Lab 1 - Redwood Data, Stat 215A, Fall 2018

Kevin Benac

October 5, 2018

1 Kernel Density and Loess Smoother Plots

1.1 Kernel Density Plots of Temperature

In this section, we look at the impact for the choice of different tuning parameters on kernel density plots and LOESS smoothing using the redwood trees dataset we used in Lab 1. We start by giving the the density plots of the temperature of all the redwood trees using a Gaussian kernel for different bandwidths. We tried 0.5,1, 2 and 5.

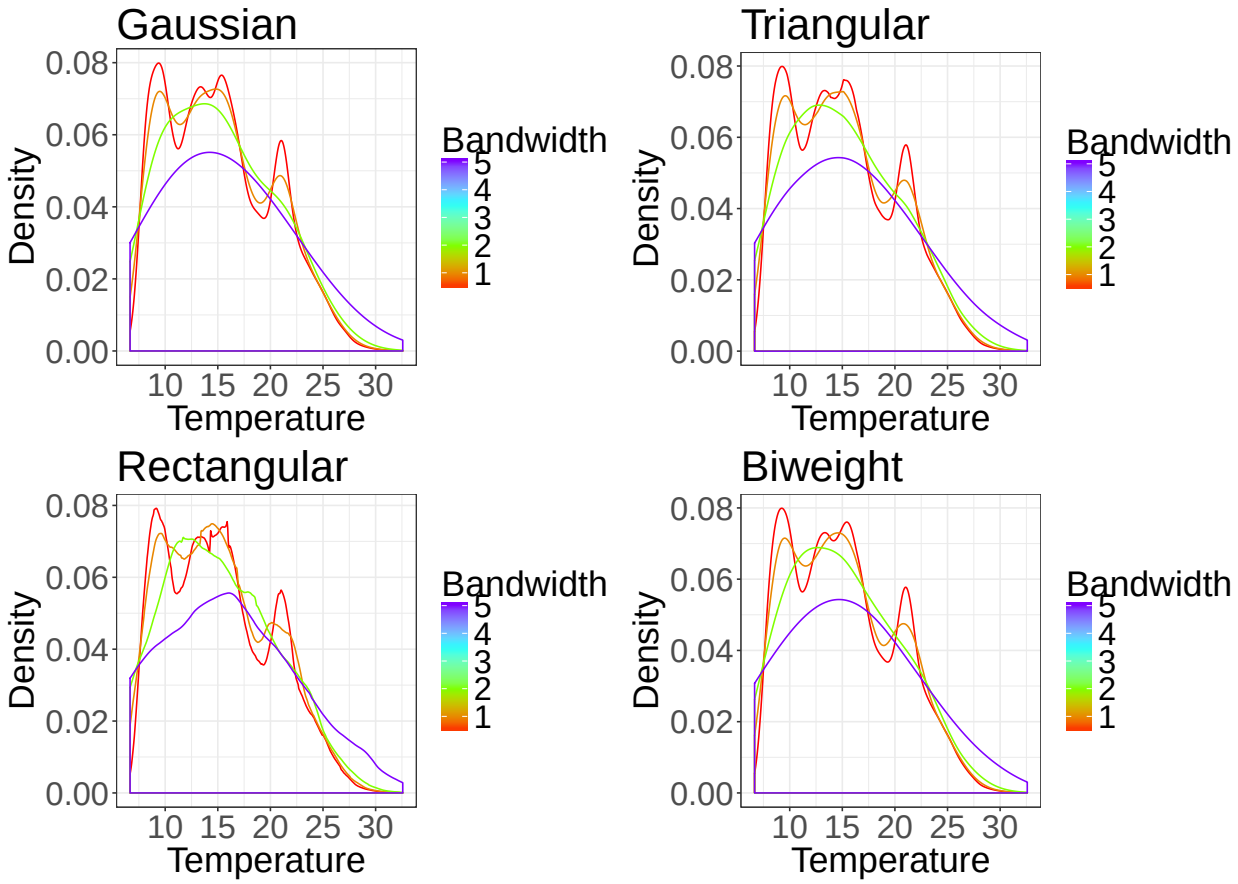

Figure 1: Kernel density estimates for the distribution of the temperature over the whole Redwood trees dataset. The four plots correspond to four different kernels. For each of them, we plot the density estimates corresponding to four different bandwidths that are 0.5, 1, 2 and 5. The curves are colored using a rainbow color gradient relative to the value of the bandwidth h.

