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

October 07, 2021

1 Introduction

To many from the United States, the existence of different dialects of English comes as no surprise. However, this linguistic variation raises many questions. Are there any non-linguistic factors (cultural, social, etc.) that are correlated with this variation? To what extent is geographic location an indicator of which label one uses for a particular concept? More generally, without prior knowledge of geographic linguistic variation in the United States, can this variation be detected solely from data-driven methods?

In this report, I will explore these questions to varying degrees using the results from a Dialect Survey conducted by linguist Bert Vaux. More specifically, I will first describe Vaux's dataset in more detail, explain my methods for cleaning the data for analysis, and provide initial exploratory plots displaying just how regional certain conceptual labels can be. Then, after experimenting with dimension reduction methods to render the dataset more manageable, I will use unsupervised learning methods, like clustering, to see if the regional variations observed for specific concepts extend to more general lexical habits. Finally, I will analyze the stability of my clustering results, indicating the validity of my conclusions.

2 The Data

The dataset used in this analysis was collected by Harvard linguist Bert Vaux as part of his Dialect Survey, conducted between 2000 and 2005. The survey asked respondents to provide their geographic location and answer a series of questions about how they pronounce certain sounds or which labels they use for different concepts. By mapping these responses to the user-provided locations, one can get a sense of the spatial-dependence of linguistic variation in the United States.

The dataset consists of 47431 respondents, each providing location information such as their city, state, and ZIP code of origin. With the help of former STAT 215A GSIs, we also have latitude and longitude information for each respondent, extracted from their provided ZIP codes. All in all, the original dataset contains each individual's responses to 67 questions, where each question aims to detect lexical, as opposed to phonetic, differences. Each question is represented by a single nominal categorical variable, where different positive integer values represent different answer choices. It is important to note that individuals' responses could be null, in that they may not have answered a question.

With such information for a large number of individuals, one may be able to detect geographic patterns in language usage. For instance, maybe one label for a concept is frequently used among respondents whose reported location is the American Southeast, but almost never used among respondents from the Northeast. Beyond exploratory analyses, such data is also ripe for unsupervised clustering methods, where features isolated in the data might coincide with geographic variables. If this is the case, we would have evidence of a relationship between geography and lexical habits.

2.1 Data Cleaning

While relatively clean already, this is data from a survey that was publicly available online, so great caution must be taken to ensure the validity of the responses provided. First, we notice that there are several columns with missing values: 537 city entries, 3 state entries, and 1020 latitude and longitude entries. Among the entries with missing latitude and longitude values, the vast majority of provided ZIP codes were not valid USPS ZIP codes. This was primarily checked through the `zipcodeR` package and the USPS website. While

many had city values provided, it was extremely difficult to attribute latitude and longitude information to the cities, since the names varied in case, spacing, and spelling. To be sure, state information was also provided for many of these entries, but this is not fine-scaled enough to reliably attribute latitude and longitude coordinates. So, all entries for which there were missing latitude and longitude data were dropped from the set.

Next, we checked the 3 observations with missing state values. Among these, the cities were listed as 'nowhere', 'nottingham', and 'RananaISRAEL'. At the very least, none of the cities matched the provided ZIP codes, so these observations were deemed untrustworthy and dropped from the dataset. After this, we notice that quite a few state observations have their states listed as 'XX', which clearly should be treated as a missing value. However, all of these observations actually did have legitimate ZIP codes, so using the `zipcodeR` package, we filled in the correct states for each ZIP code.

Next, we decided to drop the provided 'city' column in the dataset, since, as mentioned above, the city names provided had inconsistent cases, inconsistent spacing, and sometimes included state information in the city name as well. We deemed having a clean city column not worth the effort to parse and clean it, and dropping observations based on inconsistent city names to be dangerous since this is survey data, so some degree of messiness is expected. We decided that ZIP codes and the latitude and longitude information would be more reliable indicators of location.

At first, our strategy to verify the provided locations was to see if the provided ZIP codes matched with the provided states, since the city column was too difficult to parse and the latitude and longitude values seemed largely based on the ZIP codes. In this case, we would have two pieces of corroborating evidence to verify the respondents' locations. However, a large proportion of the provided ZIP codes are not valid USPS ZIP codes. In fact, using the `zipcodeR` package and checking quite a few ZIP codes on the USPS website, it seems like only about 10,000 observations have legitimate ZIP codes. We did not feel comfortable wiping out a whole 75% of the dataset just based on this one criteria. So, we opt to keep these observations, and trust the latitude and longitude information provided by the GSI's. While a possibly-shaky judgment call, we will attempt to justify it by performing a mini-reality check in the Exploratory Data Analysis section, ensuring the cleaned data still coincides with our prior knowledge about linguistic patterns in the United States.

Finally, we exclude all observations from Hawaii and Alaska, both for visual purposes (much less hassle to create a map) and because only about 200 of the total responses come from these states. Additionally, we were much more interested in studying the relationship between linguistic variation and geography in the contiguous United States.

A few more checks were also made to ensure the data was as sound as possible. For instance, it seems that all of the answer choice entries are actual choices for those questions. However, unlike other datasets for which outliers could be removed through domain knowledge or comparing to some external reality, it is much more difficult to justify outlier removal here. Who's to say that someone in the Pacific Northwest doesn't say "y'all"? Removing outliers in this way runs the risk of introducing confirmation bias, and so we opted not to remove geographic outliers.

In the final dataset, we have 46,242 observations, with 72 columns: an ID column uniquely identifying each respondent, a state column representing their provided state information, a ZIP column representing their provided ZIP codes, latitude and longitude columns, and the responses for 67 questions, categorically encoded as described above.

