Lab 2 - Linguistic Survey Stat 215A, Fall 2017

25948127
University of California, Berkeley

September 24, 2017

1 Introduction

In this lab report, I explore kernel density, smoothing, geospatial mapping, dimension reduction, and clustering techniques. This general class of procedures reflect the general notion that a primary goal of data science is sorting observations into interesting groups, and making predictions based on the insights gained from those groupings. In the first part of the lab, I apply kernel density plots and smoothing parameters to the redwood data from the last lab report. In the second part, I use data from the Linguistic Survey to understand geographic variation in American dialects. I apply geospatial analysis, and then extend the analysis with dimension reduction, and clustering techniques. Overall, I aim to illustrate the various tradeoffs associated with these methods by providing insights into both sets of data.

2 Kernel Density Plots and Smoothing

In this section, I revisit the redwoods dataset from Lab 1. I will forgo explaining the dataset in detail, and instead focus on extending the previous analysis. In particular, I experiment with various kernel and smoothing parameters to get a sense of how to visualize the data. This is an important skill because picking the appropriate kernel, bandwidth, and polynomial degree can are crucial components of designing effective algorithms to make accurate predictions.

2.1 Density Estimate of Temperature

To begin, I experiment with different kernels and bandwidths to visualize the data. These choices are important because they reflect important tradeoffs in capturing the overall shape of the data. Specifically, they reflect the "bias-variance" tradeoff, which in statistical learning theory refers to the idea that a model must tradeoff between the two sources of error. A high-bias model may miss relevant relationships between the data, and a high-variance model may capture too much random noise. However, reducing one necessarily increases the other. A kernel-density estimate is a good first-step to make choices about which combination of assumptions have the best chance at producing a useful model. Below I plot 6 different configurations of kernel and bandwidth choice.

The bandwidths I experiment with are:

- Silverman
  - $h = (\frac{4\hat{\sigma}^5}{3n})^{\frac{1}{5}}$
  - Where h is optimal bandwidth, σ is sample standard deviation, and n is sample size
  - Source: Wikipedia
- Sheather-Jones
  - Source: *Density Estimation* by Simon J. Sheather, 2004
- Unbiased Cross-Validation (UCV)
  - Source: *Biased and Unbiased Cross-Validation in Density Estimation* by David W. Scott and George R. Terrell, 1987

The kernels I experiment with are:

- Gaussian (Normal)

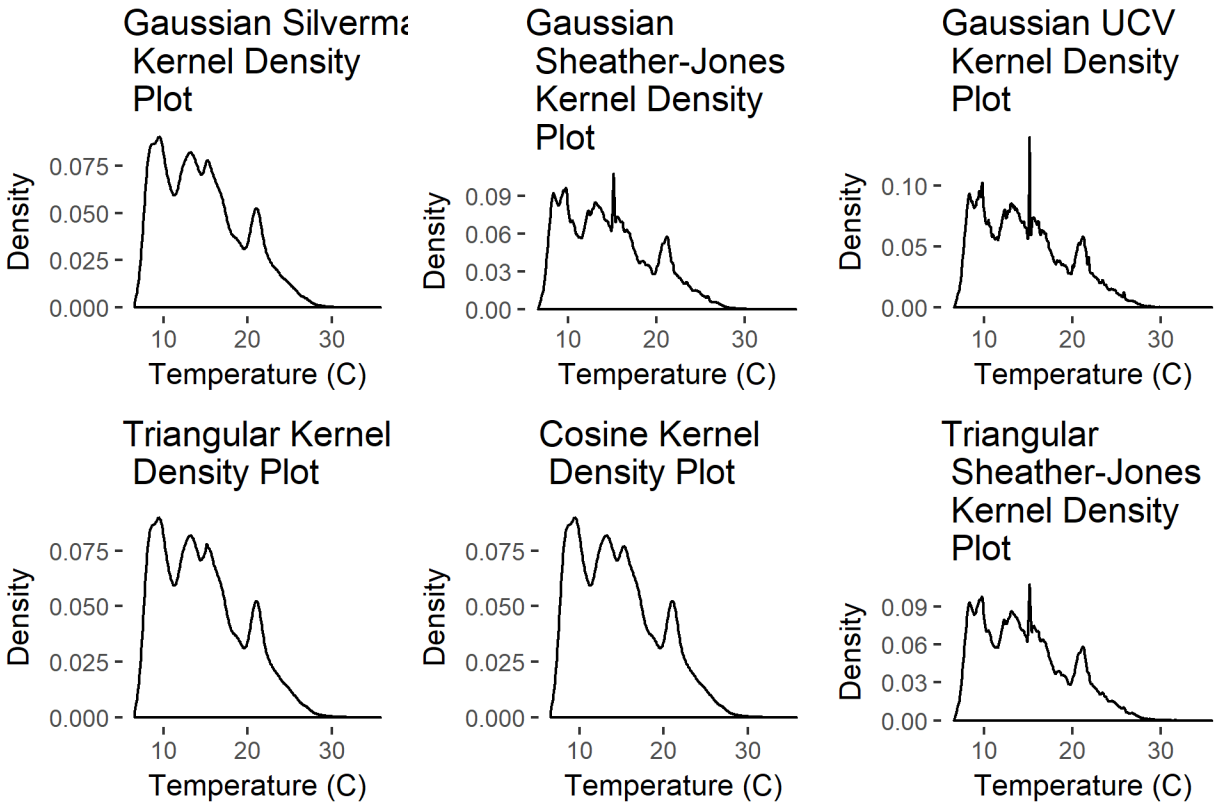
- Triangular
- Cosine

In terms of bandwidths, the Silverman bandwidth is the default, and is derived from "Silverman's rule of thumb" for bandwidth selection. Silverman's rule of thumb minimizes the standard deviation and interquartile range (IQR) to determine the approximate bandwidth. Because of this, it risks missing sharp changes in the data shape that occur within the span of a band. Because of this, I also implemented Sheather-Jones and Unbiased Cross Validation, which both used pairwise binned distances to determine bandwidth. This is a more computationally intensive approach, but has a better chance of capturing small, sharp changes.

