Lab 2 - Linguistic Survey Stat 215A, Fall 2017

3031884142

10/5/2017

1 Introduction

In this report, we first apply the density estimation and loess smoothing to the redwood data[1] from the previous lab and then we analyze the linguistic data of a 2003 Harvard Dialect Survey[2, 3]. For the Harvard Dialect Survey we discuss the issues of data cleaning, binarization and aggregation; then we use Principal Component Analysis (PCA), Multidimensional Scaling (MDS), k-means and k-medoids to perform dimension reduction and clustering; in the last part, we perturbed the answers and the questions respectively to check the robustness of an interesting finding : the 4 linguistic cluster coincide with the 4 regions defined by the US Census Bureau.

2 Kernal density plots and smoothing

The data we are using here is a sample of size 10000 from the cleaned redwood data.

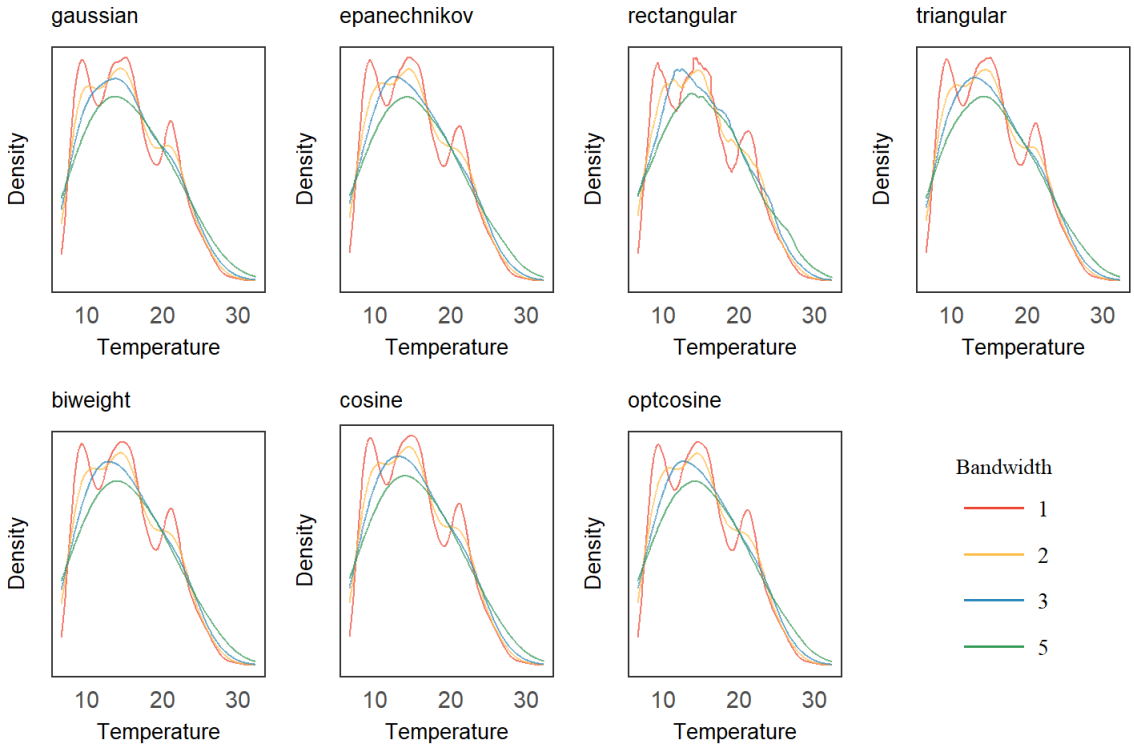

Figure 1: Density Estimation of Temperature with Different Bandwidth and Different Kernals

In Figure 1, we use 7 kernel functions: Gaussian kernal, Epanechnikov kernal, rectangular kernal, triangular kernel, biweight kernel, cosine kernal and optcosine kernel. The bandwidth is chosen to be $h, 2h, 3h$ and $5h$, where h is a rule-of-thumb baseline bandwidth defined as $0.9 \min\{\hat{\sigma}, \hat{q}/1.34\} n^{-\frac{1}{5}}$ where q is the interquartile. It turns out that a larger bandwidth produces a smoother density estimation, and vice versa. The estimation

using all these seven kernels are very similar to one another except the rectangular kernel. I think the reason is that rectangular kernel is discontinuous and hence requires larger bandwidth (which somehow 'decreases' the discontinuity) to achieve the same smoothness.

In Figure 2, we plot the loess smoothing lines for the temperature against the humidity at the 72-th epoch every day. We tried different spans and all three degrees of local polynomials. From the figure, it's seen that a higher span produces smoother lines since the estimation involves more effective data points. Besides, a lower degree produces less smooth lines, since the complexity of model is lower.

