Lab 2 - Linguistic Survey Stat 215A, Fall 2018

10/5/2018

1 Introduction

This lab is an analysis of a survey of American dialects. It looks at the relationship between nicknames for maternal grandfather and grandmother, as well as more general dialect communities and their demarcations. The results found below mostly corroborate a lot of traditional thinking around American dialects, though there are some surprising results given my own experience as someone who grew up in the American South, as well as more generally with the clustering of dialects we witness through k-means.

2 The Data

The data comes from a dialect survey conducted by Bert Vaux in 2003. The survey polled 30,788 respondents on 122 different questions having to to do with American dialects. These questions ranged from how your pronounced common words like aunt to how you refer to specialized or regionally specific things and events (e.g. the night before Halloween.) In total, the survey aims to present a fairly comprehensive view of American English and how the language differs wildly and in very interesting ways across the landmass.

2.1 Data quality and cleaning

Compared to our last project, the data quality here was much higher, with fewer missing values and inconsistencies. When looking at the summary of ling_data, for example, only 1,023 observations had missing values. I began by looking at these observations. When I isolated the 1,023 observations, I saw that the only values they were missing were for latitude and longitude. My initial plan was to provide the correct lat and long for these cities, but upon looking at missing value zip codes within the overall data set, I realized that no observation from many of these zip codes had the correct lat and long. Additionally, I saw many discrepancies in the observations that had missing lat and long, such as fictional or overly broad city and town names (e.g. "Chicagoland.") Given how messy data was, and given that it was a small fraction of the overall data, I decided to remove them. I initially considered removing all incorrect zip code observations – even the ones that had lat and long – before realizing the discrepancy arose from the first number in the zip code being a 0. Once I discovered this, I decided to retain all of these values.

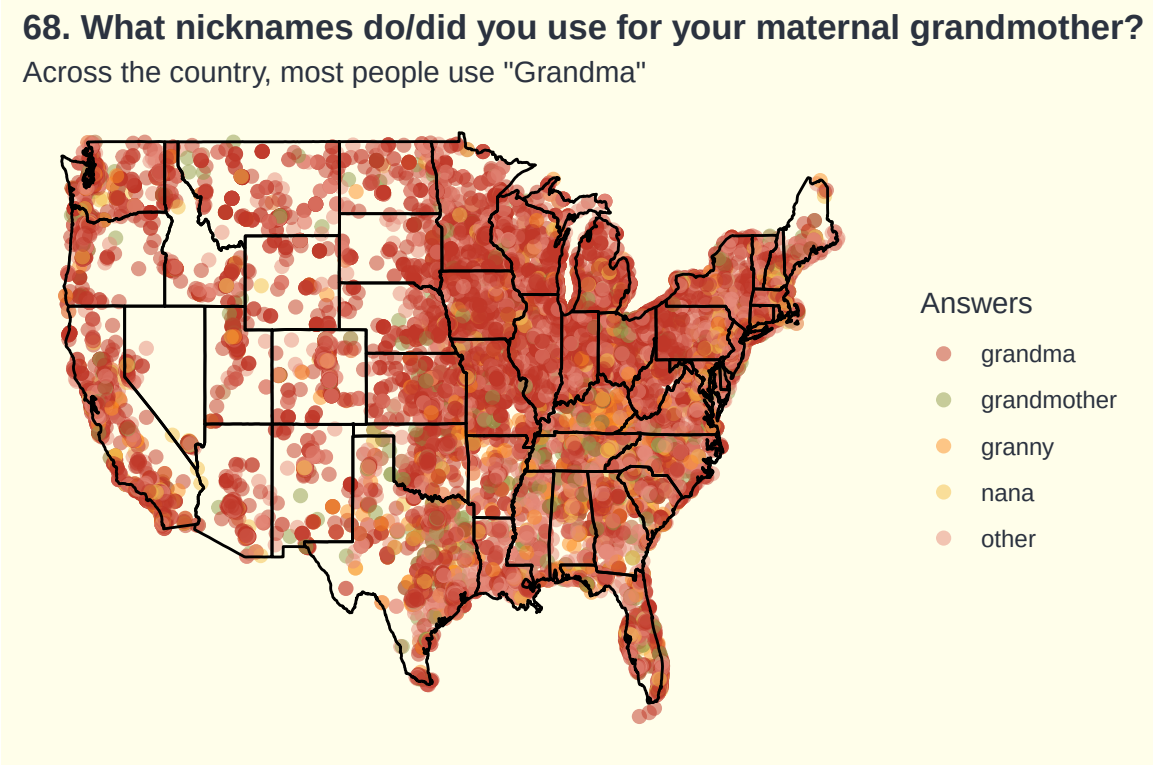
Next, I began looking for inaccurate values. I first noticed the inaccuracy for the State variable - there were over 80 different values, even though there should only be 51 (the states and the District of Columbia.) Most of these values only appear once or twice in the data, suggesting a misspelling. The only value that had a large amount of observations was 'XX.' Upon further investigation, 'XX' appears to be a stand-in for an unknown state as many of these observations were also missing their city value. With unlimited time, these observations could likely be cleaned and corrected, but given their small aggregate amount, I decided to delete them.

