# Lab 2 - Linguistic Survey
# Stat 215A, Fall 2017

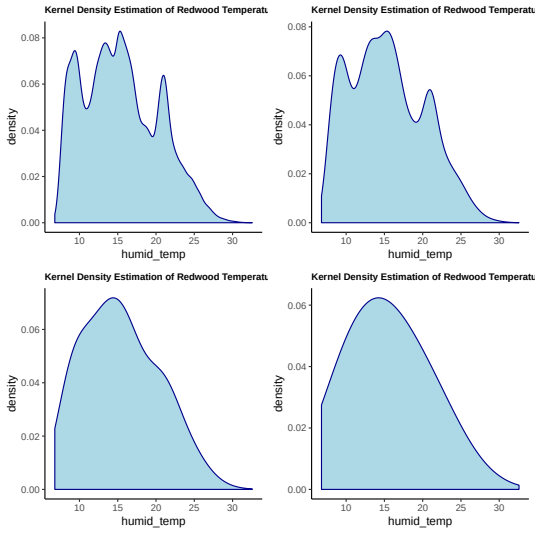
Linqing(Waverly) Wei

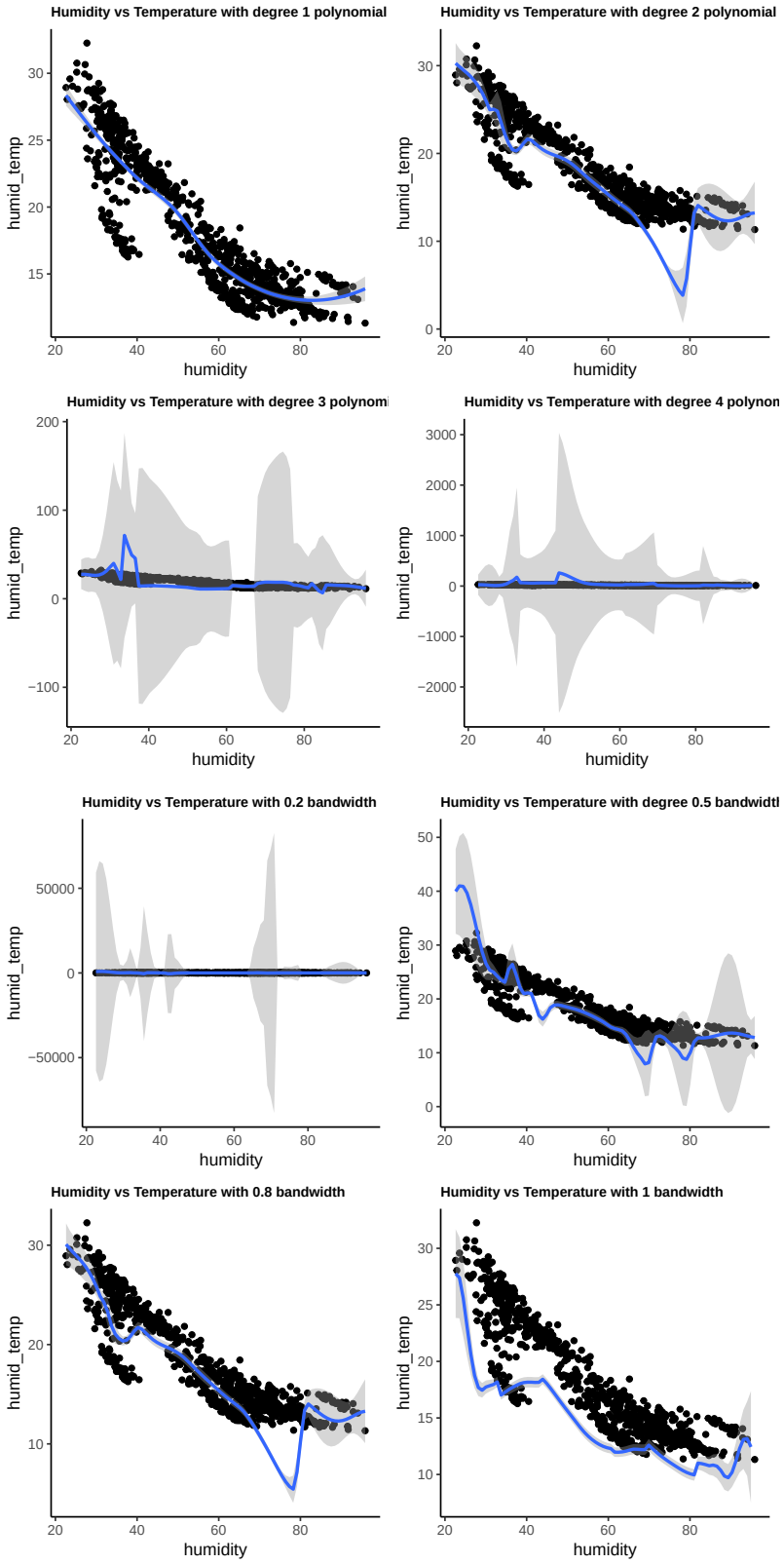
October 4, 2018

## 1 Redwood

When varying the bandwidth, we are basically changing the smoothness of the kernel density estimators. When bandwidth is 1, there are four modes shown in the plot, meaning that these four temperature bins capture the most amount of data points. When adding the bandwidth, we are smoothing over the bins, such that the plot shows three modes when bandwidth is 2 and unimode with bandwidth 5 and 10 at temperature 15°C.

When fitting a loess cruve to the data, the higher the degree of polynomial, the better the curve fits the data points, meaning a smaller bias. However, overfitting leads to very high variance. When using very small bandwidth like 0.2, the curve fits data extremely well but leads to a high variance. Increasing bandwidth captures the more variation within the data but in turn gives a large bias. Therefore, it is quite important to find the balancing point between bias and variance.




Humidity vs Temperature with degree 1 polynomial
Humidity vs Temperature with degree 2 polynomial
Humidity vs Temperature with degree 3 polynom
Humidity vs Temperature with degree 4 polynon
Humidity vs Temperature with 0.2 bandwidth
Humidity vs Temperature with degree 0.5 bandwidtl
Humidity vs Temperature with 0.8 bandwidth
Humidity vs Temperature with 1 bandwidth
humid_temp
humidity

# 2 Introduction

Modeling linguistics is like modeling human "cultural genes." Linguistic patterns are both preserved and modified over time and over human migrations. Linguistic variations are always prevalent, obvious, yet hard to quantify. Taking into individual features would lead to a hard characterization, but luckily the presence of Dialectometry handles this quantification problem