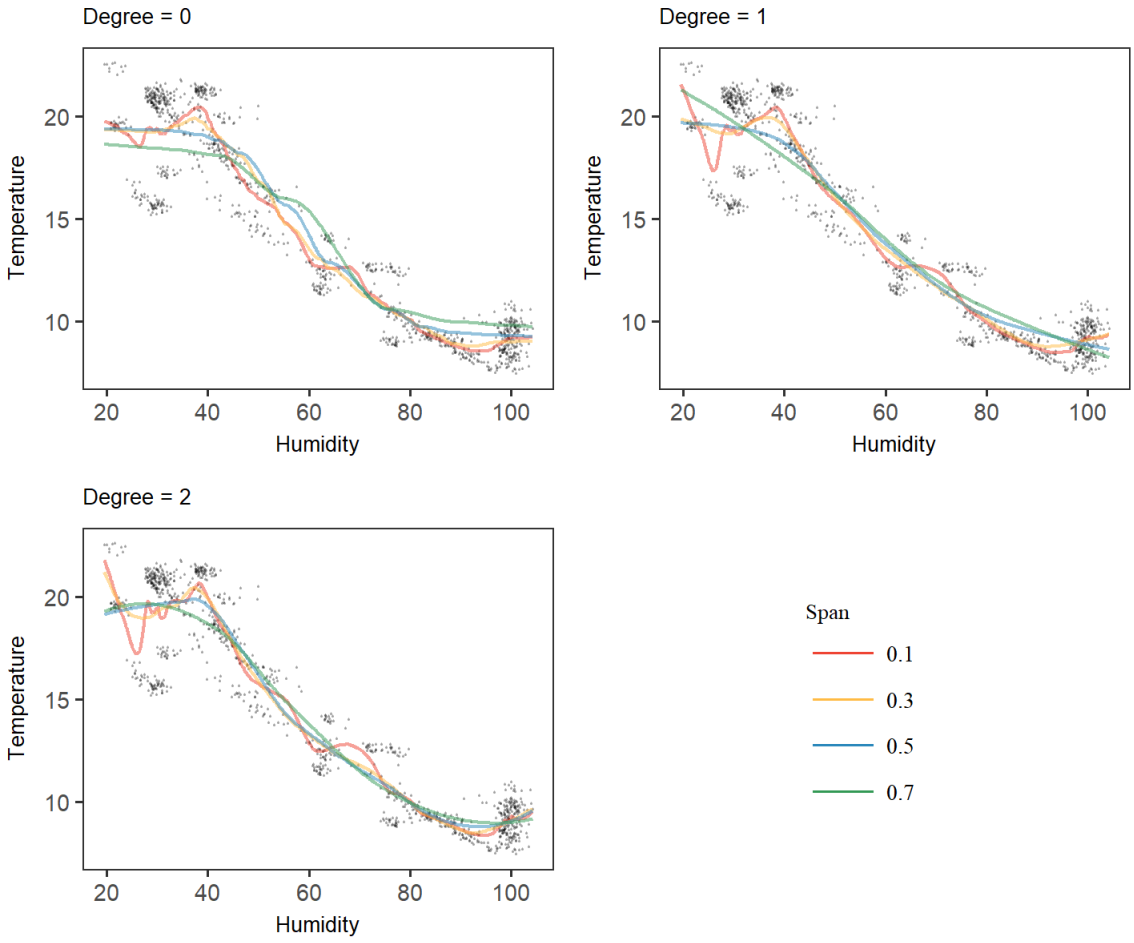

Figure 2: Loess Smoothing with Different Spans and Different Degrees of Polynomials

3 The Data

Now we move on to a new set of data. The data contains two parts, "question_data" and "lingData", corresponding to the questions and the answers from the respondents. 47471 people from across the United States participated in this linguistic research and answered all or parts of 122 questions.

In this report we are only interested in 67 questions. "lingData" contains the answers of all people as well as their ID, city, state, ZIP code, latitude and longitude. If one person does not answer one question, "0" is recorded. "question data" contains the information of the questionaire, including the questions and the answers.

3.1 Data quality and cleaning

This dataset isn't as bad as the redwood data, but there are still some issues. Therefore, we carefully check the data quality and clean it in this paper. For the sake of length, only important issues are clarified here. All other details can be found in the comments of "R/clean.R".

It is found that there is a one-to-one map from ZIP code to latitude and longitude. But there are 1020 records, from 200 different ZIP code areas, have missing values in latitude and longitude. We download a free database from http://www.boutell.com/zipcodes/ and fill in the missing values of 67 ZIP code areas. For the rest part, we write a web crawler in R trying to get them from https://usa.youbianku.com/zipcode/. Only several more is completed, therefore it is removed from the data cleaning code since it's time consuming.

After checking the left ZIP code, we find out that they are either not ZIP codes in contiguous US or even not valid at all. Besides, there are 208 more records locating Alaska or Hawaii. Since this is quite small comparing with a total of over 46 thousand records, we remove all 858 records corresponding to these ZIP codes for future convenience in processing and presentation, see Figure 3.