2.2 Exploratory Data Analysis

In this analysis, we focus on the spatial distribution of responses to questions 50, 80, and 65. Figures 1, 2, and 3 show the questions and the prevalence of certain answers. To make these figures, we first recoded the categorical variables in the cleaned dataset such that each question consisted of a binary vector, whose length was determined by how many answer choices that question had. So, if an individual responded a particular choice, instead of coding this as an integer, a 1 was filled into the column of this vector corresponding to this choice, and zeroes were filled in everywhere else. On this one-hot coded data, we rounded each latitude and longitude to the nearest degree, and grouped the observations by latitude-longitude pairings. For each

group, we summed up the number of participants and added this as a column, and then summed up the binary vectors and normalized by the number of participants. So, for each latitude-longitude pairing, we have the proportion of individuals that responded a particular choice for each question. Additionally, we exclude any latitude-longitude pairings for which the number of respondents is less than two. Two individuals are not enough to represent a whole latitude-longitude box, and thus should not be plotted at this fine scale. However, we nonetheless keep these data in the original dataset, since these responses are part of coarser-scale geographic areas that can be considered in the upcoming analysis.

Figure 1 shows the relative prevalence of using "you guys" versus "y'all" to refer to a group of two or more people. Clearly, the use of "y'all" is much more prevalent in the Southeastern United States compared to the rest of the country. Conversely, the use of 'you guys' is not so prevalent in the Southeast as compared to the rest of the country. Here, we see an example of linguistic variation that falls along somewhat stark geographic lines.

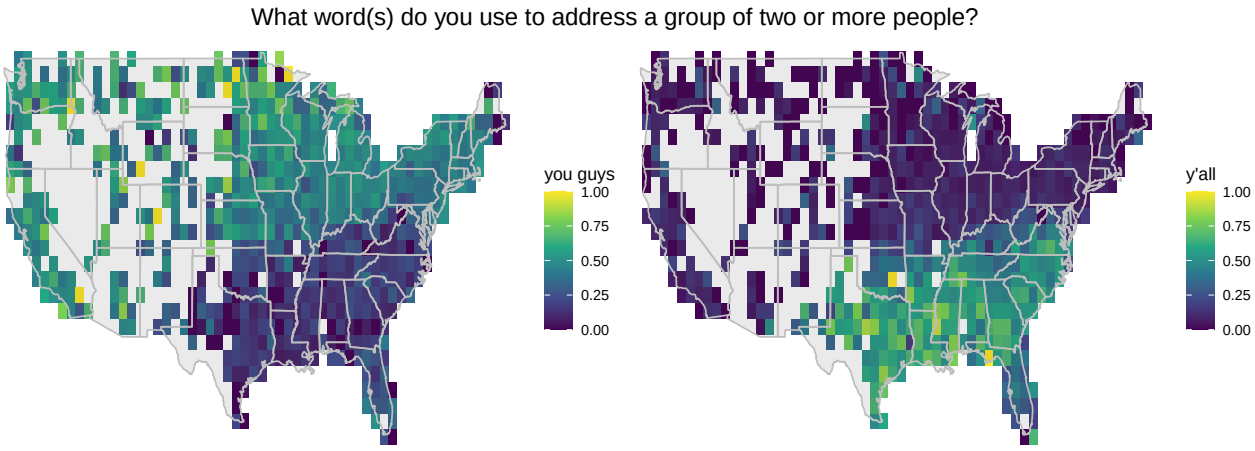

Figure 1: Proportion of respondents that anwered you guys (left) and yall (right) for each latitude-longitude box. Clearly, yall is more favored in the Southeast, whereas you guys is more favored in the rest of the country.

Figure 2 shows a similar breakdown that isolates the Southeast from the rest of the country, but also isolates the Northeast. We clearly see that the use of 'sunshower' (to describe the phenomenon of rain falling while the sun shines) is more popular in New England as well as portions of the Midwest. While there are some points in the West for which this answer seems popular, we caution that, at this fine scale, these may be outliers that are not representative of the residents of the latitude-longitude code box. However, we also see that 'the devil is beating his wife' is more prevalent in the Southeast. Once again, we see that labels for this concept have different geographic distributions.

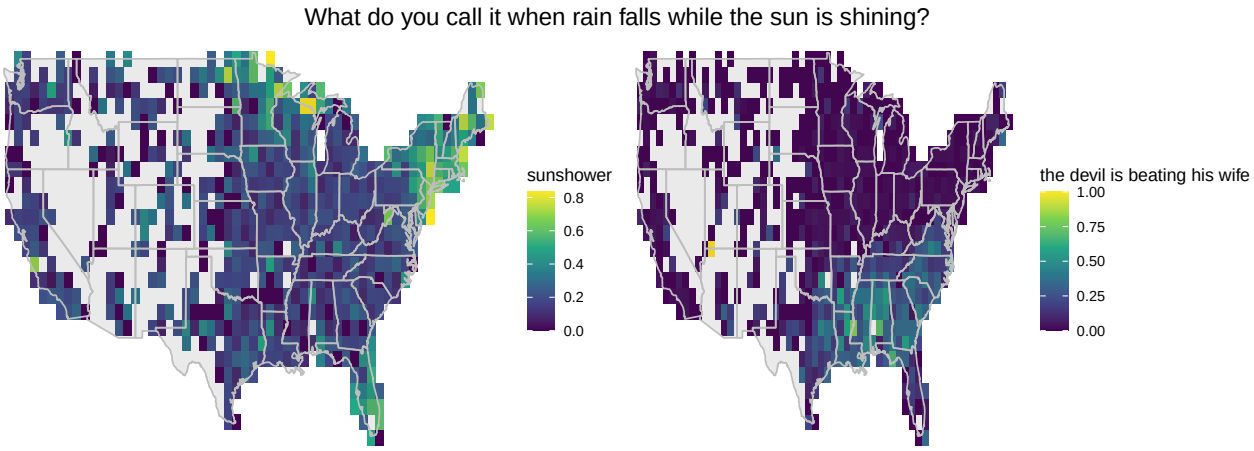

Figure 2: The use of sunshower (left) and the devil is beating his wife (right) mark both the Northeast and Southeast, respectively.