After this cleaning, there remained 537 observations that did not have a city listed. Because these observations had latitude and longitude values, I kept them because I did not think that the lack of a city value would impact analysis. After beginning my exploratory data analysis, I did decide to remove the observations from Alaska and Hawaii because of how it impacted graphing.

2.2 Exploratory Data Analysis

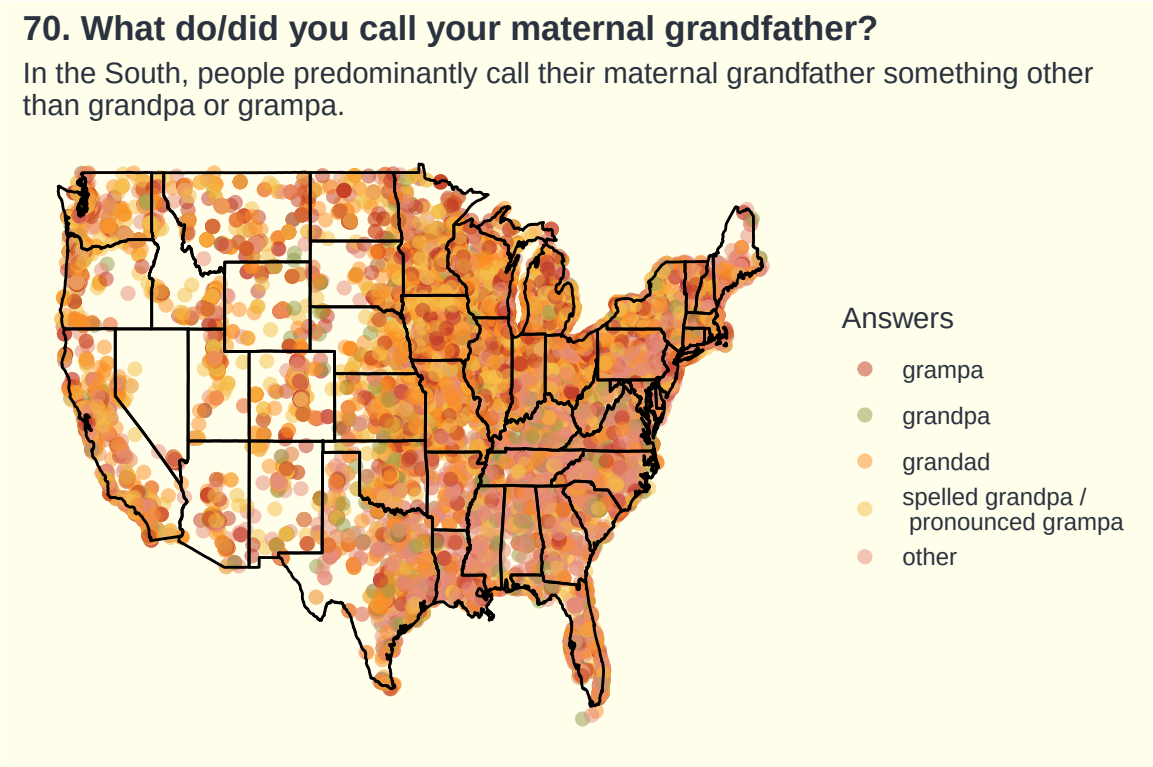
To choose the questions to analyze, I looked inwardly. As an American who spent the first 18 years of my life in South Carolina but has since lived in Massachusetts, Colorado, and now California, I have experienced a wide swath of regional dialects. Throughout the non-southern regions, people have always pointed out my use of Southern dialect, e.g. my use of "y'all." I kept my own quirks in mind as I read through the list of questions. While doing so, the questions for maternal grandmother and maternal grandfather jumped out at me. I've always called my mom's mother "Grandmother." She took this very seriously and in fact chided me once for calling her "Grandma." At the same time, I called my mom's father "Granddaddy." This always struck me as weird – why wouldn't I call him "Grandfather" if I called his wife "Grandmother." The dialect survey questions on names for maternal grandparents presented a good opportunity to see if people in my region also practiced this quirk.