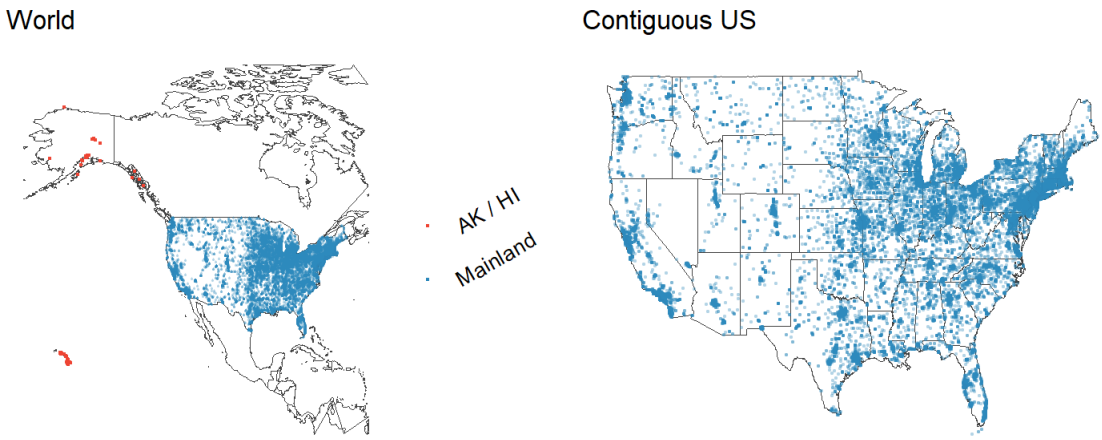

Figure 3: Geographical Distribution of Survey Answers

Then we move on to check the involvement of each person, that is, the number of questions a participant answers. If only a small proportion of questions is answered, then it is unwise to retain them in the dataset since the participant might just randomly answer several questions. It turns out that 84% of people answered all questions. Figure 1 shows the relationship between the number of answered questions and the number of people answering no more than that number of questions, that is, the $f(n) = \#\{i : i\text{-th person answers no more than n questions}\}$. It is observed that 1021 people answered no questions and there are only 456 people answering several but no more than 57 questions. We must remove the people answering small number of questions, so it delete all those who answered less than 57 questions is a proper choice. Here 57 comes from Figure 4. In this figure, we set the y axis in a logarithm form and it is seen that the slope becomes much greater when the number of questions answered becomes larger than 57. But in fact, it makes little difference if we change to any integer from 55 to 60.

3.2 Data Binarizaion

For the convenience of further analysis, we first transform "lingData" into a binary form, say, "lingData_binary". This is to create one column for each possible choice of each question. For example, if John chose the j-th answer for the k-th question which contains n choices, then his record to this question is an n-dimensional vector with its j-th entry to be 1 while all others 0. Besides, if he did not answer the

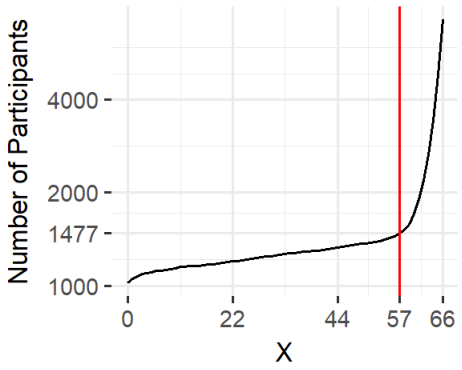

Figure 4: Number of Participants Answering No More Than X Questions

k-th question, the record is a zero vector. In this way, 67 answer columns in "lingData" are splitted into 468 columns with only entries in $\{0, 1\}$.

3.3 Data Aggregation

Over 45 thousand lines of data seems too large for the following analysis, and retaining the individual answers doesn't help analyze in any sence. Therfore, aggregation can eliminate some individual noises as well as improve the robustness. Here we use the binarized "lingData_binary" to perform aggregation. In this dataset, the average of a group of observations represents the proportion of each answer for each question, which can be regarded as a reasonable summarization of the group. In the following discussion, we will use averaging as the technique for aggregation.

A natural way to group is putting data with the same location together. In particular, we can perform aggregation with regard to zip codes. Since there are over 45 thousand lines in "lingData_binary", while the number of unique zip codes is only 11393, this zip code aggregation can reduce the number of lines by nearly three fourths.

We can aggregate the data one step further using the county map data from R package "maps". This could further reduce the time for computation while retaining visual effect and geographical information. Figure 5 visualizes the number of samples in each county across contiguous United States. It is seen that many counties have sample size 0, especially in the western part, such as Nevada, Utah and Idaho. The number of interviewees are significantly higher in some states, such as California or Florida.

3.4 Exploratory Data Analysis

The most natural way to explore (pairwise) relationship among questions is to visualize the distribution of all answers for each question. Here we use question 51 and question 52 for illustration. Question 51 is: Would you say 'Are you coming with?' as a full sentence, to mean 'Are you coming with us?'; question 52 is: Would you say 'where are you at?' to mean 'where are you?'. The answer to both questions consists of 'yes', 'no' and 'other'. These two questions seems to be very similar, as both are about the usage of prepositions. In Figure 6, each county is filled in the color corresponding to the most popular answer in that county. However, it doesn't show any clear relationship between these two questions.