Finally, we analyze another dichotomy in responses: whether or not you call the Lampyridae variant prevalent in the United States a 'firefly' or a 'lightning bug'. In Figure 3, we see that the use of 'lightning bug' is much more prevalent in the Mid-Atlantic, Midwest, and the South. The term is less prevalent out West. However, the term 'firefly' is much more commonly-used in the Western United States, and much less commonly used where 'lightning bug' is prevalent. Once again, it seems like the spatial dropoff is quite abrupt in the usage of both of these terms.

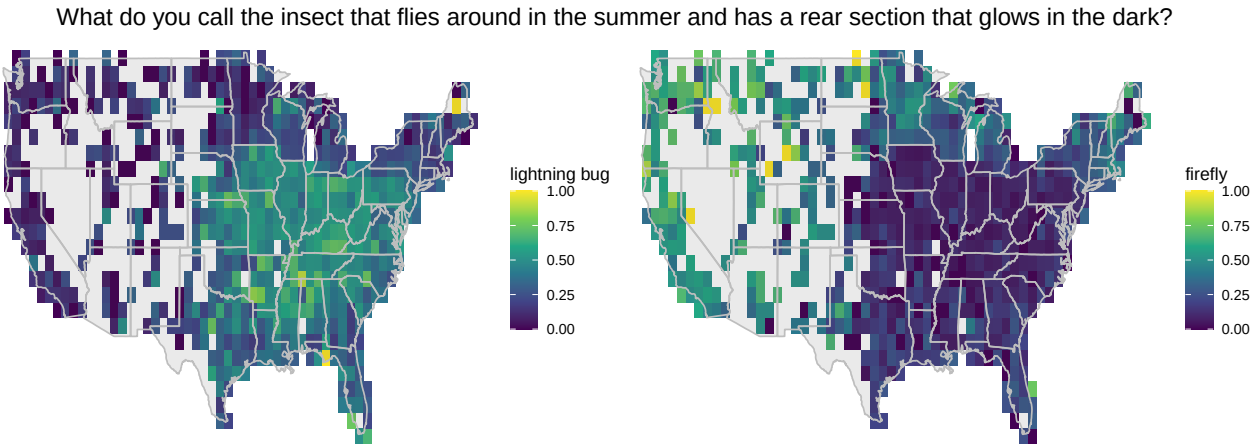

Figure 3: The use of lightning bug is much more common in the American Heartland and Mid-Atlantic and South, whereas firefly is more prevalent out West.

The above figures clearly show that there is a relationship between lexicon and geography for a few specific concepts. Circling back to the data cleaning section, these figures provide a nice reality check for my cleaning decisions. It is well-known that "y'all" is commonly used in the South, and less so in the North. In the North, "you guys" or "you all" (not shown) is generally more used. Additionally, "the devil is beating his wife" is known to be common in the South.

Taking the results in these figures together, we see that responses to these questions are not necessarily independent, and may be related through geography. For instance, if someone remarks that 'the devil is beating his wife', then this person would likely also say 'y'all' frequently. Similarly, if someone says 'firefly', then that person might also say 'you guys' instead of 'y'all'. Of course, these relationships are not certain, but this exploratory analysis demonstrates that different geographical regions preferentially respond certain

choices for certain questions in this dataset. Taking answers to all 67 questions as a representation of general lexical habits for each individual, it would be interesting to see if groupings of individuals by similar lexical habits reflect geographic groupings as well.

3 Dimension Reduction Methods

After performing the exploratory analysis, we next attempt to implement unsupervised clustering methods to see if geographic regions or responses to individual questions can be captured. However, the dimensionality of our one-hot coded data is 468, quite high and prohibitive to efficient computation. So, we will first try some dimensionality reduction techniques and display any relevant results.

However, we should first address why it was necessary to one-hot code our variables before performing dimension reduction techniques, rather than just treating every variable as categorical with nonnegative integer labels. Take, for example, PCA, which attempts to find linear combinations the variables in a dataset that maximize the variance captured. The results of PCA thus depend upon the variances of the individual variables. However, if the variables were categorically-encoded as opposed to one-hot encoded, the variances would have no meaning. The numeric labels for each answer choice are arbitrary, so an equally-valid recoding of the answer choices would yield different variance calculations. So, it is better to one-hot code the variables since a binary vector for a particular question's answer choice has a more interpretable variance.

3.1 PCA

First, we performed PCA on the one-hot coded dataset to achieve dimensionality reduction. First, we decide on whether or not we should center or scale the data. Intuitively, we should center the data because all of our data points are nonnegative, and more specifically, occupy vertices of a 468-dimensional hypercube. In finding principal components, we wish to find projections of the data onto lines through the origin, but since the data is concentrated away from the origin, it seems like it would be difficult for the principal components to pick up the variation in the dataset. So, we decide to center the data.

Next, we decide to scale the data because, while each variable has similar units (between 0 and 1), it's possible that a question answer with many 0's and many 1's only has many 0's because those other people simply did not answer the question. So, the variation in this variable is simply measuring the variability in who answered the question in the first place, as opposed to measuring the more interesting variability in who answered this particular choice among other choices. So, in order to clamp down on the former source of variation, we also decide to scale the dataset. This decision also makes sense in the context of the screeplots, which are presented in the left and middle plots of Figure 4.