We find that the bigger h, the less wiggly the density estimates. In particular, for $h \geq 2$ we see that the curve converges to that of an unimodal distribution. We also look at the effect of the choice of the kernel on the resulting density estimate. We try four of them: gaussian, triangular, rectangular and biweight. We also note that the choice of the kernel does not really matter in that case since the four plots on Figure 1 present very similar results. We only notice that the rectangular kernel yields to less smooth (sharper) estimates than the others. Let us now compare the four curves. We know that the choice of the bandwidth h impacts the bias and the variance of the resulting estimates. For a low value of h, we have low bias but high variance. In fact the extreme case would be when $h \to 0$. We would end up with spikes at every observation which is not very informative when we want to estimate the density. Instead we would like to identify some mode or maybe several if the distribution is multimodal. For small values of h (0.5 or 1) we see that we end up with very wiggly estimated of the density. The bias is smaller but the variance is big: a slightly different dataset would result in a very different density estimate. The number and the place of the modes may vary a lot.

At the other extreme, when h is too big, we end up with a very flat density estimate. Of course the variance then becomes smaller but again, it is not very informative. A completely different dataset would not change much the density estimate. The variance is low but the bias becomes high. A value of $h = 5$ seems to be a little bit too big here. Indeed, the data seems to be skewed and it does not seem to be the case anymore on the purple curve. A good trade-off between the two (bias and variance) could be $h = 2$. We still observe some skewness in the data. We have one single mode and the density is rather smooth.