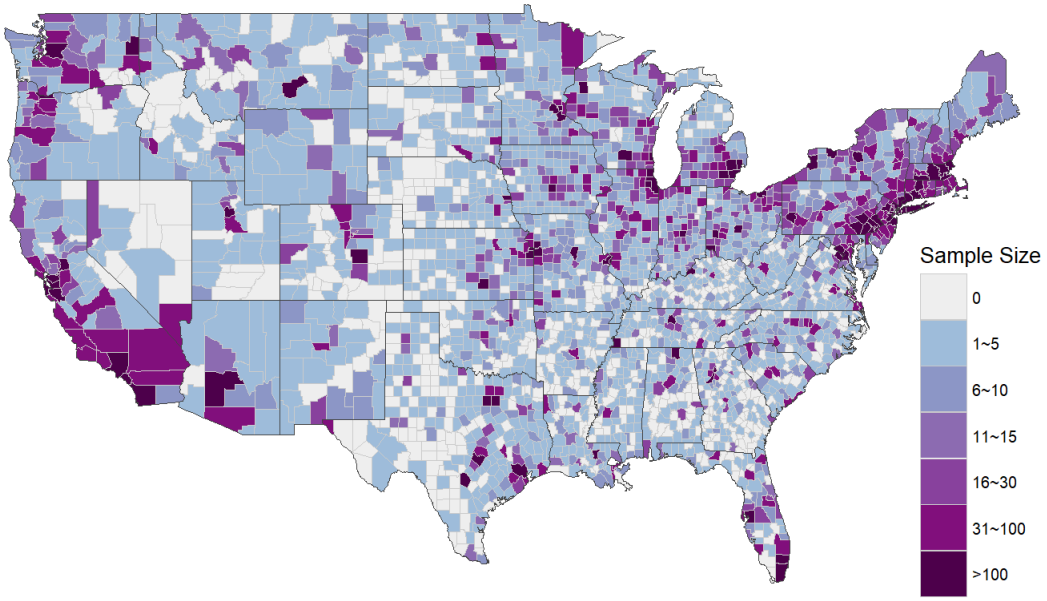

Figure 5: Sample Size in Each County

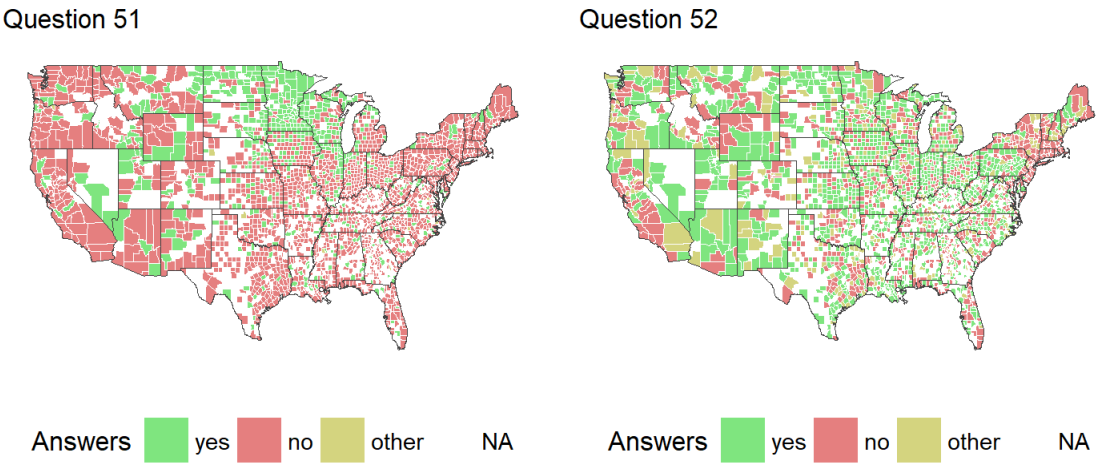

Figure 6: Answers Distribution For Question 51 and 52

4 Dimension Reduction Methods

By doing pairwise χ^2 test, it seems that every pair of questions cannot be regarded independent. This leads us to do some dimension reduction. We use the binarized data, "lingData_binary", which contains 468 columns for questions to perform dimension reduction. Although the size is not extremely large, it is still helpful to reduce the dimension, that is, the number of columns.

4.1 Principal Component Analysis

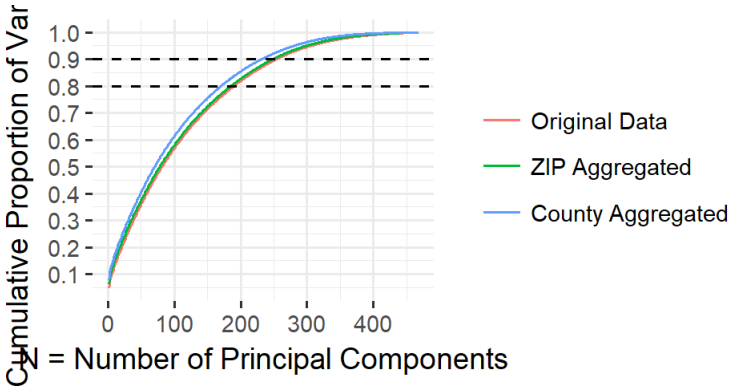

Figure 7: Cumulative Proportion of Variance of the First N Eigenvalues

One of the most natural ways towards dimension reduction is PCA, i.e. Principal Component Analysis. Figure 7 shows the cumulative proportion of variance of the first n eigenvalues. Observe that aggregation perfromed in the last section does reduce the noise, since with deeper level of aggregation, the slope becomes steeper. It is not strange that no clear turning point exists since the number of components is large. But it is seen that to preserve the 80% variance we need over 150 components no matter under which level of aggregation. Although PCA is lack of interpretability, it still reflects some interesting aspects of the data. By observing the average loadings of first twenty principal components, we found that the first several answers with highest average loadings are: the first answers for question 115, 77 and 63, the second answers for question 91 and 93.

4.2 Multidimensional Scaling

Now we move on to perform MDS, i.e. multidimensional scaling. To perform MDS, we need to calculate the distance between each pair of questions. So first of all we need to decide which metric to apply. Continuous metric including Euclidean and Minkowski is not proper since each question column is a set of discrete data. Therefore, relative variational information (RVI) is introduced to measure the discrepancy between discrete data[[3]; aghagolzadeh2007hierarchical]:

$$RVI(X,Y) = 2 - \frac{H(X)+H(Y)}{H(X,Y)},$$

where $H(X,Y)$ is the joint entropy of X and Y, and $H(\cdot)$ is the entropy. It can be check to be a metric taking values in $[0,1]$. The left part of Figure 8 displays the two dimensional MDS projection using the RVI metric distance matrix.

Once the MDS projection is obtained, we may proceed to cluster them using the projected distance. The hierarchical clustering and k-means clustering are not proper in this case since several of the 67 questions here focus on very similar questions, such as question 68, 69, 70 and 71. Besides, hierarchical clustering is somehow greedy and might no be robust while the result from k-means is hard to interpret. Therefore, we