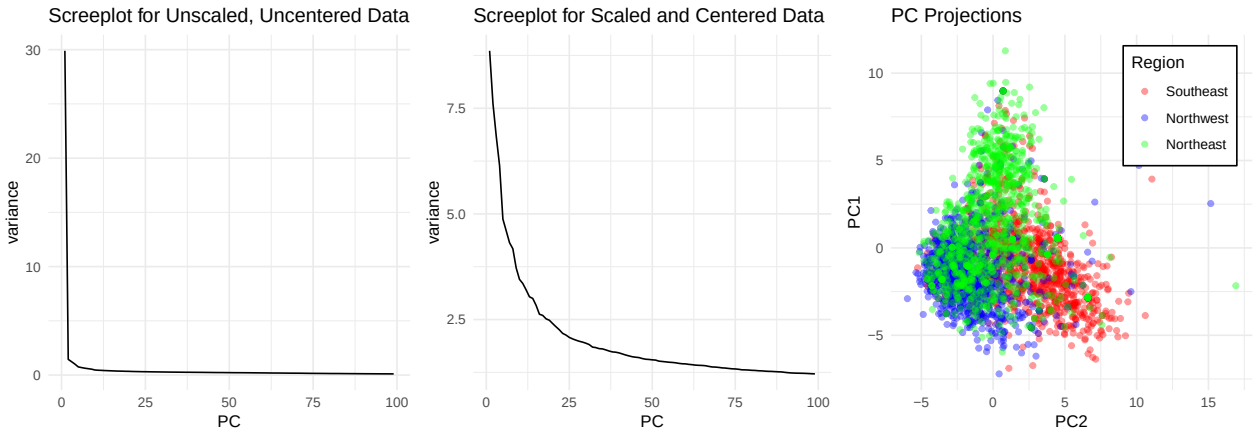

Figure 4: Eigenvalue information for both the unscaled, uncentered data, and scaled and centered data

In the unscaled and uncentered data, we see that the screeplot has an extremely abrupt dropoff in variance

explained from the first few principal components onward, meaning that the first principal component is putting weight on some variables with extremely high variances, as opposed to the other principal components. Analyzing further, we see that choice 'a' for Question 63 has the highest loading along the first principal component, but 96% of answerers chose this choice. The variation captured by this variable does not reflect variation among the answer choices, since the answer was fairly uncontroversial. Looking at the screeplot for the scaled and centered data, this abrupt drop off has disappeared. Therefore, we opt to scale and center the data for PCA. As for choosing the number of principal components, there is a huge amount of variation in this dataset from a variety of factors that will be difficult to almost completely capture with less than 100 principal components. So, we are satisfied with selecting the first 100 principal components, which roughly describe 45% of the variation in the dataset.

The right-hand plot in Figure 4 shows that, despite the reduction in dimension, a projection along the first few principal components nonetheless captures some of the geographic variation we observed in the exploratory analysis. More specifically, when projecting onto the first two principal components and randomly sampling 1000 points from each region for plotting, we see that the points are somewhat clustered by geographic region (Northeast, Northwest, Southeast; Southwest excluded to declutter plot). For instance, points from the Northeast seem to have higher PC1 scores, but PC2 scores close to zero. In contrast, observations from the Southeast seem to have higher scores on PC2 but more negative scores on PC1. Finally, the observations from the Northwest seem to have negative scores on both principal components. Of course, these divisions are not abrupt in anyway and there is significant overlap. However, this is nonetheless an indication that our dimension-reduced data captures some of the geographic variation we have already observed. Note: to define these regions, used a latitude of 37 and a longitude of -90. So, the Southeast consists of observations whose latitudes are less than 37 and longitudes greater than -90, for instance.

3.2 NMF

We also looked into NMF as a means to reduce the dimensionality of the dataset. We opt not to center the data, since NMF requires the input data to be nonnegative, and centering will introduce negative values. Scaling is a bit of a tougher decision. In PCA, it was necessary to scale since the resulting principal components, especially the first few, are so dependent on the variances of the individual variables. This in turn impacts which principal components are selected in the dimension-reduced dataset. However, in NMF, a set of principal patterns are recovered, but there is no obvious constraint that the principal patterns must explain the variation of the data in a similar fashion. The principal patterns are simply there to detect features such that linear combinations of these features minimize the difference between the original data under the Frobenius norm. So, in the absence of an obvious variance constraint like in PCA, we opt not to scale the data.

To choose the number of principal patterns, we will use the method of staNMF, in which we select the number of principal patterns corresponding to the least instability. The method was implemented as follows: for each k satisfying $1 \leq k \leq 10$, perform NMF $n = 3$ times, yielding 3 dictionary matrices. For each pair of dictionary matrices D_1 and D_2, calculate the dissimilarity score (C is the cross-correlation matrix between D_1 and D_2):

$$diss(D_1, D_2) = \frac{1}{2K}(2K - \sum_j \max_i C_{ij} - \sum_i \max_j C_{ij})$$

Then, we average the dissimilarity scores for this k. We repeat for all other values of k, and plot the average dissimilarities for each k. The results are shown on the right-hand side of Figure 5.

It seems like there is a minimum around 9 principal patterns. So, we opt to choose 9 principal patterns. It should be noted that, given additional computational resources, we would be able to try more values of k and average more dissimilarity scores together to achieve more robust and convincing results. However, we will be satisfied with this plot. We can also make some plots to see what the data looks like in this reduced space, as we did with PCA. Figure 5 shows some of these results.