1.2 LOESS Fits

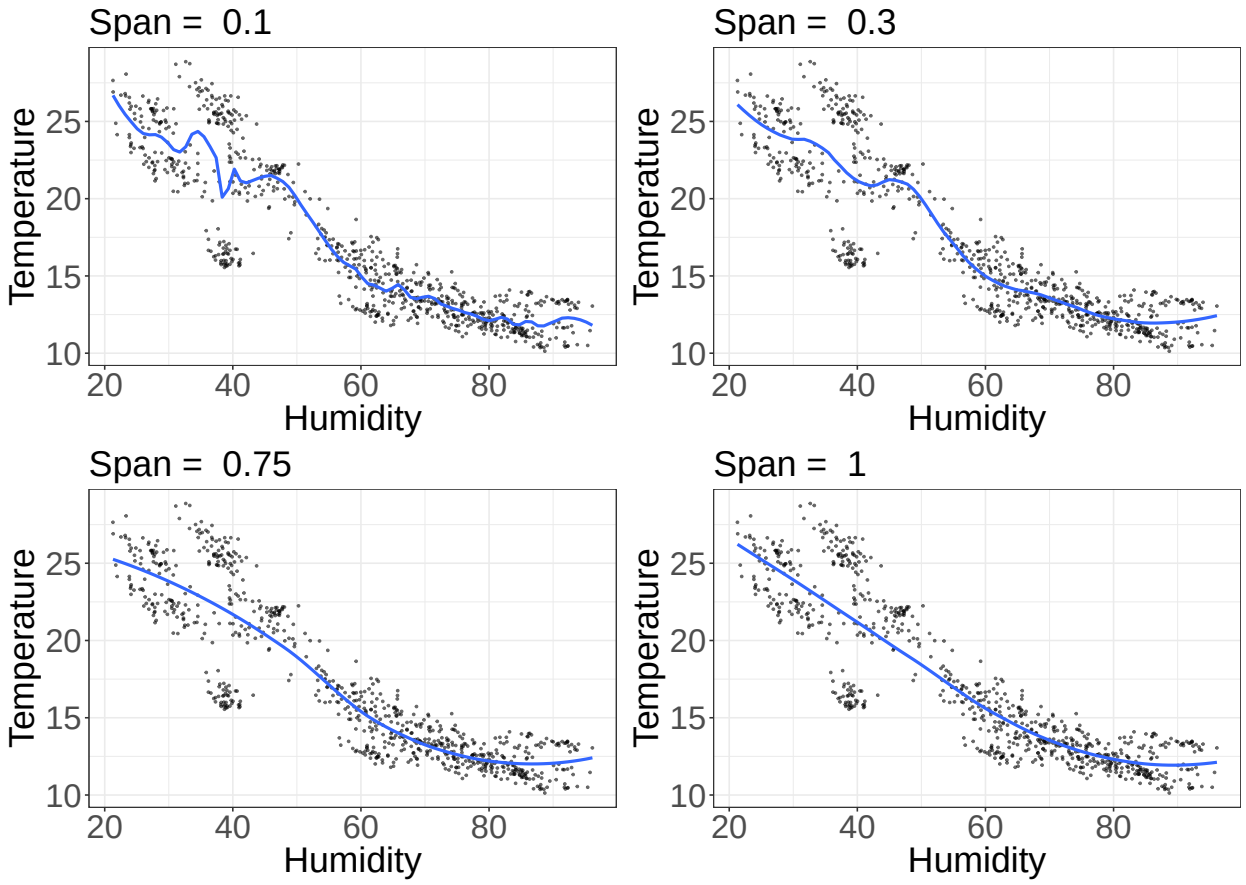

Figure 2: Scatterplots of the temperature against humidity every day at 6:00 pm. The four plots correspond to four different spans: 0.1, 0.3, 0.75 and 1. For each of them, the blue curve is the linear loess fit.

We then look at the impact that the span tuning parameter on a loess smoother has on the plot of temperature

against humidity for the same time every day. We arbitrarily chose 6:00 pm. The span parameter corresponds to the proportion α of the points in the local neighbourhood for each point (if $\alpha < 1$). The span parameter for the loess can be interpreted like the bandwidth in kernel density estimation. For a linear loess smoother with spans 0.1, 0.3, 0.75 and 1, as in Figure **??**, as the bandwidth gets larger, we see that we get something approaching linear. We observe that choosing a span 0.1 or 0.3 gives a fit that is much more local and then the resulting curve is much more wiggly than when choosing a span 0.75 or 1.