quite well by realibly aggregating over geographical patterns. This study investigates people's dialectology patterns in the United States and maps onto geographical areas to detect the underlying pattern of what has been shared and what leads to the differentiations? The questions are designed to capture all aspects of people's social and cultural habits in order to refelct the linguistic pattern underneath.

# 3 The Data

There are three datasets: question, which contains the survey questions and answers; lingData, which includes subject ID, states, zipcode, city, each person's response to the questions and geographical coordinates. lingLocation expands all the answer choices and count the number of people fall under each answer choice within a small block of geographical location.

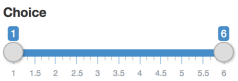
## 3.1 Data quality and cleaning

This dataset has two major issues: large amount of missing values in longitude and latitude columns and incorrect geocode in STATE column. I removed the rows containing missing values in latitude and longitude. For states, I removed the ones with apprent wrong code like "!L". Then I noticed that some observations have two-letter state code but not in one of the 50 states, possibly some territories I've never heard of. I inspected the number of people falling under those codes, which is quite small. Therefore, I made a decision to remove those unknown code since those wrongly-coded observations won't impact the overall underlying trend of the data.

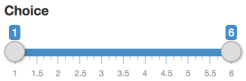
## 3.2 Exploratory Data Analysis

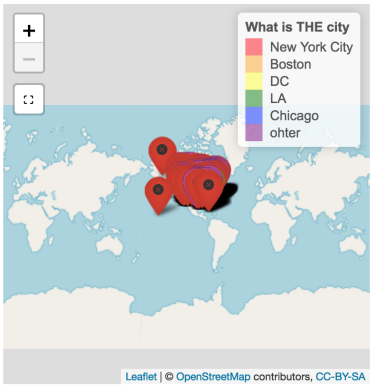
I decided to choose question 104 and question 95. Q95 is: What is THE city? Q104 is: What do you call a public railway system. I chose those two questions because I suspected that they will both display some localized trends.