These differences in approaches manifest themselves in the plots below. The Gaussian Silverman plot is fairly smooth throughout, with rolling hills. The SJ and UCV estimates show a generally similar shape, with a few more jagged edges, but notably capture a sharp change in the data around 15 degrees Celsius. This is likely because the pairwise-approach is better able to detect a subtle change that occurs at just one x-axis point. There is a "bump" around the same region in the Silverman plot, but it probably smooths the change too much given the information gleaned from the other estiamtes.

Similarly, I also experimented with the kernel choice. The primary motivation here is to select a kernel that will best approximate the underlying probability density function. In this case, I did not see much evidence of the choice of kernel making a difference. All three kernels produced virtually identical plots (illustrated below with a Silverman bandwidth). I suspect that this may be because there are not a large number of separators in the data structure, thus making the kernel choice inconsequential.

Figure 1: Kernel Density Plots



2.2 LOESS on Temperature vs. Humidity

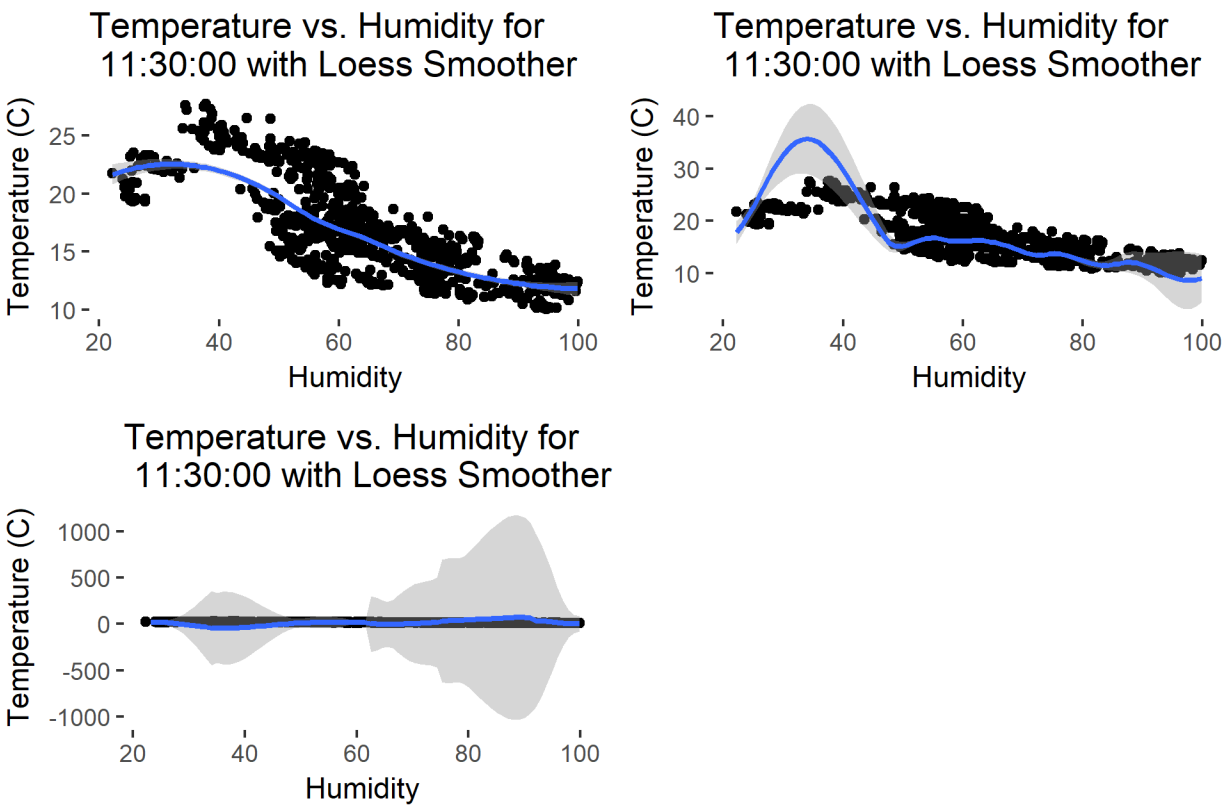
Next, I experiment with how the choice of a model can affect the interpretation of the data. Here, I plot Temperature against Humidity, and then model several different locally weighted scatterplot smoothing

(LOESS) options. I arbitrarily chose to plot observations for time "11:30:00." The default parameters specify a linear smoother, and I also experimented with polynomials of degree 2 and 3.

The default linear model does a fairly good job of fitting most of the data. Between the 40 and 60 percent humidity points however, the model underestimates the data by about 5 degrees Celsius. Attempting to correct this by fitting with a polynomial degree two creates the opposite problem, as this model overshoots the temperature in the 20-40 degree range, and still underestimates the average temperature around 50 degrees Celsius. Wrenching up to polynomial degree three totally eliminates all of the useful variation in the data. Higher degree polynomials exhibit basically the same behavior as a third degree one.

These plots illustrate the bias-variance tradeoff problem quite well. Increasing the polynomial degree lowers the bias of the estimate, but also creates a high variance estimate that makes poor predictions. Conversely, a more biased estimator is a poor fit for the data within certain intervals of the dataframe. Weighing these tradeoffs, I would select the the first plot (LOESS with a linear model) because it would have the best predictive capability. While the second degree polynomial model is not bad, it looks like it would underestimate an out-of-sample prediction.

Figure 2: LOESS and Polynomial Fits



3 The Data

Turning to the linguistic data, I examine the relationships between the answers to the survey questions and geography. The motivation for the survey was to investigate the differences between American etymology for different concepts. The data also provided information about the city, state, zip code, longitude, and latitude for the respondents, making it possible to examine overall differences by geography.

3.1 Data Quality and Cleaning

The data come generally already clean, as columns correspond to variables and rows correspond to observations. The data provided are the question-answer key (all.ans), the subset of questions to actually use (quest.use), the survey responses (lingData), and the responses aggregated into longitude/latitude bins (lingLocation). In terms of additional work, the main tasks are:

1. Compile all of the datasets containing information on the question wording and answer choices into one master dataset, and then subsetting that dataset to the questions that I am examining. In particular, I only looked at questions that dealt with different names for things, as opposed to pronunciations.
2. Rename variable names to allow for matching between the master question-answer dataset and the survey response dataset.
3. Create a new dataset that transforms categorical responses into binary responses. Basically, this meant extending the number of columns in the lingData dataset from 67 to 468, creating a column for each answer choice, with each cell containing a logical (true(1) or false(0)) indicator for whether the respondent chose that particular answer.
4. Tag each observation with a specific region.