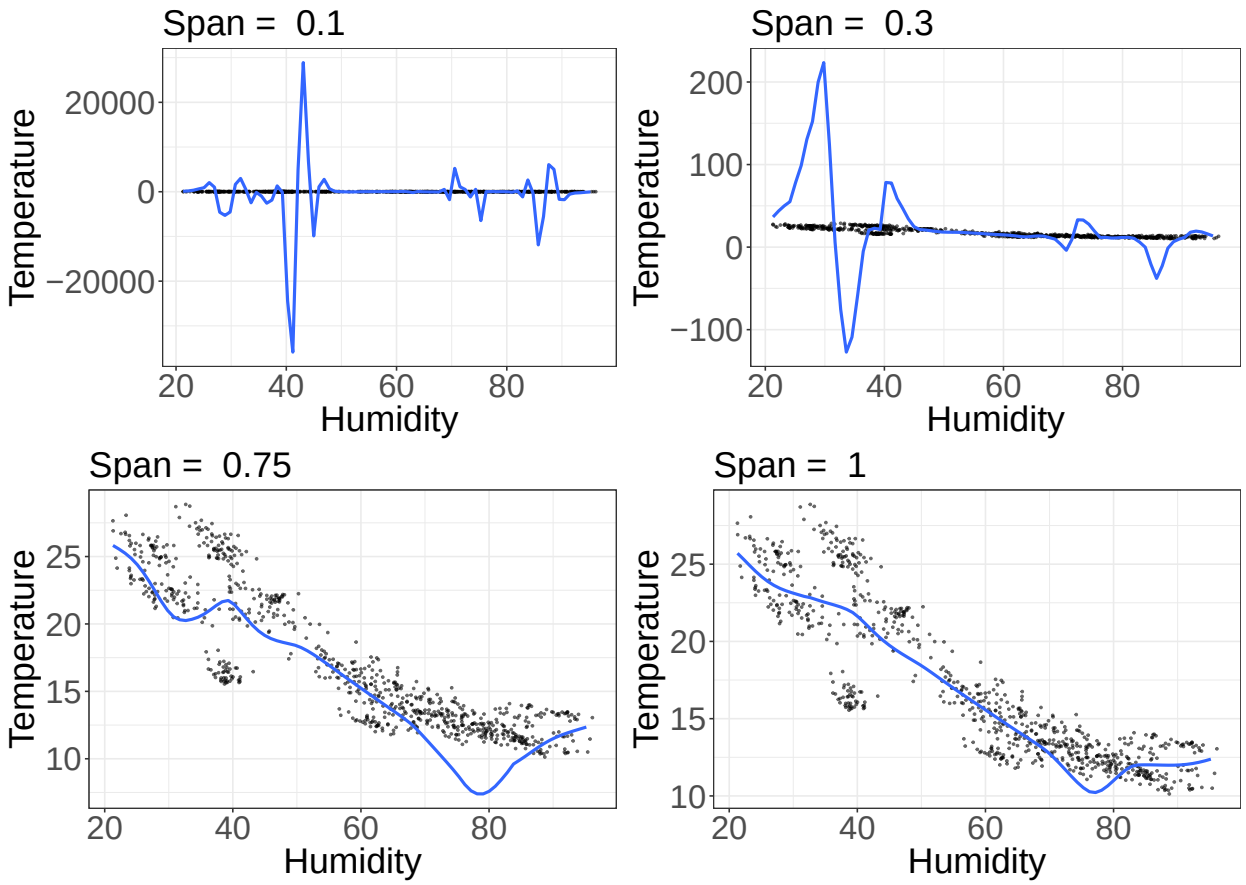

Figure 3: Scatterplots of the temperature against humidity every day at 6:00 pm. The four plots correspond to four different spans: 0.1, 0.3, 0.75 and 1. For each of them, the blue curve is the linear loess fit.

For a quadratic loess smoother, we observe very weird behaviors when the span is too small (Figure 5). We see that the loess tries to locally fit something quadratic, but it doesn't really do well and it is getting worse and worse when we increase the degree of the polynomial. The problem likely is that it tends to locally overfit the neighbourhood but we end up with a global fit that does not closely follows the scatterplot. However, it improves when we increase the span. When we increase the degree of the polynomial we should as well increase the span. The global take-away message of all these LOESS plots is: the warmer, the less humid and conversely.

2 Introduction

In a country as huge as the United States, it seems natural that each region has their own dialect for some words. Some drink soda, others call it pop. Some wear sneakers, others wear tennis shoes. The answers may vary a lot depending on where we go in the US. As an international student, I might not be the best persone to debate about this but all of these differences have resulted in the study of dialects and hence, the study of computational dialectometry has emerged, which aims at measuring variations in dialect based on

geography. In this report, we analyze survey results from the Dialect Survey conducted by Bert Vaux and Scott Golder in 2003 using dimension reduction and clustering methods to present our own findings about different dialects in the US.

3 The Data

The data was gathered from a survey conducted by Bert Vaux and Scott Golder with 122 questions relating to both phonetic and lexical differences. In this report, we only focus on the questions related to lexical differences, which are questions 50-121. We have responses to the questions from $n = 47,471$ participants as well as their self-reported city, state and zipcode data. A previous GSI for STAT 215A also collected their latitude, and longitude. This data has also been binned into one degree latitude by one degree longitude sequences where in each bin, binary response vectors were summed over individuals, which gave a second dataset.