First, I looked at Question 68, which asks, "What nicknames do/did you use for your maternal grandmother?" Though the question has seven total answers, I only included the top five to make the graph a little clearer.



As the graph above shows, "Grandma" is definitely the most popular name countrywide for your maternal grandmother. In terms of numbers, "Grandma" claims a simple majority with 50.67% of respondents giving that answer, with "Other" receiving 30.79%, "Nana" receiving 5.77%, "Grandmother" receiving 4.78%, and "Granny" receiving 3.77%. Though "Grandma" appears to do well in every region, it has a particular stronghold in the Midwest and Great Lakes region, where it dominates all other responses. Additionally, "Other" was a popular answer in the South and Gulf region, as well on parts of the Eastern Seaboard. Outside of these two clear trends, the rest are more varied. "Granny" has a stronghold in parts of Kentucky but, like "Nana," is sprinkled throughout the map. "Grandmother" is also sprinkled throughout but does not appear to identify strongly with any region, including South Carolina or any part of the South.

Next, I looked at Question 70, which was the companion question to Question 68. It asked, "What do/did you call your maternal grandfather?" Like Question 68, I only included the top five to make the graph a little clearer.

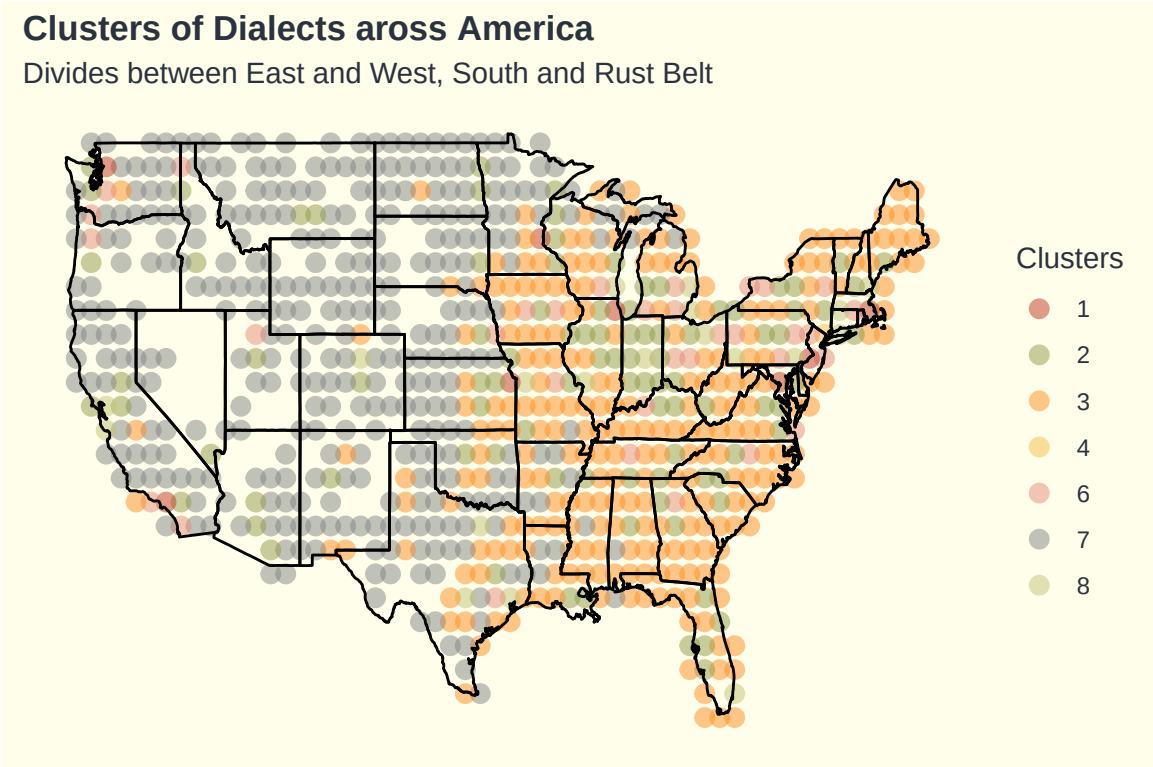


For maternal grandfather, the most common response was "Other," with 32.26% of respondents. "Spelled grandpa/pronounced grampa" came in second, with 25.90% of respondents, while "Grandpa" had 21.05%. 'Grampa' had 13.86%, and "Granddad" had 5.07%. Q70, like Q68,has a clear trend in the South. "Other" dominates this area – possibly (but likely not) fitting with my childhood use of "Granddaddy" – even though each response makes an appearance in the area. Though not as pronounced as the trend in Q68, the Midwest also has its own unique trend, with both "Granddad" and "Spelled grandpa" appearing across the region. "Grampa" also appears in the Midwest, specifically in the north of Michigan and a few pockets of Iowa and Indiana. Overall, the story of maternal grandfather resembles that of maternal grandmother: The South and Midwest each demonstrate their own answers, while all five answers appear across the country to some extent. In both questions, a large percentage of respondents said "Other," suggesting either a diverse set of names for a grandparents or an inadequate amount of options on the original survey.

3 Clustering

Next, I began clustering to see if I could identify similar groups of general dialects across the country. I decided to use ling_location instead of ling_data because the data had already aggregated total answers for each possible answer in each grid, meaning I didn't need to spend time encoding the categorical data. I did need to normalize the dataset, but that was very easy with the scale function.