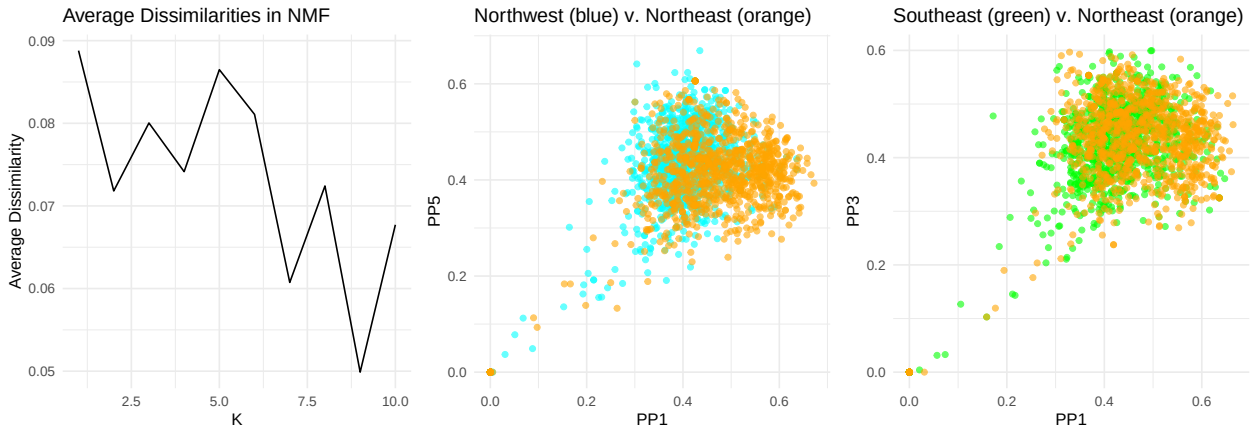

Figure 5: (left) staNMF results show 9 is a good choice of principal patterns (middle, right) Projections of the data onto the principal patterns may capture geographic variation, but it is slight

To produce the middle and right plots in Figure 5, we randomly sampled 1000 observations from each region (regional bounds defined above) and plotted the observations from the Northwest/Northeast projected onto the plane spanned by Principal Pattern 1 and Principal Pattern 5 (left) and observations from the Southeast/Northeast projected onto the plane spanned by Principal Pattern 1 and Principal Pattern 3. We see that these different projections of the data somewhat separate out these geographic regions (although there is certainly overlap). So, we see that, like in PCA, this dimension-reduced data might still capture some of the geographic differences we observed in the exploratory analysis.

4 Clustering

Once we reduced the dimension of the data, we were able to try out a few clustering methods to see if observations could be grouped into general linguistic patterns and whether these groups reflect geography. We first attempt to implement spectral clustering, whereby we perform k-means clustering on the PCA dimension-reduced dataset. Subsequently, we use NMF. Decisions on how many clusters to choose, as well as discussion and interpretation of results, are provided.

4.1 K-Means

First, we performed k-means on the PCA dimension-reduced dataset. To choose the number of clusters k, we first attempted the silhouette method, whereby for a sequence of k values, we perform k-means, calculate a pairwise distance matrix for each observation in the PCA dimension-reduced space, and use this information to calculate average silhouette scores. However, the size of the dataset was computationally prohibitive, especially when calculating pairwise distances between more than 40,000 observations. Subsampling was attempted to make these calculations more feasible, but the optimal silhouette scores were highly unstable to different subsamplings. So, this was abandoned.

Instead, we opted to use a simple elbow method, whereby for $2 \leq k \leq 8$ clusters, we performed k-means and plotted the total within-cluster sum of squares as a function of k. An elbow in the plot should indicate an optimal choice of k. The elbow plot is provided in the left-hand plot of Figure 6. It seems that the elbow in this plot is around 6 clusters, so we will choose $k = 6$. The middle and right-hand plots in Figure 6 show the clustering results projected onto two planes: the PC2-PC3 and PC1-PC3 planes.

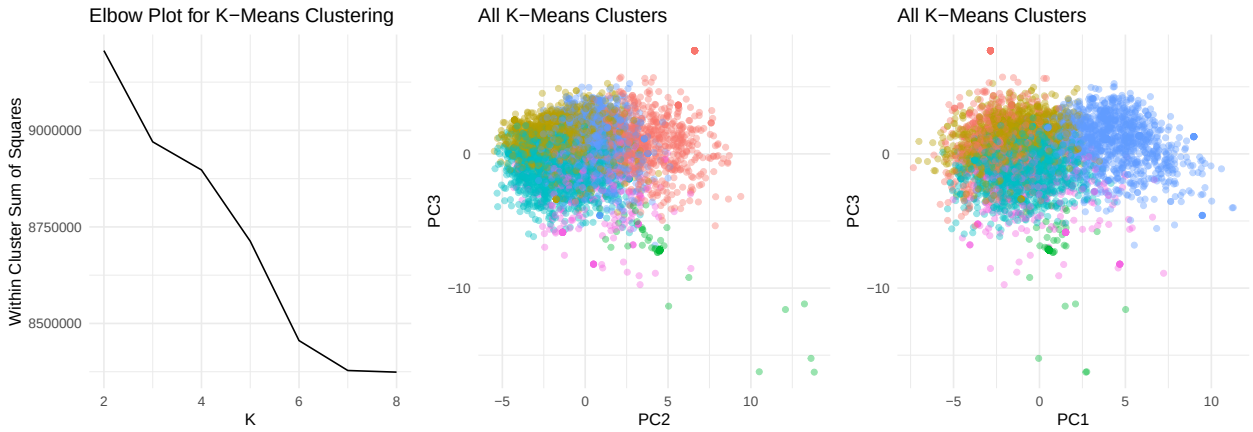

Figure 6: (left) 6 clusters seems optimal using the elbow method (middle, right) clusters are distinct from one another in different projections on the principal components

In Figure 6, we see that k-means has effectively clustered all of the survey respondents into distinct clusters. Furthermore, it seems these clusters, while not located far from one another in each projection, do not significantly overlap. Indeed, in the PC2-PC3 plane, it seems there is significant overlap between cluster 5 and the other clusters, but projecting onto the PC1-PC3 plane shows that cluster 5 is actually fairly distinct from the rest.