3.1 Data Quality and Cleaning

Globally, the data is not too dirty. It was submitted online and a program compiled the answers to the questions, which yielded to the `lingLocation` dataset. Nevertheless, ther are still some inconsistencies in the information on city, state and zipcode in `lingData`. Some people gave some fake information. For instance: one person put 'expatriatecommunityinHongK' as their city in Massachusetts.

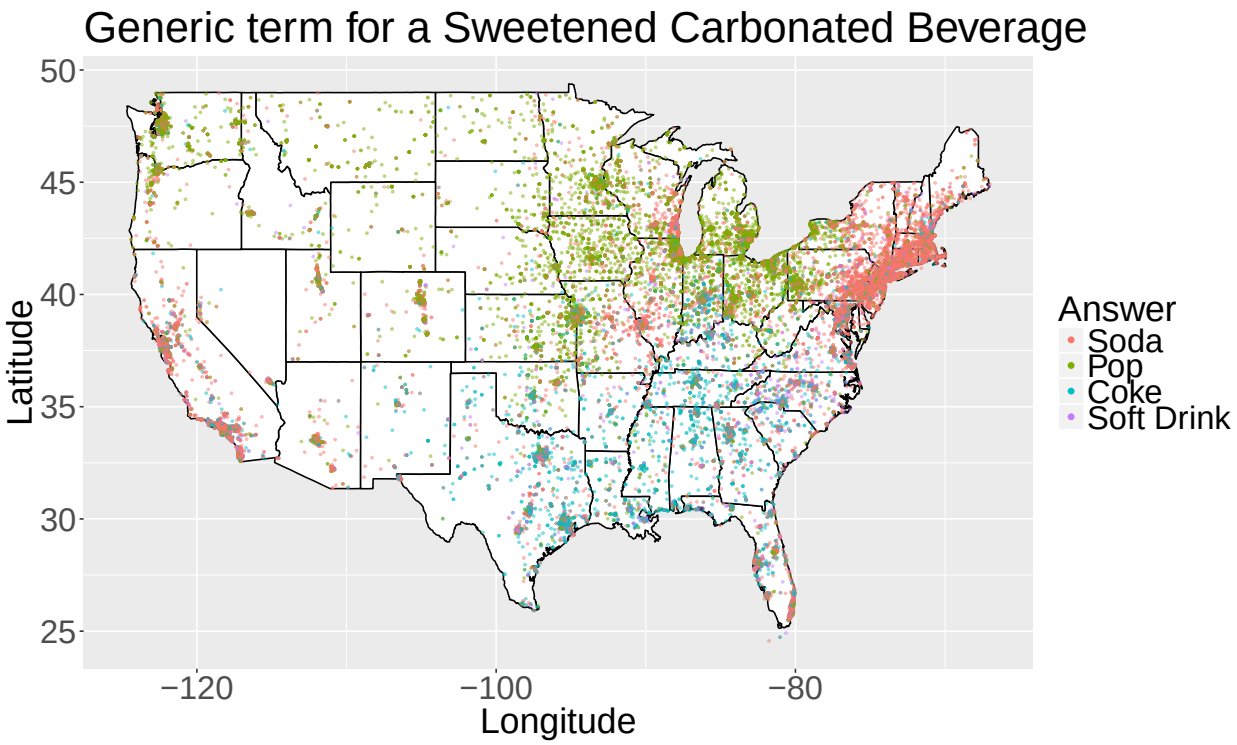

Figure 4: Question 105: What is your generic term for a sweetened carbonated beverage? – Map plot with the answers. Each dot corresponds to a different participant. Colors correspond to the different answers.

Fortunately, there is a `zipcode` package in `R` that lists all the zipcodes with their corresponding cities, states, latitude and longitude. We then decide to not take their information about city, state, latitude and longitude into account but to trust only their zipcode information which we match with the `zipcode` dataset. The only problem is that some people entered a zipcode that does not exist. However, it is only 641 observations out of the 47,471 so it does not have a big impact on the data.