The data cleaning code can be viewed in the accompanying RMD file.

3.2 Exploratory Data Analysis

For my first stab at the data, I chose questions 118 and 119 (**Note**: Only the map for question 119 is displayed because of space constraints, but code to generate the map for 118 is provided in the accompanything RMD file), which examined "What Do You Call a Drive-Through Liquor Store" and "What Do You Call Food That You Buy at A Restaurant but then Eat At Home?" respectively. I picked the first question mainly out of personal interest - I never heard of the concept before and was curious about where in the U.S. it existed. I picked the second question because of its similar theme as the first, but my hunch was that to-go food is more universal than to-go alcohol, so I was interested to see if there was still regional variation in the response to that question.

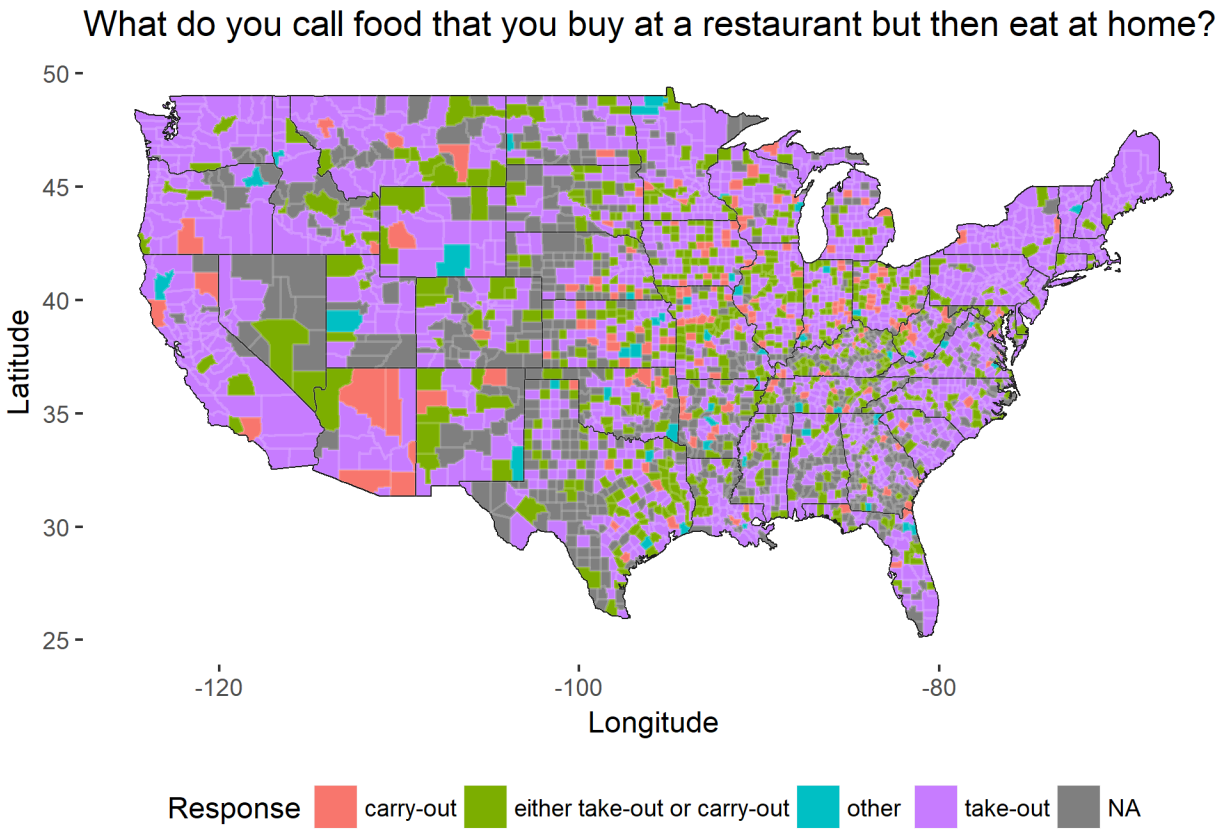
Before continuing, I want to pause to define what I generally mean when I refer to regional areas in the U.S.:

- *New England*: Upstate New York, Massachusetts, Connecticut, and other states in the upper Northeast
- *Mid-Atlantic*: New York City, New Jersey, Philadelphia and its suburbs, Maryland, Washington D.C. and its suburbs
- *South*: The East Coast from Virginia (excluding D.C. suburbs) through Northern Florida, and Alabama, West Virginia, Mississippi etc.
- *Midwest*: Pennsylvania west of Philadelphia, Great Lakes States, Iowa, Missouri
- *Southwest*: Texas, Louisiana, Arizona, New Mexico
- *Mountain*: Colorado, Nevada, Utah, Wyoming, Idaho, Montana, parts of California
- *West*: California Coast
- *Pacific Northwest*: Oregon and Washington State

The graph for the most common response to question 119 (again aggregated at the county level) shows that the U.S. overwhelmingly prefers the term "take-out" to refer to food taken from a restaurant and eaten at home. That being said, a counties in the Midwest and Southwest will use either "carry-out" or "take-out" to refer to the concept.

This initial exploration largely confirmed some of my pre-existing notions about the spatial distribution of American linguistic quirks. Having lived in the New York and San Francisco metropolitan areas, I had never heard of a drive-through liquor store, so I was unsurprised to see that people in those counties similarly expressed that sentiment. Meanwhile, I am unsurprised that the concept exists in other parts of the country. Similarly, I rarely hear the phrase "carry-out" (though I am familiar with it), and was not surprised to see that "take-out" was the most common answer across several geographic areas.