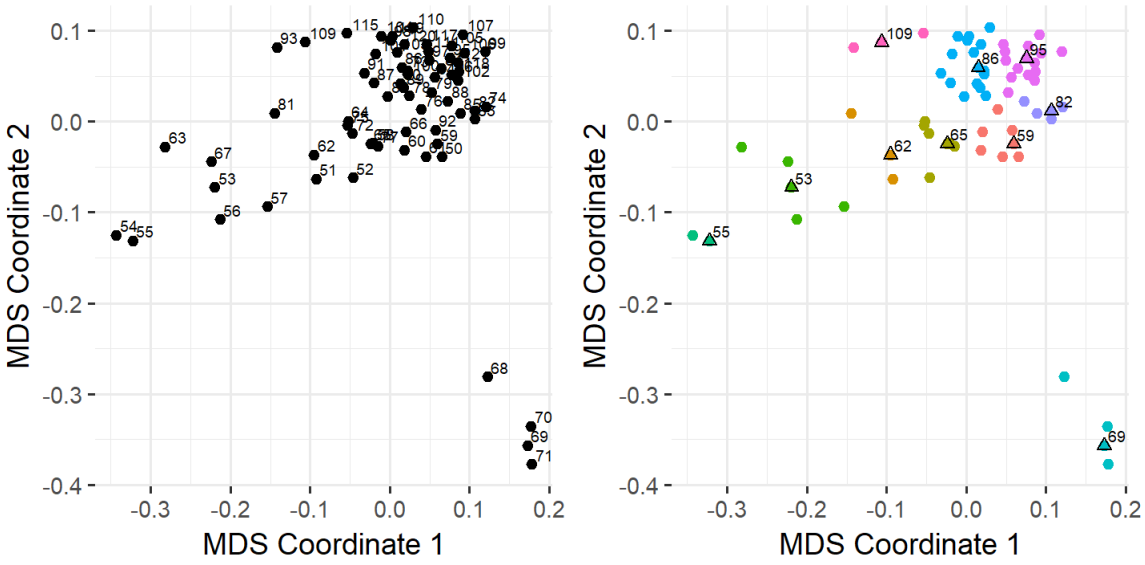

Figure 8: Two Dimensional Projection of MDS with RVI Metric (Right: with cluster result)

use k-medoids clustering algorithm, which is the same as k-means except that it only uses the sample point as cluster centers. Thus, the center can be regarded as the representative of that cluster and hence more interpretable. These centers are then treated as topics and the other questions in its cluster as the ones related to this topic. When perfroming k-mediods, we use the result from k-means as the initial guess of medoids to achieve quick convergence as well as better clustering result.

To determine the number of clusters, we use the average silhouette width criterion. When plotting the average silhouette width versus the number of clusters, it is seen that 8 or 10 might be good choices even though less than 5 topic might give large silhouette width, since we believe that there should be more topics in this research and we want more clusters to capture these topics. We select 10 as the number of clusters. In fact, one of these clusters contains two questions about the usage of "anymore" and another cluster contains four questions about the appellation of grandparents. The right part of Figure 8 displays each clusters together with their medoids. The selected questions are: 53, 54, 59, 69, 72, 82, 89, 93, 95 and 98.

4.3 Sample Clustering

After trying to reduce the dimension regarding the questions, we now move on to cluster the participants and try extracting cultural differences among people. Hopefully, this research can be used to study the separation of (maybe geographically) different subcultures.

In this section, we use PCA, and then k-means on the countylevel aggregated binary data for illustration and visualization, see Figure 9 to 11. Figure 9 is a set of density plots (upper triangular parts) and scatterplots (lower triangular parts) of first four principal components. It seems that the first two components define three clusters since the scatterplot has a clover shape. But the third and the fourth components are not informative. Thus we conduct PCA on the first two dimensions. Figure 10 displays the average silhouette width for different number of clusters. Obviously the best choice is to go with 4 clusters. Figure 11 plots the silhouette width for each cluster (left panel) as well as the clustering result in the space of first two principal components (right panel). The pattern shows that 4 clusters are all desirable in terms of both the group size and the average silhouette width. Notably, those points lying on the boundaries indicate the existence of continuum. Theoretically, boundary points usually have small silhouette value, thus can be used to find possible locations of linguistic continuum.

Figure 12 shows the clustering in the map. The yellow areas correspond to eastern parts, the red parts