We then tried to see if any of the lexical clusters corresponded to certain geographic regions or answers to certain questions. Informed by the exploratory analysis, where we saw significant geographic variation for a few questions between the Northeast and the Southeast, for instance, we attempted to link one of the clusters with either region. We also wanted to find a link between a cluster and a common region-dependent response to the questions from the exploratory analysis, primarily to bolster any claim that the lexical clusters correspond to certain geographies. Figure 7 shows some results.

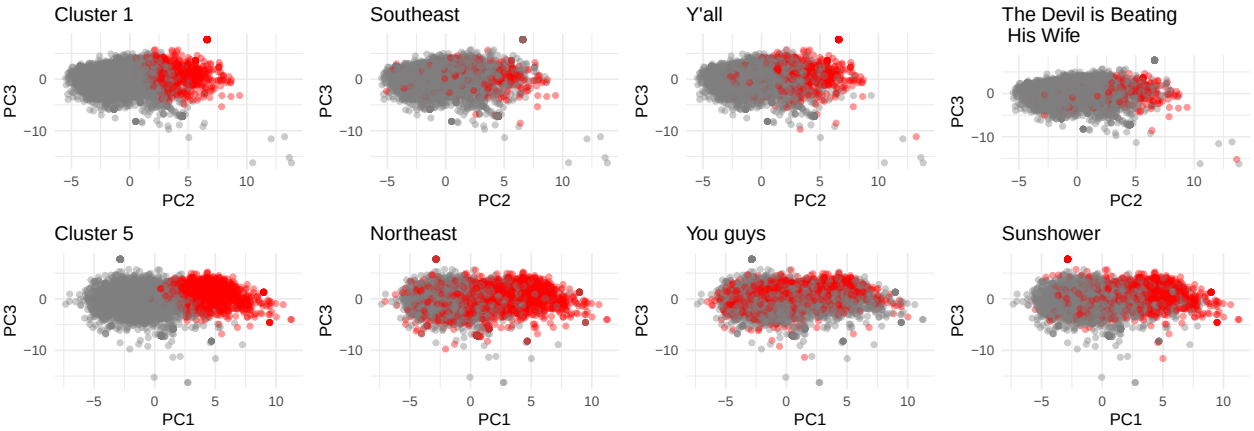

Figure 7: Top: In the PC2-PC3 plane, Cluster 1 overlaps significantly with respondents from the Southeast, and respondents who say yall and the devil is beating his wife. Bottom: In the PC1-PC3 plane, Cluster 5 largely consists of individuals from the Northeast who say you guys and sunshower.

Looking at the top row of Figure 7, the group of respondents that the k-means algorithm has decided to cluster into Cluster 1 seem to be predominantly from the Southeast, say “y’all” to refer to a group of people, and say “the devil is beating his wife” when the sun is out while it is raining. Of course, there is not perfect overlap: Cluster 1 contains respondents that are neither from the Southeast nor say “y’all”, for instance. However, it does seem like this lexical cluster has significant representation from the Southeast.

Looking at the bottom row of Figure 7, we see that respondents in Cluster 5 are predominantly located in the Northeast and say 'sunshower' when it is raining while it is sunny outside. Many of those in Cluster 5 do say 'you guys', but 'you guys' is also popularly used outside of the Northeast as noted in the exploratory analysis section, so this is not the best marker of a Northeasterner. Nonetheless, it seems that the lexical grouping produced by this method has significant representation from the Northeast. More specifically, Cluster 5 has significant overlap with 'Sunshower', which Figure 2 showed is popular in New England. Since the points in Cluster 5 seem to be a subset of the points in my defined Northeast region, Cluster 5 probably more specifically refers to New England.

These clusters also make sense in the context of my dimension-reduction strategy. Figure 4 in the Dimension Reduction section showed that, rightfully so, PCA captured some of the variation that fell along geographic lines. So, it makes sense that a clustering algorithm like k-means on this PCA data would also detect this geographic variation.

There are quite a few advantages to this clustering method. First, it is very computationally feasible. Performing PCA on this dataset was not particularly expensive, and performing k-means on this dimension-reduced dataset was also quite fast. This method also has the advantage of using data projected onto axes representing large proportions of variation in the original data. This enables k-means to detect meaningful clusters that could succinctly reflect the variation observed in the original dataset. Disadvantages primarily have to do with PCA not being designed for binary data. While our results were somewhat interpretable, caution should be taken in applying this method to other binary settings. Also, the different starting points in the k-means algorithm can yield different clustering results, as well as different optimal k from the elbow method. This will be discussed further in the stability analysis section.

4.2 NMF

The last clustering method we attempted was NMF. Taking the NMF-reduced data from the prior section, we can cluster each observation by the principal pattern number the observation gives the most weight to. In our case, we chose 9 principal patterns via staNMF, so this will be our choice of k for this clustering procedure. For ease of comparison with the last clustering method, we will present our NMF results with a similar suite of plots. Figure 8 shows each respondent projected into two principal pattern subspaces, colored by their cluster membership.