The main drawback of relying on graphs like these is that the data is aggregated to the county level, and only reports the most common answer. This masks some potentially interesting variation, and also skews a reader's perception about the frequency of word usage. For example, if the second-most common word was only 1% less common, that information is not reflected in these plots. Furthermore, any geographic map of the U.S. runs the risk of overstating the popularity of given word choices because of wide variations in population density. For instnace, if there was a phrase that was primarily used on the coasts, but totally unused in the country's interior, an uninformed reader may mistakenly believe that the phrase is not that common. Such an inference would be inappropriate however, as 40% of the U.S. population lives in a county that borders a coastline (NOAA). Similarly, the urban-rural divide can be masked here. Nearly 63% of the U.S. population lives on approximately 3.5% of its land (United States Census Bureau), and just 146 counties house half of the U.S. population (Business Insider). Despite these facts, the county-level maps do not indicate population density. I still choose to proceed with county-level maps as this method creates much clearer looking maps than ones drawn based on the "lingLocation" dataset which creates hundreds of square bins (and therefore distorts the shape of the U.S.), but I flag the context.



4 Dimension Reduction Techniques

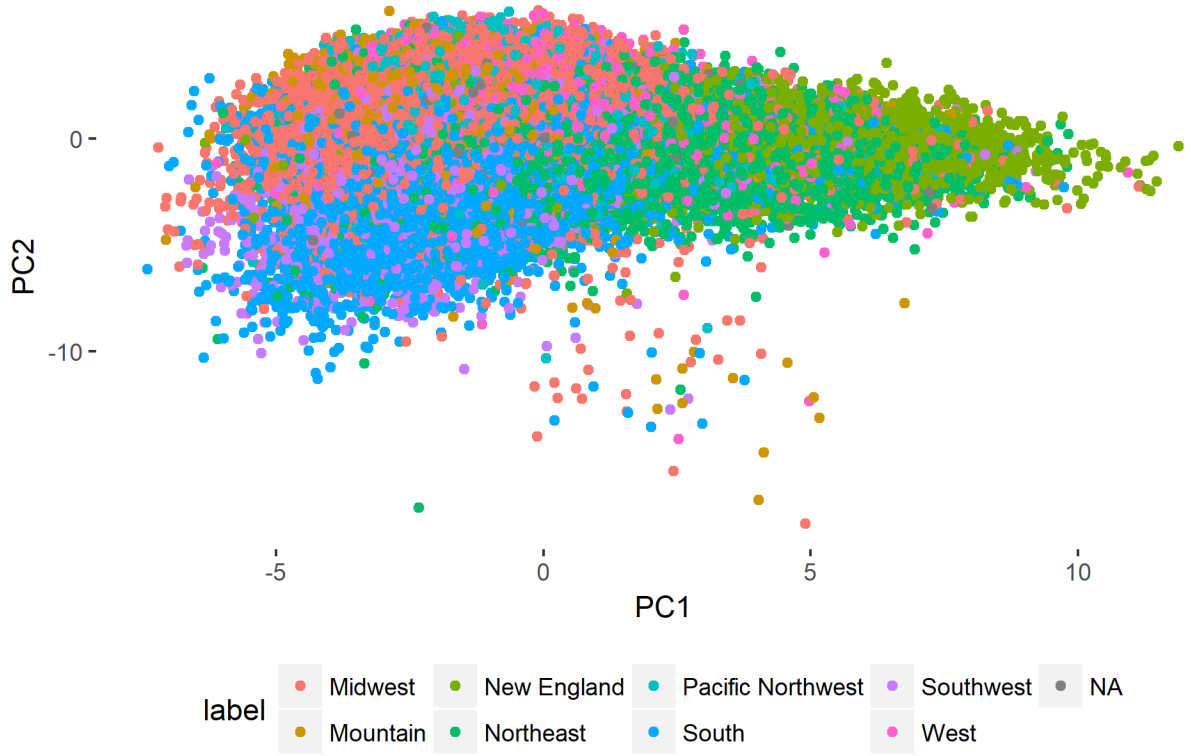
4.1 Principal Component Analysis (PCA)

Because there are 468 features in the binary-converted dataset, I experimented with dimension reduction techniques. Although the graphs above allow a peak at regional variations, there are simply too many questions to vizualize the differences in their answers in a cohesive manner. Because the human eye would struggle to discern patterns from several dozen individual plots, I use dimension reduction techniques to visualize and analyze regional dialectical variation.

First, I performed a PCA on the binarized data, and then plotted the results. Prior to plotting this, I added a variable to the original "lingData" dataset that indicated the observation's geographical region (corresponding with the regions I outlined above). In the plot, I then colored each point by its respective region.

A few clear trends show up with this crude colorization. Most importantly, the graph suggests that there are genuine linguistic differences between the various regions, as clusters naturally show up. The New England and Northeast regions cluster near each other on the righthand side of the graph. Meanwhile, the South seems to be its own distinctive group, with the Southwest blended in. The Midwest and Mountain regions are similarly intermixed. These clusters are still very close to one another (and in fact, overlap), indicating that Americans across the country share linguistic similarities, but there are still enough differences to create noticeable groupings.