**NOTE:**
**These maps are interactive plots. Please find the .Rmd and .html files to reproduce the maps**

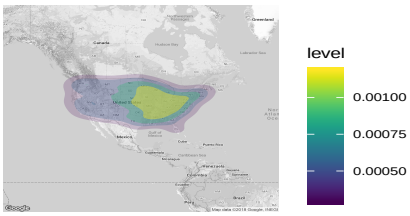




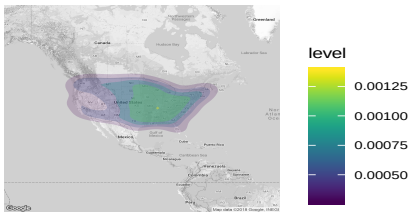


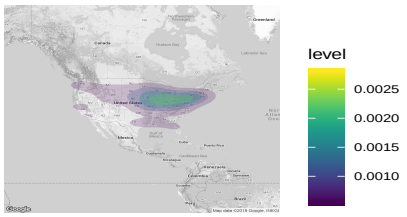

**The Subway**



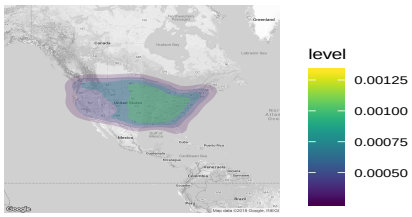

**the L, or the El**



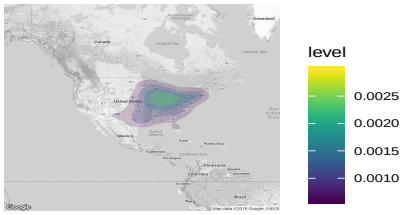

**the T**



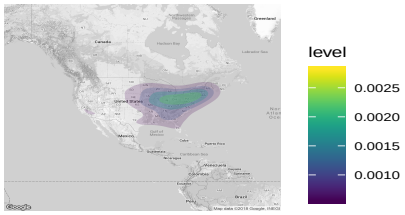

**the metro**

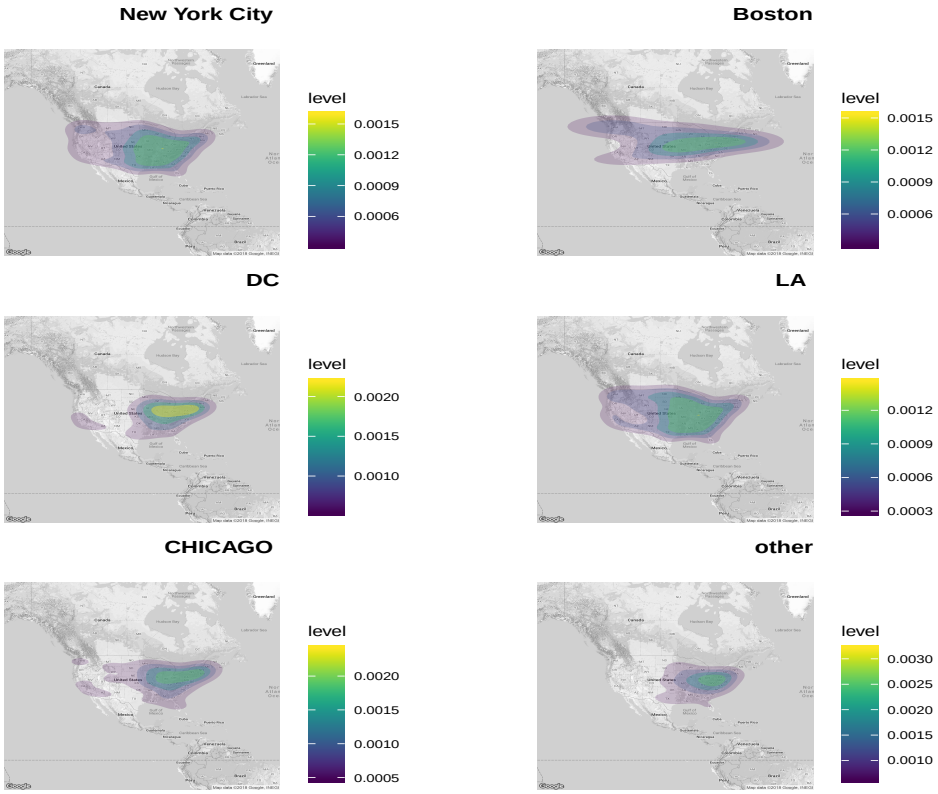



**BART**



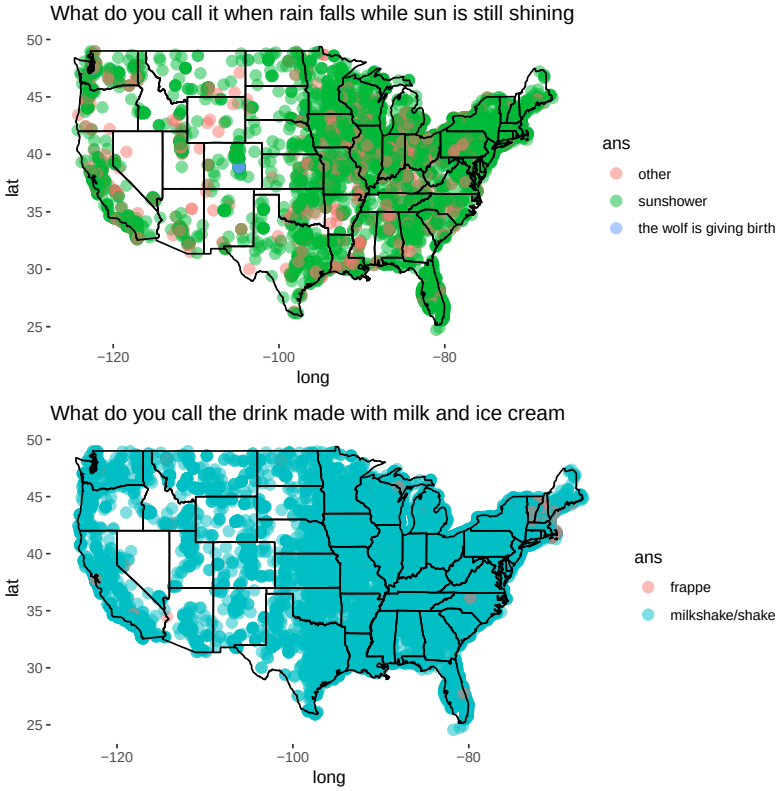

**other**






From the plots, we can tell eople do have a shared answer for "the public railway system" which is "The subway." However, small clusters do form by geographical areas. For example, people who answered "The Metro" are more centered around northeast. We only see tiny amount of people choose "BART" on the west coast. It doesn't mean that BART is not a prevalent term in the bay area, but the data is not weighted and less people on the west coast participated into the research compared to the east. This reflects an interesting aspect of dialectometry: although many geographical areas do have a locally shared linguistic register, for commonly used social constructions, we do have a shared term to identify the object.

