-hot encoded binary matrix. Also we could run stability check in a more quantitative manner for model selection.

7 Academic Integrity Statement

Here I make the truthful statements: I myself designed and performed all the data analysis procedures presented in this report. I myself wrote all the texts and produced all the figures in this report. I have included and documented all the procedures in the workflow and the results can be fully reproduced. Wherever I included the work from others, I cited the sources.

I think academic honesty is an essential prerequisite for conducting any form of research. First, it is important that we take responsibility for our research results. Because science is built upon collaboration. Your results will be the foundation of other people's research. A dishonest or unreproducible research will possibly lead to a cascade of false results. Also, it is important that your work is original. Plagiarizing is not contributing anything new to the scientific community and it is disrespectful to the original author as you are taking their credit. Therefore, we should always keep our research honest, transparent and reproducible.

8 References

1. Nerbonne, J., & Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. Computers and the Humanities, 37(3), 245-255.
2. Nerbonne, J., & Kretzschmar, W. (2006). Progress in dialectometry: Toward explanation. Literary and Linguistic Computing, 21(4), 387-397.
3. Dialect Survey. http://dialect.redlog.net/index.html