After normalization, I ran the k-means algorithm with eight centroids. Even though the elbow method I ran suggested a lower k, when I plotted a lower k, I found the data hard to interpret. Almost all of the points fell into one bin, making the graph much more homogeneous than I expected. As I increased k, categories became more clear, and groups started to break into regions that make sense given the context. After deciding on eight as my k, I plotted the algorithm on a map of the United States, which can be seen below:



The above graph identifies a few clusters that show distinct dialects. The clearest is the divide between East and West. As one crosses into more culturally "Western" states like Kansas, Nebraska, Oklahoma, and Texas, we see a new dialect cluster emerge that dominates the entire Western United States. Additionally, we see another divide between the South and the Rust Belt (which consists of some states in the Midwest like Illinois, Indiana, Ohio, and Western Pennsylvania.) This does conform with the conventional wisdom of both the South and the Midwest having unique dialects, but an unexpected result is many parts in the Mid-Atlantic and New England sharing the same cluster as the majority of the South. This is surprising, especially in New England, which is often considered a cultural enclave. Though I did not test stability, my inference given this result is that the projection is not very stable, and that I would ultimately need to adjust my projection to obtain a more accurate picture of American dialects.

4 Dimension reduction methods

I ran out of time and did not do this section.

5 Stability of findings to perturbation

I ran out of time and did not do this section.

6 Conclusion

Above is an analysis of Bert Vaux's monumental American dialect survey. While the analysis is unfinished, we still have arrived at some interesting insights. First, we saw that my nicknames for my maternal grandparents were unique for my place of birth - I used "Grandmother" and "Granddaddy" in an area dominated by

"Grandma" and an unspecified "other" which is likely a combination of a handful of different nicknames. Next, we saw clusters of dialects in America. Some of these formed in predictable ways – for example, the divide between the Eastern and Western half of the country. At the same time, we also witnessed some surprises, most notably the similarities between the Northeastern and Southern parts of the country.

While this paper provides a good start to the analysis of the dialect survey, it is only just a start. It would benefit from a deeper analysis, including dimension reduction methods and perturbation tests to better understand its stability.