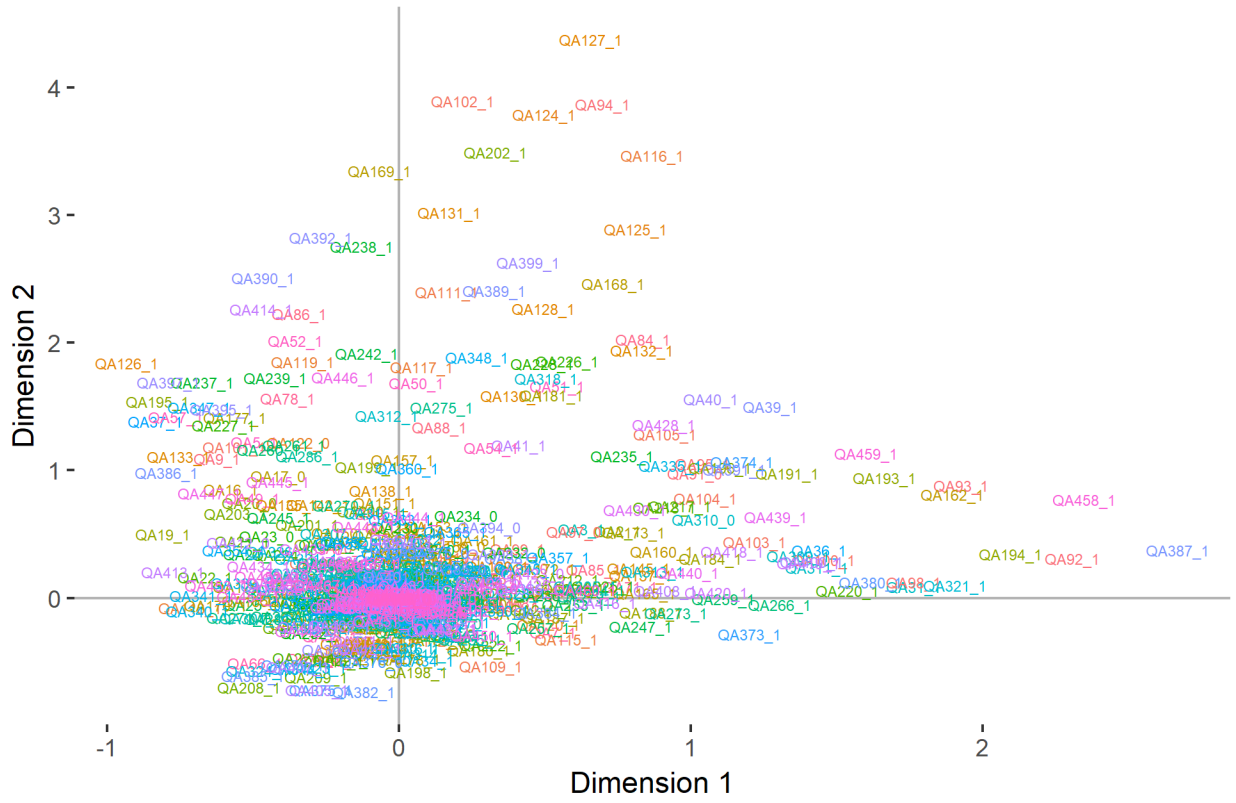
Figure 4: Groupings Based on Region



4.2 Multiple Correspondence Analysis (MCA)

I also performed a Multiple Correspondence Analysis (MCA), which is statistically similar to the PCA, but is better suited for categorical variables (Wikipedia). Similar to the PCA, I performed the MCA on the binarized question-answer data. In this context, the results of this analysis show which particular question-answer combinations are most influential in separating points from each other. The basic intuition here is that each question-answer pair represents a dimension, and the MCA maps the distances in the n-dimensional space to a 2D plane. The axes are scores that measure how influential the question-answer pair is. Here, I show an illustration to give the reader a sense of how many questions are truly influential in separating the respondents. Question-answer pairs that are close to the origin are not particularly good at separating the data, whereas those that are farther away are good seperators.

Figure 5: MCA Plot



On the first dimension, there are a handful of observations with scores above 2, and on the second dimension there are observations with scores higher than 3. I look at these observations to gain insight into which questions distinguish parts of the dataset. I select observations with scores higher than 2 on the first dimension, and higher than 3 on the second dimension, and pulled a handful of observations that were just below the cutoff. I arbitrarily chose these cutoffs to illustrate the most extreme examples; there are a multitude of question-answer pairs with lower scores but were nonetheless influential.

A few interesting trends emerge. Below is a table that lists some of the most influential question-answer pairs that separated the data the most. I exceprt the table here because of space constraints, but the full table is available in the supplemental lab folders.

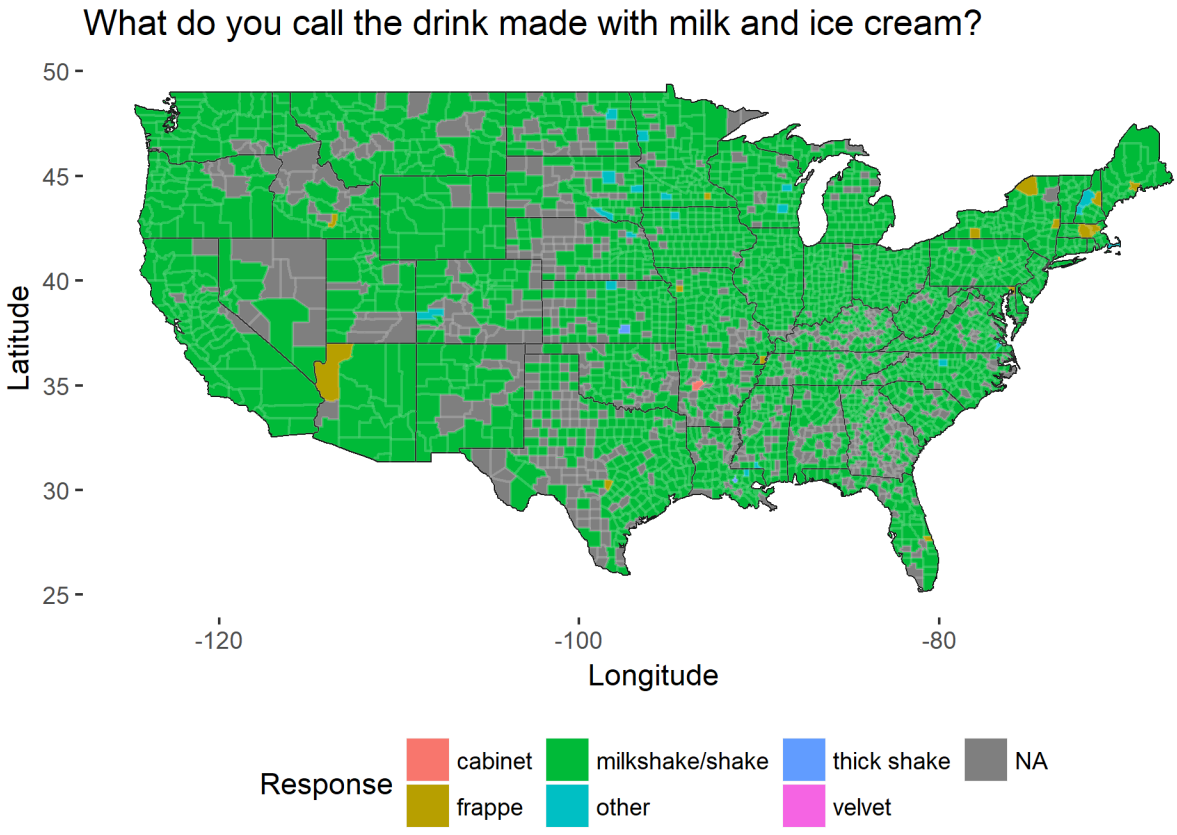
Overall, there were relatively few (12) question-answer pairs that drove the separations in the data. I plot some of the most interesting results that popped out. In general, the extreme MCA values seem to have captured dialectic variations that are isolated to very small areas (in these visualizations, one county). If a respondent did not provide a particularly rare answer to the question, it is easy to separate them from the handful of respondents who do provide the rare answer. Again, this table only reflects the most extreme examples. There are a multitude of observations with scores greater than 1 on both axes that also explain quite a bit of the variation, and likely cover larger areas.