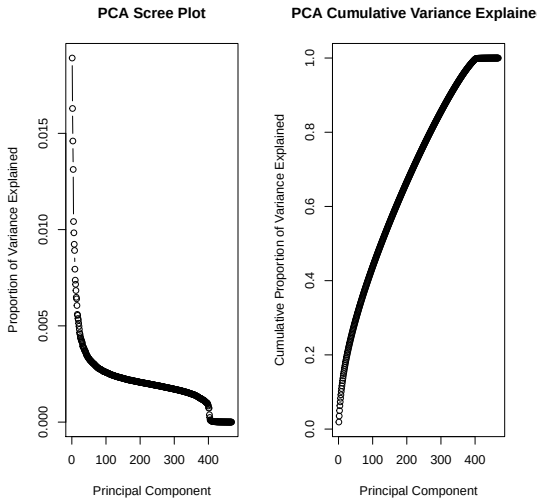
Question 95 "What is THE city?" is more interesting. "City" is not like a public railway system, it is more of a descriptive term, very self-defined. In this case, we do see a larger variation among the answers. People refer to the largest and closest city as "The City." Initially we might suspect that we would to see very distinct geographical clusters, but in fact, there is still a commonly shared term: "New York City." Many people on the west coast did choose NYC as the city. Rather than defining this as a linguistic habit, I view this as more of a cultural habit. New York City is culturally defined and publicized with a metropolitan image, and that made people think out of the adjacent city and pick NYC as the city. These two questions can be used to PARTIALLY predict one another since they both define similar geographical clusters. For example, the area with the high intensity of people choosing "Metro" also has high intensity of choosing "NYC." However, since the west coast has a smaller amount of participants, we fail to see distinct correlation between the two questions' answer choice on the west side.