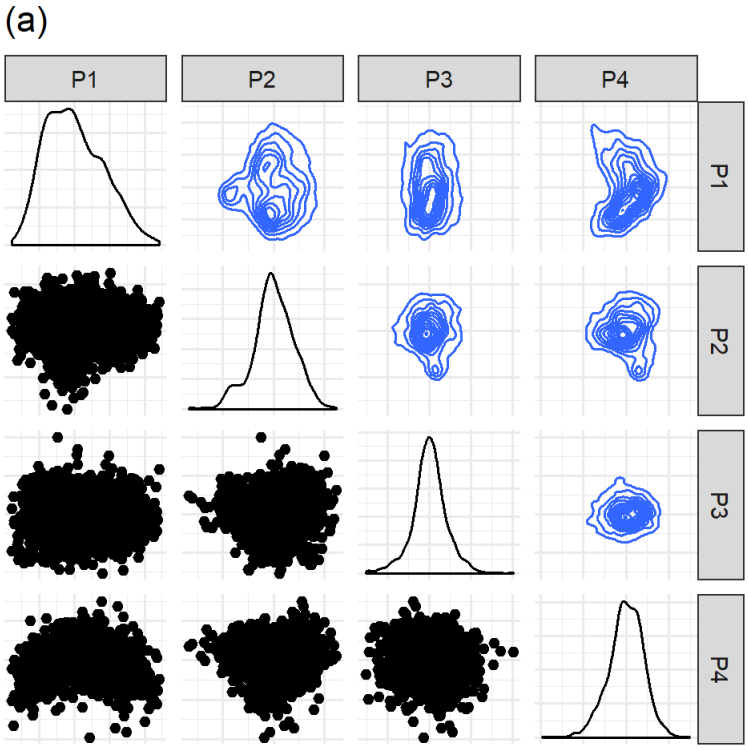

Figure 9: Relationship among the First Four Principle Components

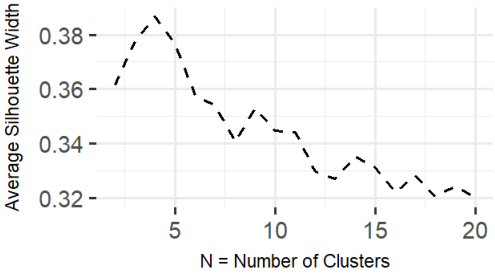

Figure 10: Average Silhouette Width using N Clusters

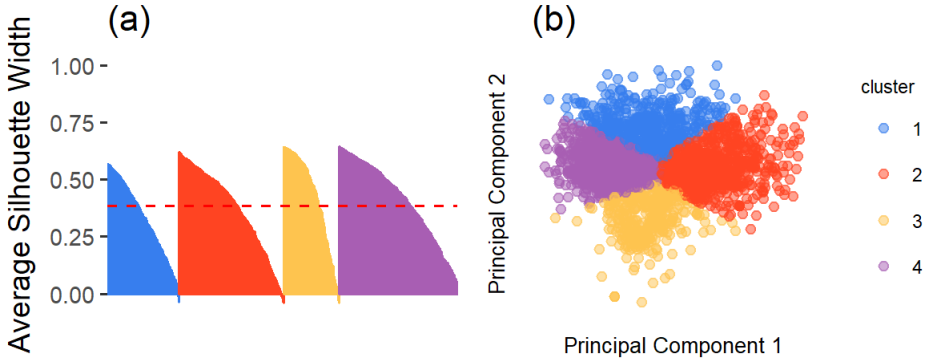

Figure 11: K-means Clustering Result with 4 Clusters

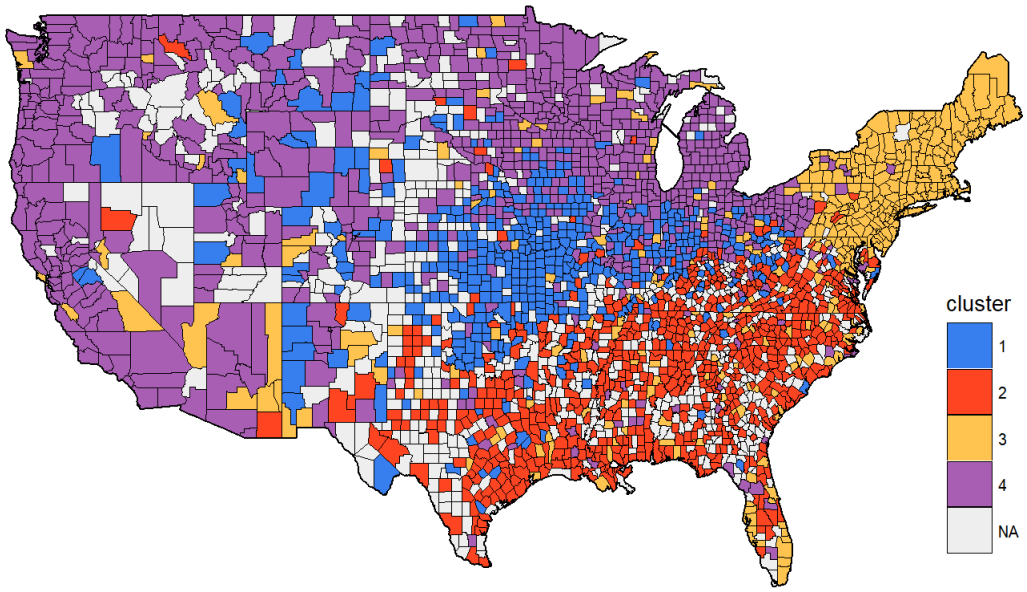

Figure 12: K-Means Clustering Result with 4 Clusters on Contiguous US Map

correspond to southern part, the blue areas correspond to western parts, mid-western parts and Great Lake district, and the green parts correspond to middle parts of US. The gray areas have no interviewees. This clustering is reasonable. Figure 13 displays the silhouette width of each county. The light areas have high silhouette width and hence can be regarded as "pure" in its cluster, while dark areas have low silhouette and can be regarded as the "continuum" between two clusters.

An interesting fact is, the clustering is also somehow geographical, since the 4 linguistic cluster of contiguous US is very similar to the four statistical census regions defined by United States Census Bureau: Northeast, Midwest, South and West. Only that the north part of Midwest region is combined with the West region in our clustering. We will discuss the robustness of this finding in the next section.

To study the robustness of the clustering, we also present the result using 3 instead of 4 clusters in Figure 14. It is observed clearly that the color of silhouette map is much deeper. Moreover, the new cluster #1 is roughly the union of old cluster #2 and #3 while the other two clusters are quite similar. Both facts tell the fact that 4 clusters is an optimal choice for k-means algorithm here.

5 Stability of findings to perturbation

As discussed in the last section, the 4 linguistic cluster is very similar to the 4 regions defined by the US Census Bureau. To analyze its robustness, perturbation analysis is often introduced. Usually, the problem involves a lot of samples to be clustered and hence it is not easy to identify which ones of them or which factors influence the clustering result significantly. Thus, we need to test the stability of the clustering. A natural way is to perturbate the data and see the difference of clustering results.

In this section, we will focus on perturbing both questions and answers. To perturb questions, we randomly delete a proportion α of columns (questions); as for perturbing answers, we randomly alter the values of a proportion α of entries from each column. α is a parameter controling the perturbation level. Just a reminder, here we are going to use different levels of both pertubation to test the stability of k-means clustering based on the first two principle components from PCA to see how the result is changed.