Going through each of these examples, I was surprised by Questions 63 and 64 (the term for a milkshake and the term for a shopping cart respectively). I never heard any other names for a milkshake, so seeing the handful of counties that use terms like “frappe,” “cabinet,” and “velvet” would make it very easy to distinguish them. Similarly, the term “shopping varriage” seems to only exist in parts of Massachusetts and Rhode Island. Question 64 (not pictured, available in markdown file) about the name for cold-cut sandwiches surprised me less as I lived in New Jersey and am familiar with the fierce divisions between people from North Jersey preferring “sub” and South Jersey preferring “hoagie,” though I was surprised by some of the other less common phrases that were limited to smaller and less populuous areas. Generally, these extreme MCA scores demonstrated how ultra-localized dialectic patterns provided critical information for separating

Table 1: Influential Question-Answer Pairs

Question Number	Question-Answer Pair	Question Wording	Answer
Q063	QA92	What do you call the drink made with milk and ice cream?	frappe
Q063	QA93	What do you call the drink made with milk and ice cream?	cabinet
Q063	QA94	What do you call the drink made with milk and ice cream?	velvet
Q064	QA102	What do you call the long sandwich that contains cold cuts, lettuce, and so on?	bomber
Q064	QA98	What do you call the long sandwich that contains cold cuts, lettuce, and so on?	grinder
Q075	QA190	What do you call the wheeled contraption in which you carry groceries at the supermarket?	shopping cart
Q075	QA193	What do you call the wheeled contraption in which you carry groceries at the supermarket?	shopping carriage
Q075	QA194	What do you call the wheeled contraption in which you carry groceries at the supermarket?	carriage

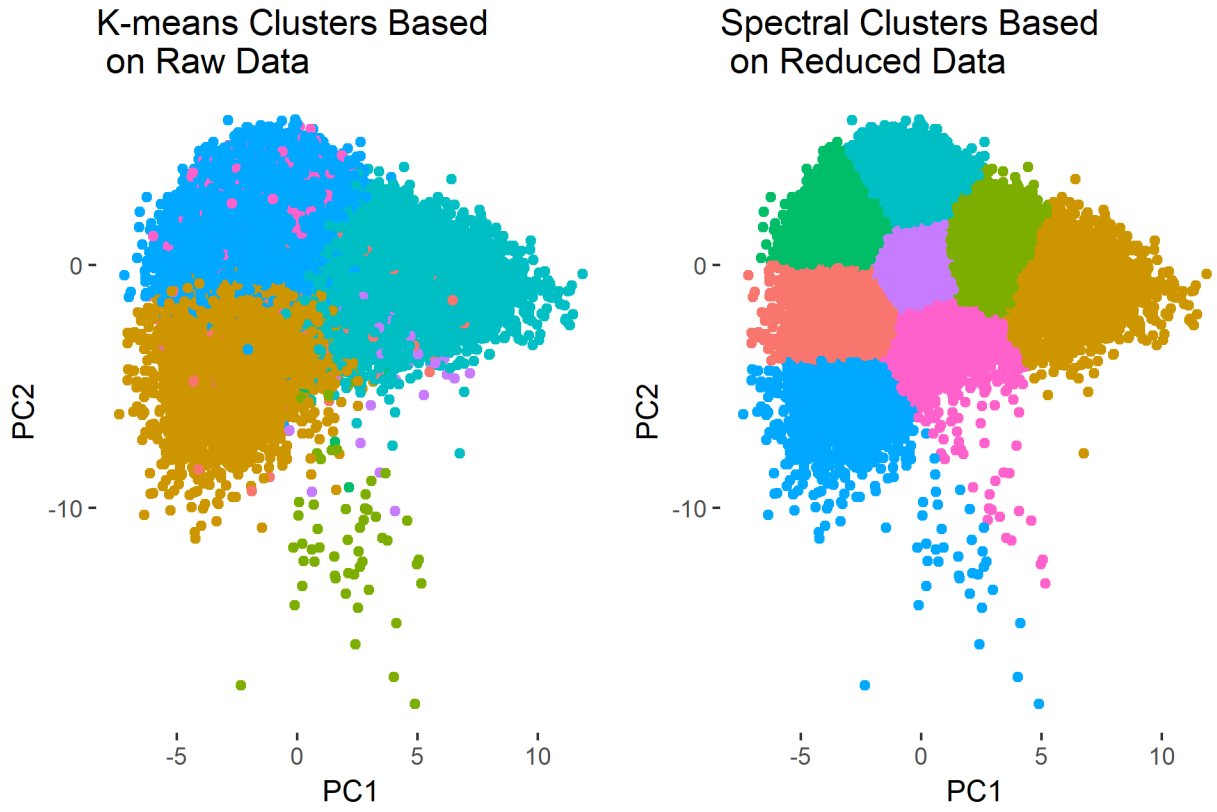
the geography of the respondents.