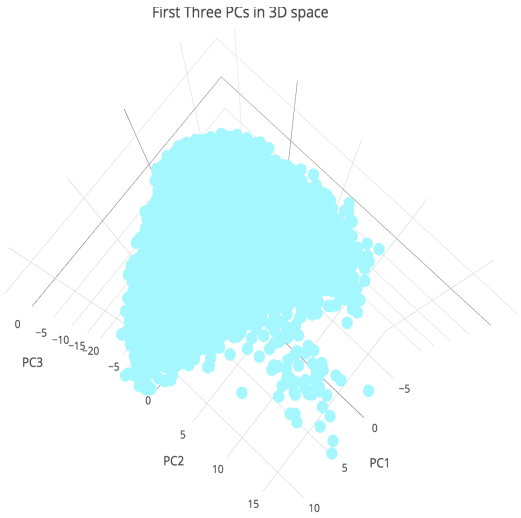



Many other questions do not show any geographical distinctions. It seems like people show consistency of dialect bahaviors when it comes to food and weather. Question 80 "What do you call it when rain falls and sun is shining" shows a very prevalent trend. People call it sunshower across all the states. The answer answer choice "The devil is beating his wife" appears in the southern part with a small cluster. It could be a very specific and localized term but not prevalent. Question 63 shows similar result. Almost all participants call "the drink made with milk and ice cream" as milkshake, despite a few people on the east coast call it "frappe."

# 4 Dimension reduction methods

Before proceeding with PCA, we conducted binarization because in the original dataset, multiple answer choices are grouped together under each question. Binarization helps with expanding the variation in the original dataset. Also, a scaled and centered binary matrix would make PCA more computationally efficient.

