Declan Shener
Joel Graycar
Summer 2014

Statistics 133 Final Project Analysis

We began our analysis by first writing some functions to map out every single question, color-coded by response, with the hopes that some obvious trends between geographic location and response might jump out at us. For this we used the original response data, rather than the binary data, since it allowed us to easily map each response to a different color, with palettes being generated by the RColorBrewer package. The results of that process can be found in the file all_maps.pdf, and though it worked as we had wanted, we discovered that the huge number of responses made it nearly undecipherable. What was immediately clear, however, was that there were many questions, such as Q054 and Q081, where one answer was by far the most common across the entire country.

With this in mind, we decided to proceed by using the binary data to find which response was most common for each question, and to then plot it and see what trends emerged. The resulting file (most_frequent.pdf) revealed that for many of the questions, the most frequent response would arise in the same approximate geographic location. For example, Q054.2, Q055.2, Q056.2, and Q057.2 are each the most popular responses to the questions, and each has a geographic distribution extremely similar to the others.

In order to get more rigorous results, we decided to apply hierarchical cluster analysis to these most frequent results, and see how stable the resulting clusters are. While running our code we discovered, unfortunately, that performing Euclidean distance and clustering calculations on all responses took an infeasible amount of time, and so decided to run it on a random sample of 3000 responses. We then ran the code with three levels of clustering: 8, 6, and 4 clusters. The resulting maps, most_frequent_k8.pdf, most_frequent_k6.pdf, and most_frequent_k4.pdf reveal a number of stable clusters. In particular, the regions of the West Coast, Midwest, South, and Northeast recur in almost every single question, most obviously visible in the 6 cluster maps. This is because many of the questions had responders in Alaska and Hawaii, who, in nearly all cases, formed their own two clusters. Thus, for the 4 cluster maps the result in nearly all cases was Alaska, Hawaii, the Western half of the country and the Eastern half.

Another common trend was for the clusters to appear based primarily on longitude, with the vertical dimension of the cluster spanning the entire length of the country North to South (seeQ096.4 and Q102.1 of the 6 cluster maps). This seems to imply that linguistic trends often occur based on longitudinal location.

After completing our analysis of the most frequent responses, we decided to analyze the least frequent responses in the same way, since these were more likely to be unique to certain regions. After a very simple change to our code, we ran the analysis method, but quickly realized our options for amounts of clusters was much more limited. This was due to a few of the least common responses having only received a few votes, thus limiting our maximum number of clusters. We found we could still run the analysis method with 4 clusters, with produced very helpful

results (least_frequent_k4.pdf). For instance, questions Q070 and Q71 both have clusters right on the Pennsylvania-Maryland region, implying the region there has a few unique regional dialects.

In order to help with easily examining the data for certain questions, we came up with a couple methods, plotBinaryQuestion and plotOriginalQuestion, that take in a question name, such as "Q070.2" for Binary or "Q091" for original, and map out all the responses to that input. plotBinaryQuestion simply displays all the locations of those who picked it, while plotOriginalQuestion color codes the points by response, and provides a corresponding legend. This allowed us to easily examine a response more closely when other analysis indicated it as a potential point of interest; for instance, when Q075.4 was picked up as one of the least frequent responses that seemed heavily clustered in a single area, running plotBinaryQuestion("Q075.4") let us immediately see that yes, response 4 of question 75 does seem to be concentrate almost entirely in the Northeast, more specifically Massachusetts, Connecticut, and Rhode Island.

By taking the absolute value of the loadings of each value in a specific principal component, we were able to gain insight as to which questions played a significant role in determining the variability of our data. We decided to take limit our exploration to the the first 10 principal components, as well the 10 questions that produced the most absolute variation within each. In order to do this, we ran prcomp on the binary answers to the questions and the sort the rotation of the principal components. We then subsetted the data to the first ten for reasons stated above. We tried to look for two answers to the same question that produced a relatively significant amount of variation to the same question. This allowed us to skip over comparing each answer to every question, and instead pushed us towards looking at solid examples of regional dialects. Sometimes the results were not necessarily clear, but usually we were able to do some analysis on the two maps we were given.

An example of this was question 56: "Pantyhose are so expensive anymore that I just try to get a good suntan and forget about it." The answers "acceptable" and "not acceptable" were important in determining PC1 in our analysis, so we pulled up maps of each. The map of people who answered "unacceptable" was pretty evenly distributed around the United States in addition to being the most popular answer. This did not give us much insight into whether this was acceptable regionally or not. However, by pulling up a map of the people who answered "unacceptable", we saw that the change people who had answered "acceptable" vs. "unacceptable" for each region, did not change much except for the south. What this could possibly show is that this statement is more likely to be viewed as "unacceptable" in the south.

Another interesting analysis this led us to was the use of the term "tennis shoe" vs. "sneaker". These two answers both produced significance in our analysis, and when we looked at a map there was a definite regional difference in how people answered this question. The term "sneaker" is more common in the northeast, while "tennis shoes" were more common in the rest of the country.