4.3 Clustering

I now turn to clustering methods to gain more insight into the data. Here I experiment with K-means and spectral clustering to evaluate each method's ability to discern distinct clusters. For both of these plots, I specified 8 centers. This choice arises from the self-defined continental U.S. regions above (New England, Mid-Atlantic, South, Southwest, Midwest, Mountain, West, Pacific Northwest). After running this simulation several times, the k-means clustering generates 3-4 distinct clusters, whereas the spectral clustering consistently produces 8 distinct clusters. Furthermore, the k-means clusters are not as cleanly separated as the spectral clusters, as there is some overlapping between the clusters and a handful of non-clustered points that show up within a given cluster. Meanwhile, the divisions between the spectral clusters are clean, and most of the points are tightly concentrated around their respective clusters. This suggests that the spectral clustering did a better job of identifying the geographic clusters that I suspect exist in the data. That being said, there are some points that are not especially close to their clusters, so 8 centers may still not be optimal.

Figure 7: K-Means and Spectral Clusters



5 Stability of Findings to Perturbation

In this section, I assess the stability of the clustering done in the previous section. In simple language, stability basically means that multiple runs of the simultion produce similar-looking clusters. This is an important step in the process because unstable clusters would suggest that the clustering algorithm does not do a consistent job finding groups. In turn, instability suggests that the groups are not all that strong, and no one simulation's resulting clusters can be trusted as the basis for drawing inferences or making predictions.

First, I assess the stability of the k-means clustering. As suggested above, I found this method to be quite unstable. As illustrated below, it produces either between 3 and 5 clear clusters when specified to find 8, and these clusters have significant overlap. The spectral clustering perturbation (not pictured) performs much

better as it consistently creates eight distinct clusters that are separated from each other. However, the exact shape of these clusters varies a bit, which suggests to me that eight is probably not the optimal number. Doubling the number of spectral clusters to 16 considerably improved the performance, as it produces 16 clusters with fairly consistent shapes.

Figure 8: 8 K-Means Cluster Simulations

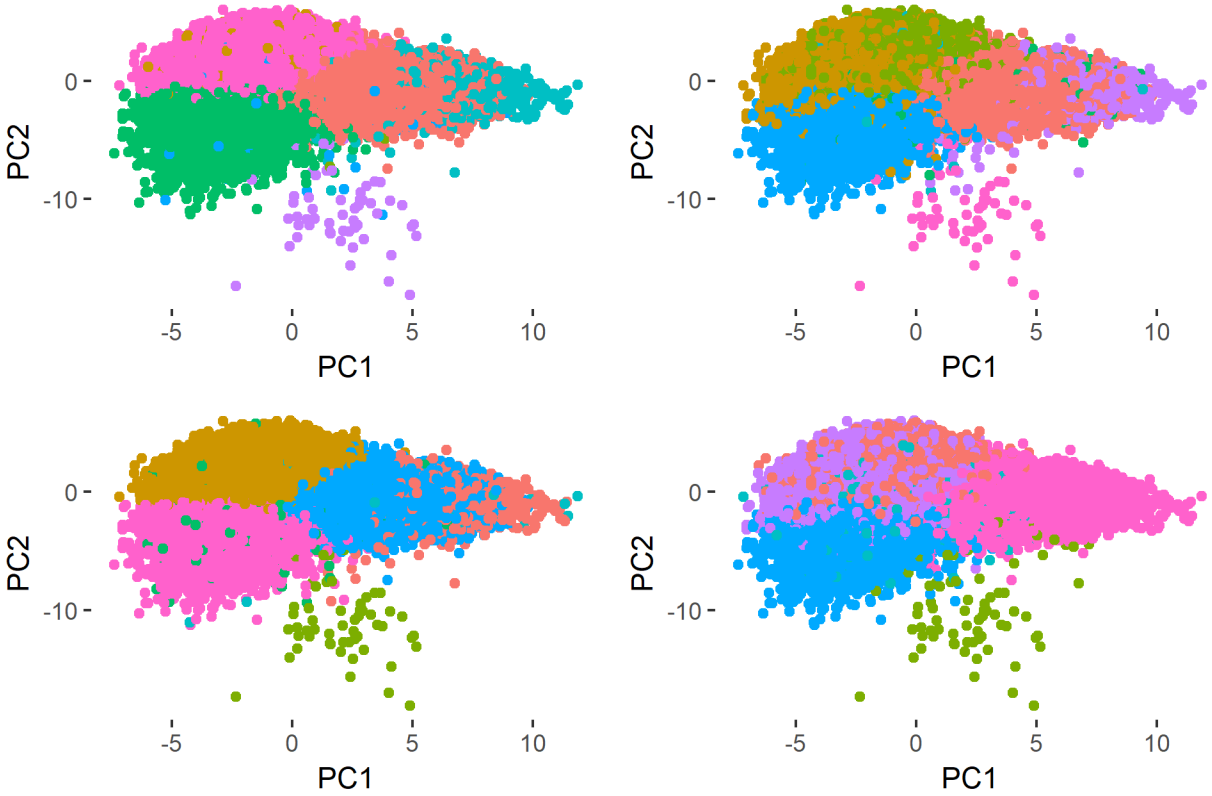
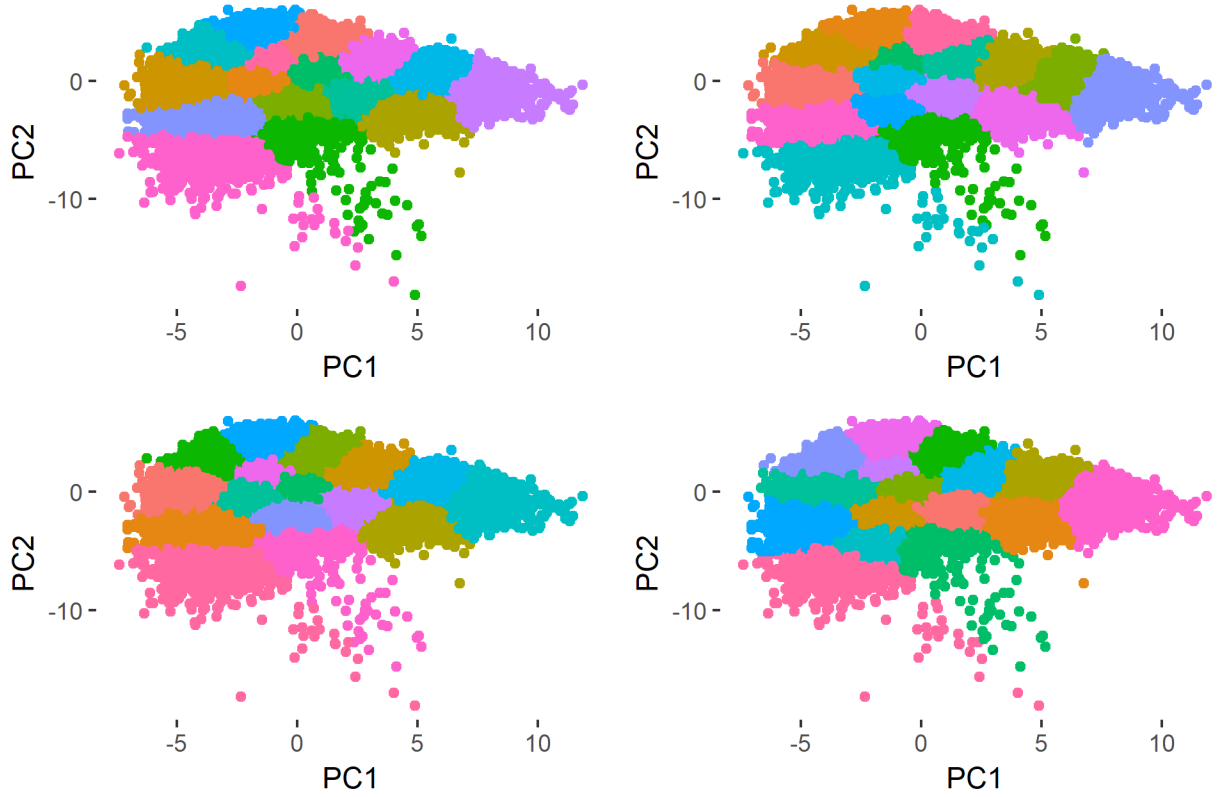