Finally, we create a binary vector for each user submitted entry for the questions. In order to do so, we look at the number of answer choices each question has then take a cumulative sum. Then, adding the cumulative sum (except the last one which will be 468) to the answer choices (except the first one) tells us the positions of the vector where the one should be. We then end up with a big binary matrix with 468 column.

3.2 Exploratory Data Analysis

We look at two questions: Question 105 regarding the generic term of a sweetened carbonated drink and Question 78 regarding scratch paper vs scrap paper. In both cases, instead of plotting all the answers, we focus on the most popular ones. For Question 105 we consider answers Soda, Pop, Coke, and Soft Drink. From Figure 3, we can see that soda is present almost everywhere, which is not surprising as it is the most generic term. Pop is found primarily in the midwest, but there is a small cluster in the North-West. Coke is rather in the South though. Soft drink is less common, but it primarily shows up at the East.

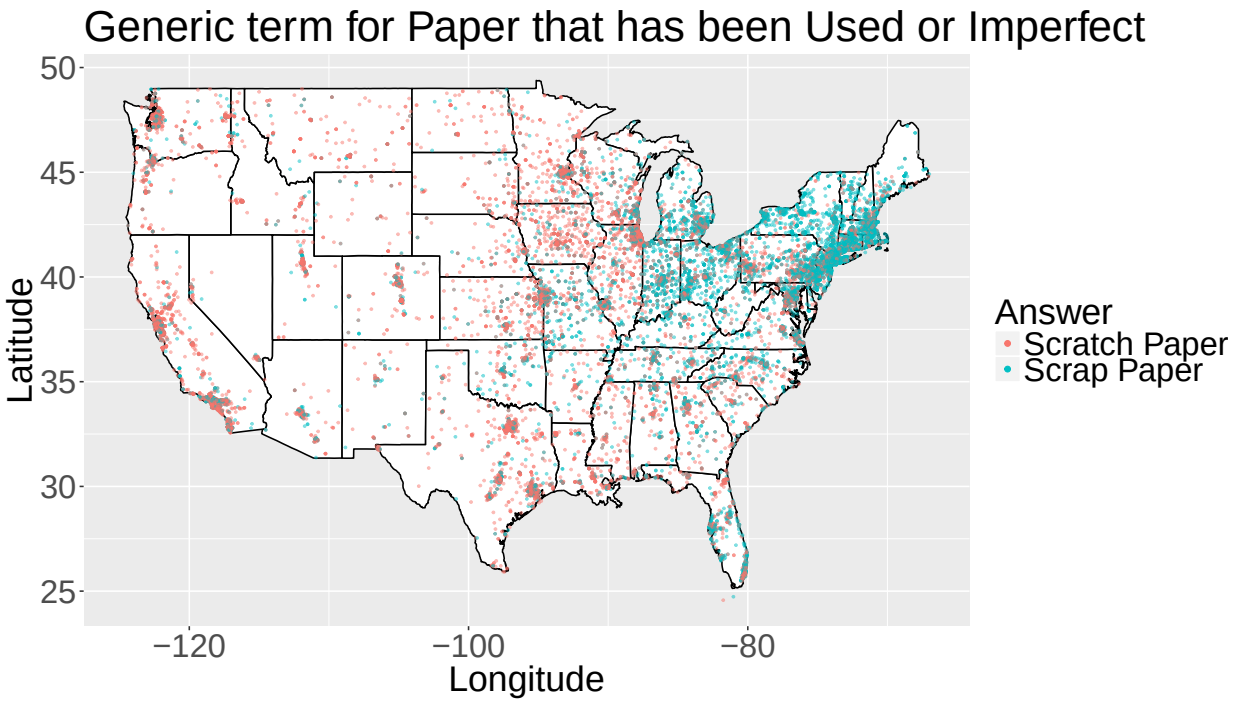

Figure 5: Question 78: What do you call paper that has already been used for something or is otherwise imperfect? – Map plot with the answers. Each dot corresponds to a different participant. Colors correspond to the different answers.