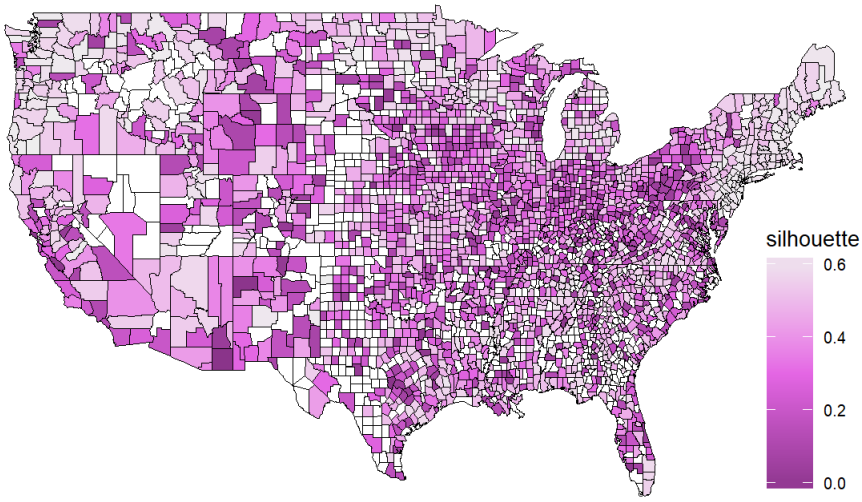

Figure 13: Silhouette Width with 4 Clusters on Contiguous US Map

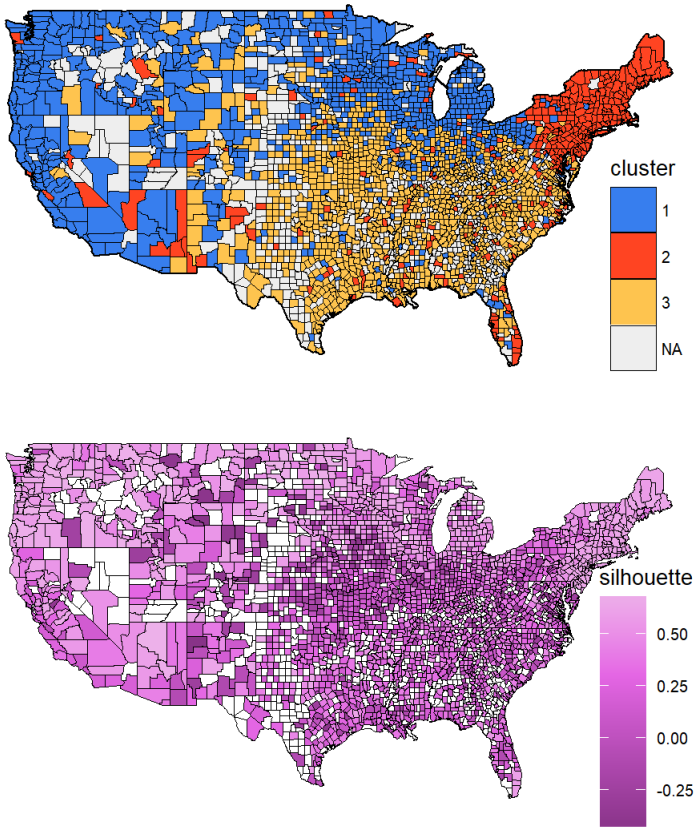

Figure 14: K-Means Clustering Result (Upper) and Silhouette Width (Lower) with 3 Clusters on Contiguous US Map

After obtaining the clustering result of pertubated data, we calculate the dissimilarity among results to see whether or not it is robust. A natural way to calculate the dissimilarity is simply using the proportion of coincidence, i.e. $\#\{R_1(i) \neq R_2(i)\}/n$, where $R_k(i)$ is the cluster to which the i-th datapoint belongs under the k-th clustering result and n is the number of points being clustered. Since the labels of clusters is interchangeable, we need to take the maximum over all possible permutations of cluster labels, which is exactly the assignment problem and can be calculated with Hungarian method[4]. We can tell the robustness of our finding by comparing the clusters of original data and those of perturbated data. Figure 15 displays the dissimilarity of clusters with respect to perturbation levels, 20 repetitions are done per level for both questions and answers perturbation. It is seen clearly from the boxplot that the result is stable since when the perturbation level is low the dissimilarity is quite low and when the perturbation level is high, the dissimilarity is still not high: only very few of them goes beyond 20% even when the perturbation level is as high as 32% under both types of perturbation.