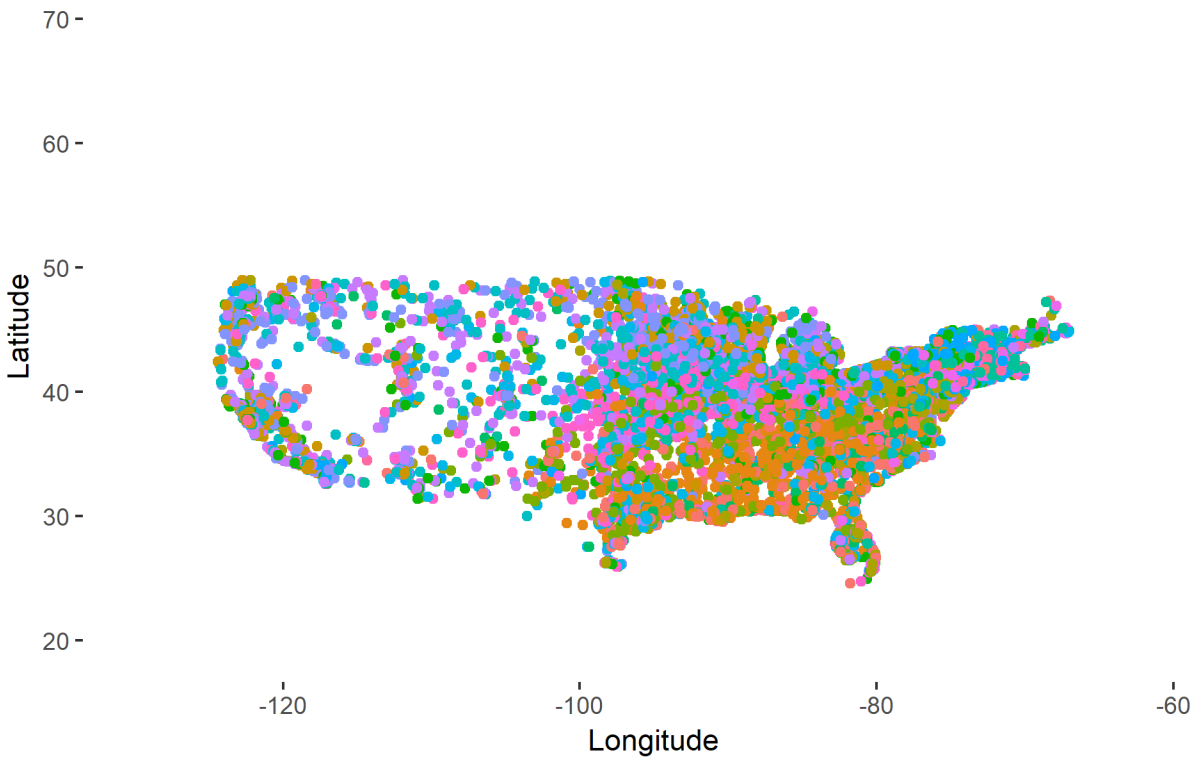
Figure 9: 16 Spectral Clusters Simulation



5.1 Geographic Mapping

To check this result, I project the 16 spectral clusters onto a map of the continental United States. The clusters take on abnormal shapes due to population dispersion, but they nonetheless emerge. There does seem to be a genuine distinction between the Northeast, South, and West, but the clusters within these regions are a bit muddled.

Figure 10: 16 Spectral Clusters Projected Onto Map of U.S.



6 Conclusion

To conclude, I will briefly sum up my most interesting findings. My main finding was identifying the most extreme contributors to separating the data, along with providing a method for identifying other contributers. The MCA analysis is a powerful tool because it leverages the straightforward intuition that the more unique an answer was, the better it is for distinguishing those respondents from others. Using PCA analysis and clustering, I was able to connect these aggregate differences to geography. Although based on toying with the clustering parameters, my hypothesized regions seem to be underinclusive, I still saw strong evidence that New England/the Mid-Atlantic are linguistically distinct from the South and Midwest, and the West/Pacific Northwest lies in between them. That being said, the clusters are still close to each other. Given this fact and the results from the MCA, I conclude that Americans share deep linguistic similarities, and the differences are largely driven by highly localized offshoots of common phrases.