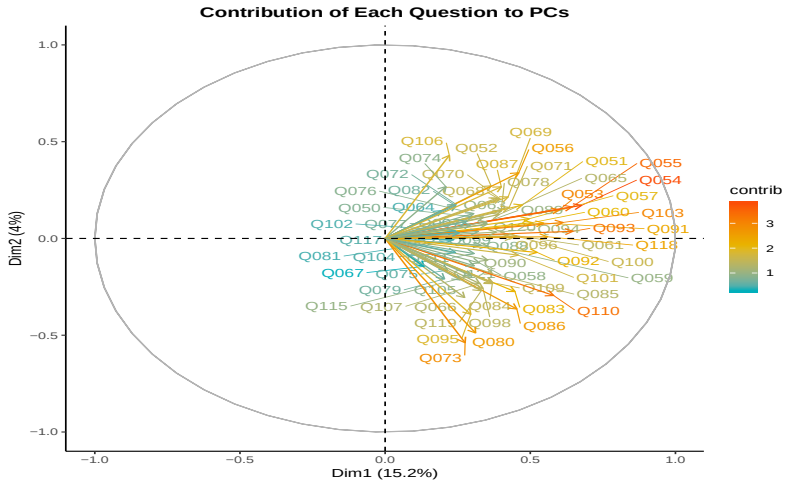


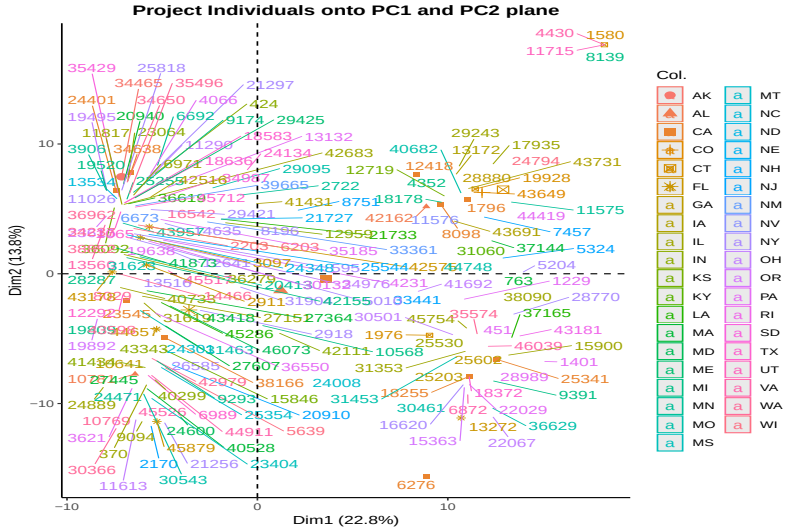



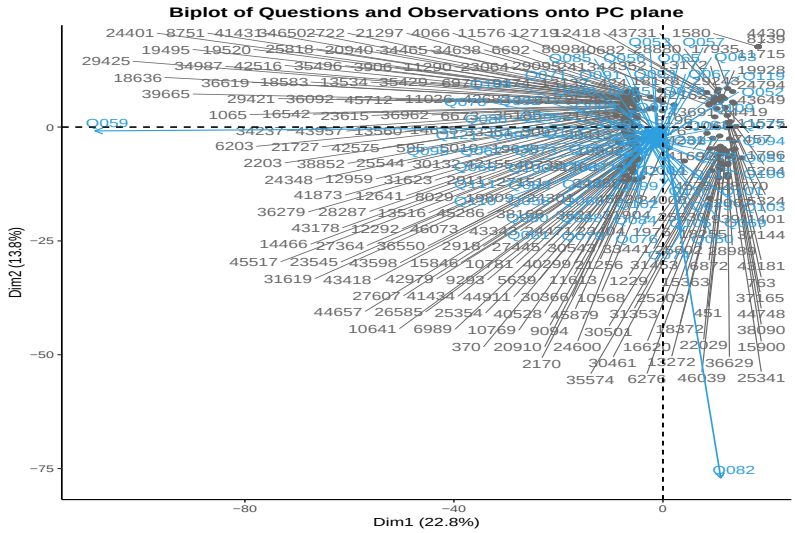

**NOTE:**
**This is an interactive plot. Please click the link below.**

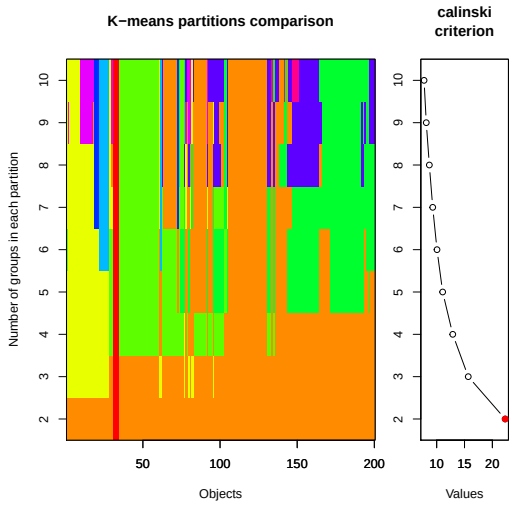
`https://plot.ly/~linqing_wei/19.embed`

After calculating the principle components of the binarized dataset, PC1 only captures 1.5% of the total variance. This is very reasonable since after binarizing the dataset, each variable represents only one answer choice of one single question. The amount of variation a single principal component is able to capture would be very limited. The threshold appears at 400, meaning that roughly 400 principal componenets would be able to capture more than 98% of the total variance. After projecting the data onto PC1 and PC2, it's hard to evaluate whether projecting onto this plane would be ideal. Therefore, I created a 3D plot to explore different ways of projecting. It seems that projecting onto PC1 PC2 is actually quite ideal. The data is separated into one huge cluster and one sparse, smaller cluster. This makes sense considering how the data was collected. Many participants centered around east coast. Less people took the surveys on west coast. If projecting onto the plane made by PC1 and PC3, we would only be able to see one large chunk,which is not very meaningful. Roughly, the data should be culstered into two geographic categories. In the next section, we will dive more into the data for a better view of the principal components.


Contribution of Each Question to PCs
Dim2 (4%)
Dim1 (15.2%)
contrib
1.0
0.5
0.0
−0.5
−1.0
3
2
1
Q106 Q052 Q069 Q056 Q074 Q087 Q071 Q051 Q055 Q072 Q070 Q078 Q065 Q054 Q076 Q082 Q053 Q057 Q050 Q064 Q060 Q103 Q102 Q093 Q091 Q081 Q104 Q061 Q118 Q067 Q090 Q092 Q100 Q058 Q101 Q059 Q079 Q105 Q199 Q085 Q115 Q107 Q066 Q084 Q083 Q110 Q119 Q098 Q086 Q095 Q080 Q073