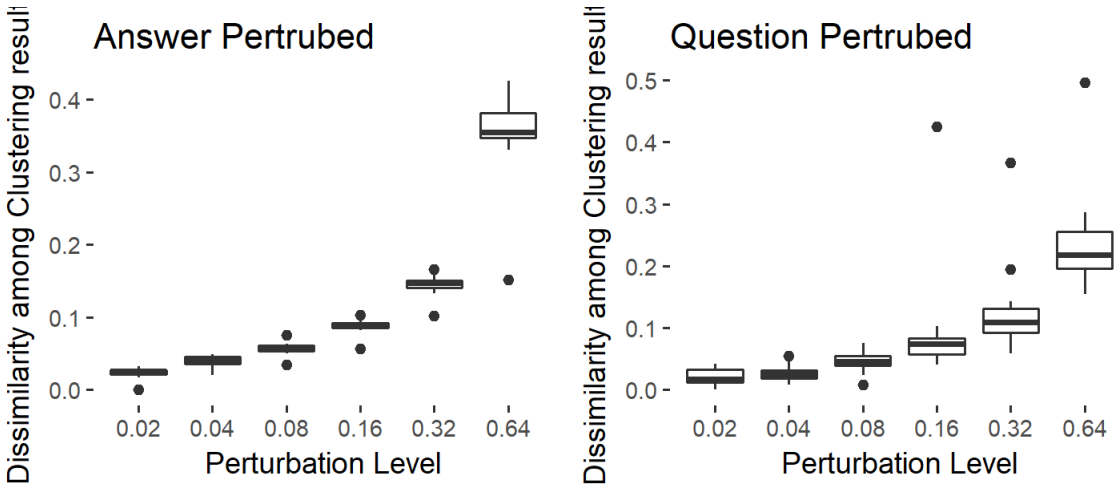

Figure 15: Dissimilarity among Clustering results

Figure 16 shows the clustering result on contiguous US map with 32% and 64% of the questions and the corresponding answers deleted, respectively. It is clear that even with pertrubation level as high as 64%, the geographical pattern still largely holds. This once again proved the robustness of the finding.

In fact, the perturbed data and the perturbation analysis can be performed to (and not limited to) everything we mentioned in the previous sections. But we are not going to discuss them in this report.

Reference

[1] Tolle, G., Polastre, J., Szewczyk, R., Culler, D., Turner, N., Tu, K., Burgess, S., Dawson, T., Buonadonna, P., Gay, D. and others. (2005). A macroscope in the redwoods. In *Proceedings of the 3rd international conference on embedded networked sensor systems* pp 51–63. ACM.

[2] Nerbonne, J. and Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. *Computers and the Humanities* **37** 245–55.

[3] Nerbonne, J. and Kretzschmar, W. (2006). Progress in dialectometry: Toward explanation. *Literary and Linguistic Computing* **21** 387–97.

[4] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics (NRL)* **2** 83–97.

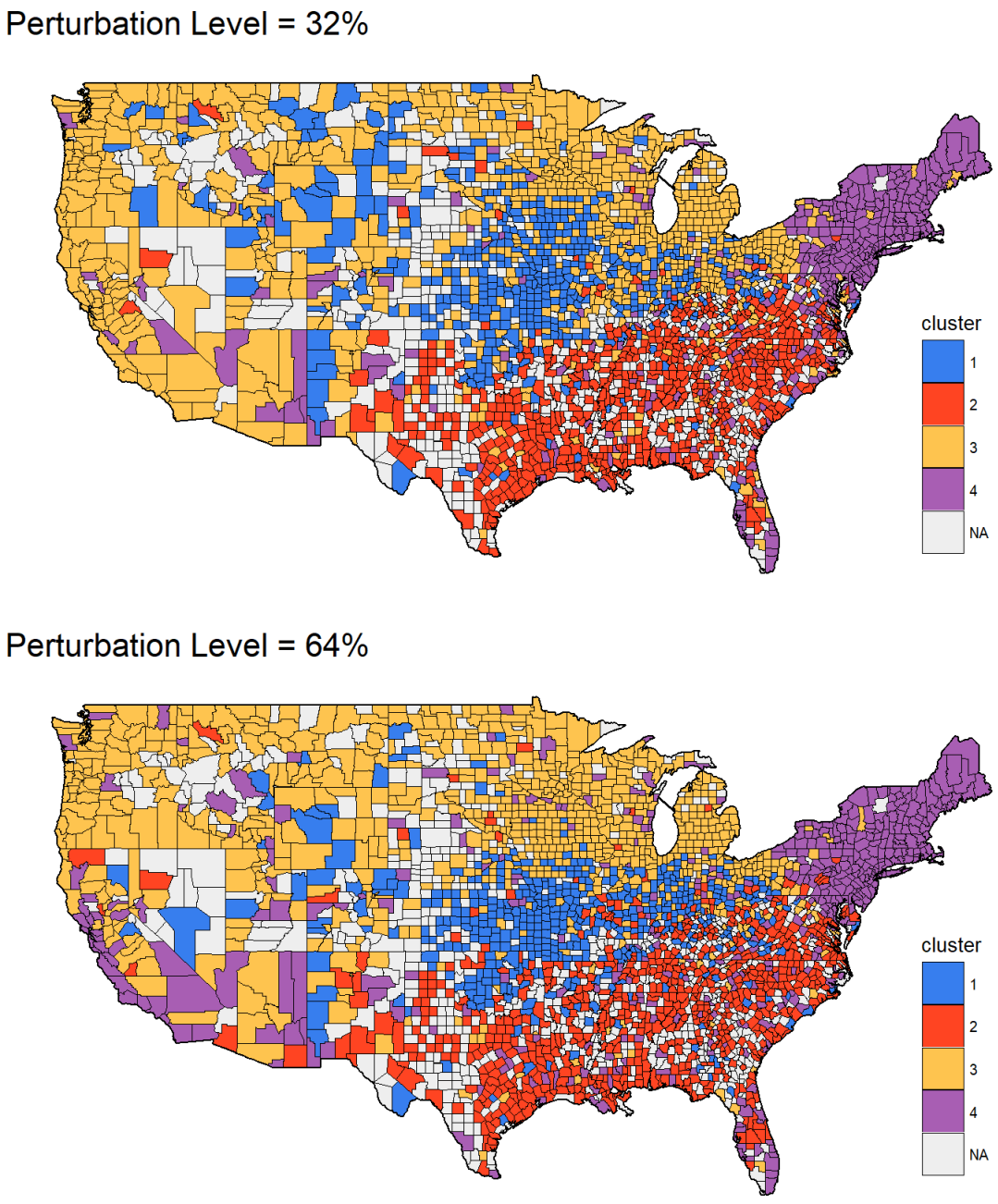

Figure 16: K-Means Clustering Result with 4 Clusters on Contiguous US Map with Perturbation