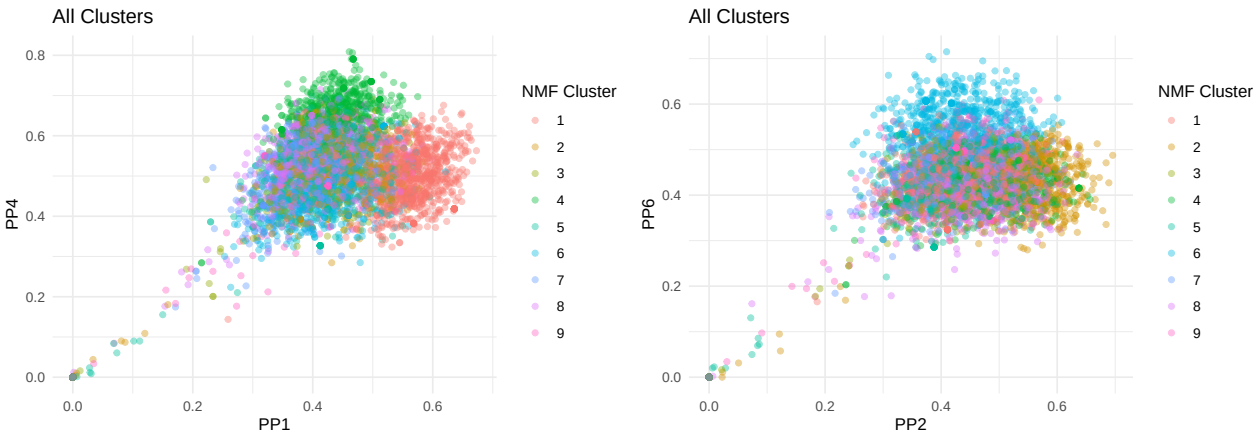

Figure 8: Clusters assigned via NMF are separated from one another in different projections of the data onto the principal patterns

In Figure 8, we see that the clusters assigned via NMF are somewhat distinct from one another in 2D projections of the data onto various principal patterns. For instance, while cluster 4 or cluster 5 in the PP1-PP4 projection seems mostly distinct from the other clusters, there is nonetheless some overlap with the others. It is possible that other projections of the data will display more separation, but at least some degree

of separation is visible. As with k-means, we then attempted to match some of these clusters with geographic regions or common region-specific responses to questions. As before, the goal is to see if these lexical clusters correspond to certain geogrraphic regions. Figure 9 shows some of these results.

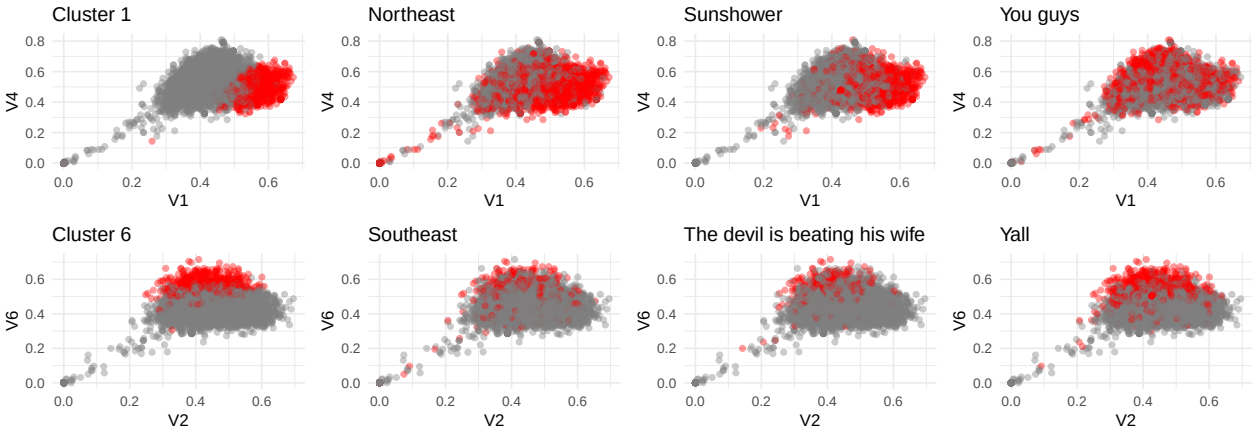

Figure 9: Top row: In the PP1-PP4 plane, individuals in cluster 1 tend to be from the Northeast, say sunshower, and say you guys. Bottom row: In the PP2-PP6 plane, many individuals in cluster 6 are from the Southeast, say the devil is beating his wife, and say yall.

Like with k-means, we do not see a perfect correspondence of some cluster with a geographically defined region (at least with our self-defined regions). However, we do see certain clusters that overlap significantly with geographic regions and have individuals that share lexical quirks with a particular region. For instance, we see that in the top row, Cluster 1 contains individuals from the Northeast that like to say 'sunshower', and a significant number that like to say 'you guys'. So, this cluster seems to represent Northeastern speech patterns. Similarly, in the bottom row, we see that individuals in cluster 6 tend to say "y'all", many say the "the devil is beating his wife", and many are also from the Southeast. So, it seems like this cluster represents, at least somewhat, the Southern speech pattern. Of course, there are points in the cluster that are not geographically Southeastern by our definition, but playing around with definitions of geographic regions could yield a cleaner correspondence.

The mathematical model behind this dimension reduction strategy makes sense for these clusters. In NMF, we simply find some dictionary of features for which the data can be suitably approximated by some linear combination of these features. While there is no explicit variance maximizing criteria like in PCA, features of the data could nonetheless consist of quirky regional responses to questions, like saying "the devil is beating his wife". In this way, observations could be clustered by the degree to which this feature applies to them, allowing for clustering on the basis of lexical patterns. This in turn can reflect geography, since many lexical quirks are regional.

NMF proved to be quite amenable to clustering. The dimension-reduction and clustering process was baked into one procedure, so it is efficient in this regard. Also, it does not seem like the results are as sensitive to scaling as PCA, averting a potentially bad judgment call. However, an issue might be that NMF is not necessarily guaranteed to pick up on features that explain much variation in the data. In this way, our dimension-reduction and thus clustering results could just be based on noise, and not anything actually important in the data.