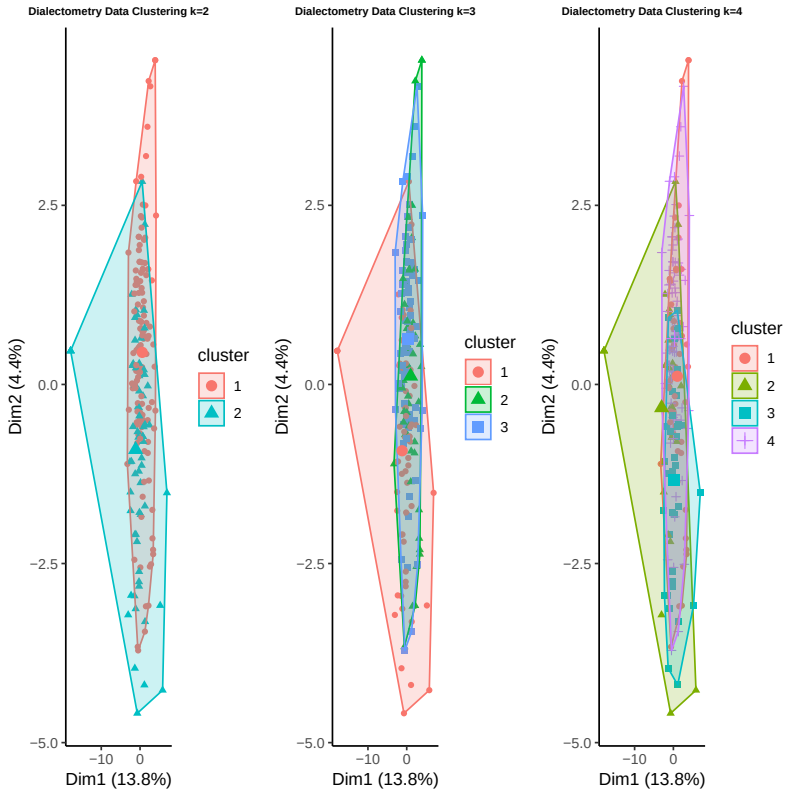

Project Individuals onto PC1 and PC2 plane
Dim2 (13.8%)
Dim1 (22.8%)
Col.
AK AL CA CO CT FL GA IA IL IN KS KY LA MA MD ME MI MN MO MS
MT NC ND NE NH NJ NM NV NY OH OR PA RI SD TX UT VA WA WI


Biplot of Questions and Observations onto PC plane
Dim2 (13.8%)
Dim1 (22.8%)
Q059
Q082



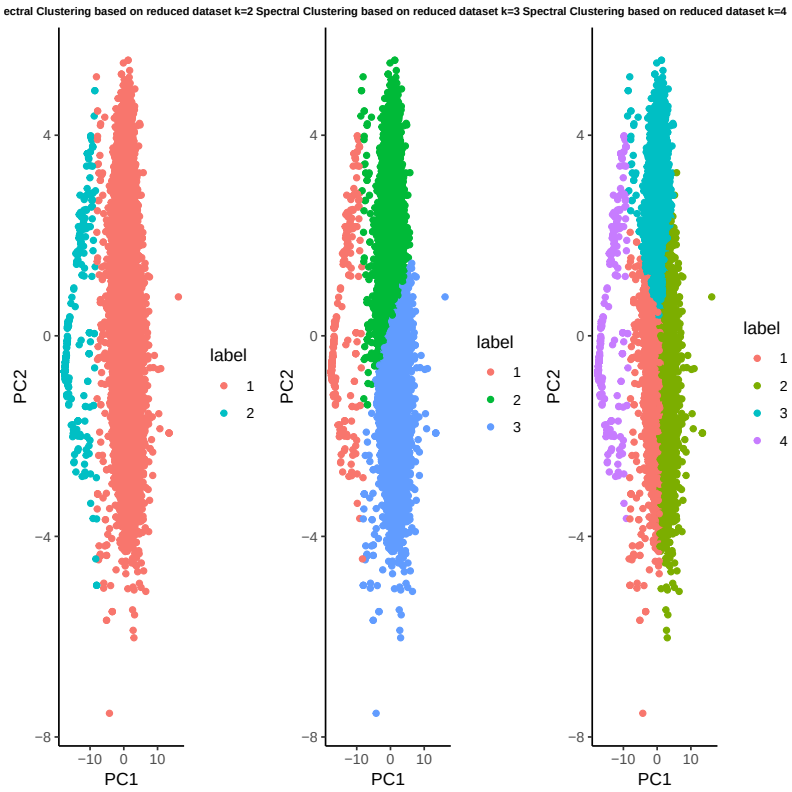

By mapping the full dataset onto PC1 and PC2, we found that individuals are clustered into three geographic groups. The result is seen most clearly only when we subset the data rather than plotting all the observations. One giant culster side by side with a mid-size cluster, and another tiny cluster in the corner.Q54,Q55 and Q93 contribute the most to the first two principal components, meaning that those questions are the major varaibles potentially capturing large amount of variations. Q54 and Q55 asked about whether it's appropriate to plug in "anymore" into the sentence..Q93 aksed about whether people use "on line" or "in line." These questions all capture people's grammatical pattern.

Another interesting finding is that Q59 and Q82 are the two major questions which separate the geographical groups.Q59 asks about "What do you call the game wherein the participants see who can throw a knife closest to the other person (or alternately, get a jackknife to stick into the ground or a piece of wood)?" It contains 21 answer choices, all colloquial. It makes sense that this question is able to distinguish geographical areas. Same thing as Q82, Q82 asks about"What do you call the gooey or dry matter that collects in the corners of your eyes, especially while you are sleeping?" It also has 21 choices with very localized expressions.One conclusion we can make from this finding is that: the more colloquial question, the more likely it would be able to separate the geographical groups.




The clinski criterion chooses k =2 as the optimal number of clusters. From the previous findings, we suspect that there should be three culsters. We could therefore varify the results by using cluster number of 2,3,and 4. Intuitively, k =3 would be a better number of clusters,roughly middle, east, and west three geographical areas. However, based on the clustering results shown in the plots, k =2 well captured the major variations.

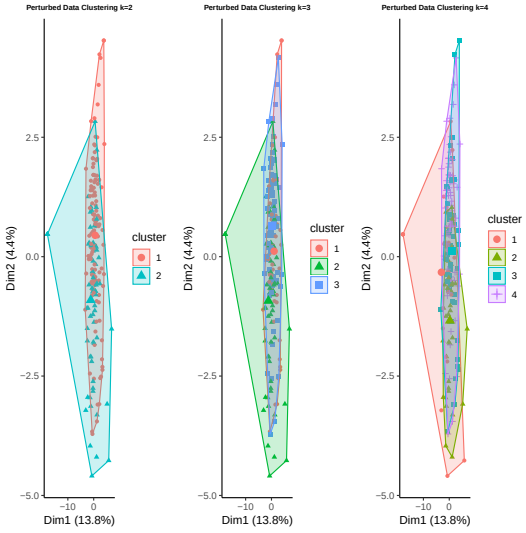
Now, we run Spectral Clustering on the PCA results, the reduced data, to see if there's any interesting clustering results we failed to capture in the previous analysis.



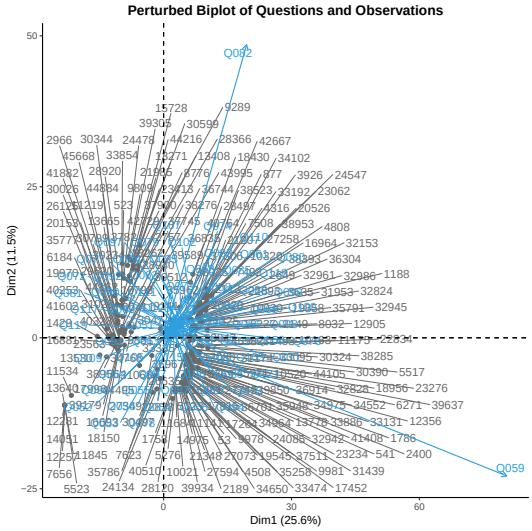

The spectral clustering results are quite interesting. It not only confirms the previous geographical separation but also adds another layer. When k=3, we found that there's also a north-south distinction on the eastern part, which we failed to capture in the previous plots.

# 5 Stability of findings to perturbation

The first perturbation is to perturb the k-means clustering results. Instead of using the starting point 20, we use 10 instead to see whether the culstering results are significantly different from the previous ones. When k=2 and k=3, the clustering results are essentially the same but when k is 4, instead of culstering the small group of points as a line, the perturbed one defines a more inclusive cluster to capture that minor group.



The second perturbation is conducted to varify the PCA results. Recall that to convery the PCA results in a more clear and interpretable way, we made a subset of data to see which questions better separate the geographical groups. Now, we perturb the subset to see if we can get the same results. Based on the perturbed result, we still get Q59 and Q82 as the two questions separating the geographical groups, which proves that our analysis is very stable under perturbation.




# 6 Conclusion

Dialectometry does provide insight on the quantification of linguistic variations with respect to geographical areas. By mapping the distribution and intensity of people's responses, we found that people do have a shared vocabulary across the country but with interesting localized variations. Principal component analysis does help with reducing the dimensionality and narrowing down the couple of vectors that capture the largest and most significant variations in the massive space. The questions that are designed to invoke more colloquial responses actually have an impressive performance of separating geographical groups. Observations are separated into three major clusters with the biggest one on the east coast. K-mean clustering and Spectral clustering also help with verifying that finding and reveal additional north-south geographical distinction. Since the results are fairly consistent under perturbations, we could reasonably conclude that our findings do show significant and well-separated geographical clusters representing people's varied dialectology.