5 Stability Check

I find my most interesting clustering result was the coincidence between Cluster 1 in Figure 7 and the number of people who say 'y'all'. In that example, this one cluster captures more than 60% of the respondents who say 'y'all', among the 5000 randomly selected individuals that were plotted. The next cluster captures 15% of respondents who say y'all. So, I will define this clustering result to be robust if, subject to perturbations,

there still exists one cluster which dominates all other clusters in capturing respondents who say 'y'all'. Since Figure 7 used a random sampling of the already-clustered data to show this result, a natural starting point for a stability check is to repeat this sampling many times and see if the distribution of respondents who say 'y'all' across the clusters is similar. The left-hand column of Figure 10 shows these results. Next, since the k-means EM algorithm begins with random cluster assignments, we also perform k-means a number of times on the same dataset and display the distribution of participants who say "y'all" across the different clusters. This is shown in the middle column of Figure 10. Finally, we can further check for stability by performing k-means on random subsamples of the data a number of times and checking the distribution of "y'all" respondents across clusters. We opt for random 95% subsamples of the original data. These results are shown in the right-hand column of Figure 10.

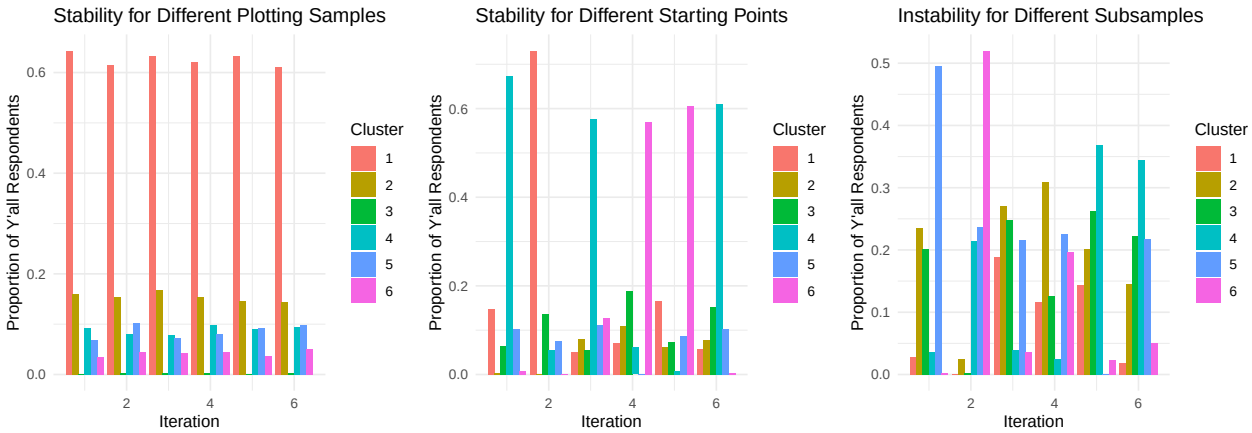

Figure 10: (left) Cluster 1 dominates yall respondents over different random samples of already clustered data; (middle) For different k-means starting points, there is a dominant cluster that captures most yall respondents; (right) Dominant cluster comes and goes when performing k-means on different 95% subsamples of the original data.

It seems our clustering observation is stable to random selections of 5000 points post-clustering, as shown in the left-hand plot of Figure 10. Looking at the middle plot, it seems our clustering observation is stable to 6 different starting points of the k-means EM algorithm. Note that multiple iterations of k-means do not necessarily preserve cluster labels. All that matters for our definition of stability is that there is one cluster that by far captures most of the "y'all" respondents. However, in the right-hand plot, we see that, for 6 different random 95% subsamples of the original data on which k-means was performed, there is sometimes a cluster that contains most of the "y'all" respondents, and sometimes not. So, we conclude our clustering observation is not stable to this perturbation.

6 Conclusion

This report fits well into the three realms of data science. We have data that was collected between 2000 and 2005 that is supposed to represent some reality (the linguistic landscape of the United States), we used unsupervised learning techniques and algorithms to cluster the data and see if any lexical groupings could be mapped to geography. Using these clusters and a correspondence between cluster and geographic region, we might be able to predict someone's home region from which cluster their lexical habits place them in. However, we do not think this data is useful for future decision-making purposes. This data was collected when the Internet was still fledgling, with not nearly as much communication as there is today. Since so much person-to-person contact occurs over the Internet these days, lexical variations along geographic lines might be blurred as people are constantly exposed to unfamiliar ways of speaking. Thus, we also think my clusters would not also be particularly useful for future data. Case-in-point: "y'all" was a distinctly Southern phenomenon in this dataset, but from our experience, it is currently prevalent throughout the United States. An external reality check to our k-means clustering results could be that we randomly sample

American individuals and ask them the same 67 questions with the same choices. Then, we could project their responses onto my PCA dimension-reduced space, and assign their observations to clusters using the existing k-means object trained on this older data. However, our clusters should be taken with a grain of salt, since our attribution of a single lexical cluster to the Southeast was unstable to perturbations in the dataset.

Given more time, we would explore the instability issue presented in Figure 10 with more robust clustering methods. We might also have explored more answers to different survey questions and their geographic dependence, as well as other notions of geographic location of the respondents. For instance, we considered which quadrant a respondent lived in as their geographic location, but a future analysis could try to use the home states of the respondents, or some other finer and more informative notion of location. We also might consider the phonetic questions in the dataset and see if this more complete assessment of dialect leads to more conclusive geographic clustering results. However, as Nerbonne and Kretzschmar note, dialects are notoriously hard to map to geography.

7 Academic Integrity Statement

I attest that the work in this project is my own work, and that any other student or individual I consulted with have been acknowledged, and any source I used has also been acknowledged. Academic research honesty is a bedrock principle in academia because it drives us to give credit where credit is due, to acknowledge the work of our peers, and to drive the pursuit of knowledge forward.

8 Bibliography

Wu, S. et al., "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks". PNAS, vol. 113, no. 16.

Nerbonne, J. and Kretzschmar W., "Introducing Computational Techniques in Dialectometry". Computers and the Humanities, vol. 37.

Nerbonne, J. and Kretzchmar W., "Progress in Dialectometry: Towards Explanation". Literary and Linguistic Computing, vol. 21.

I also used Stackexchange for advice on how to accomplish certain coding tasks in R