We also look at Question 78 regarding Scratch Paper vs Scrap Paper. Scratch Paper seems more or less universal, but Scrap Paper seems to be more popular in the North-Eastern states. In looking at this particular

question, we might be able to predict who is from the North-East.

Globally it is hard to find a good model to accurately predict answers to one question using another. However, we can see that the non-response rate can be predicted using the other questions since oftentimes, people who do not answer one question are more likely to not answer others. Finally, we look at a third question. This was a sentence that was shown to the participants and they were asked whether the use of the word "anymore" is acceptable or not (Figure 6). It is interesting to see that most people think it is not and they are mostly located along all the borders of the United States while people who think it is acceptable are mostly at the interior of the country.

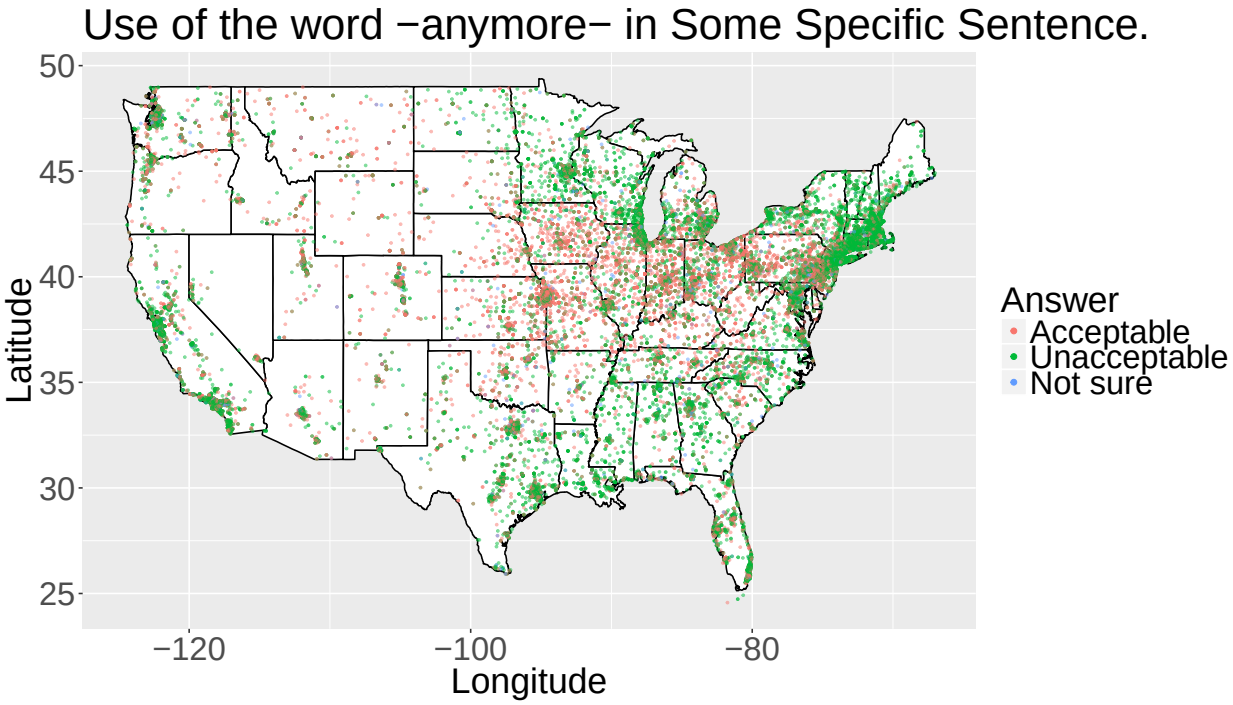

Figure 6: Question 56: Is the use of "anymore" acceptable in the following sentence: "He used to nap on the couch, but he sprawls out in that new lounge chair anymore" ? – Map plot with the answers. Each dot corresponds to a different participant. Colors correspond to the different answers.

4 Dimension Reduction Methods and Clustering

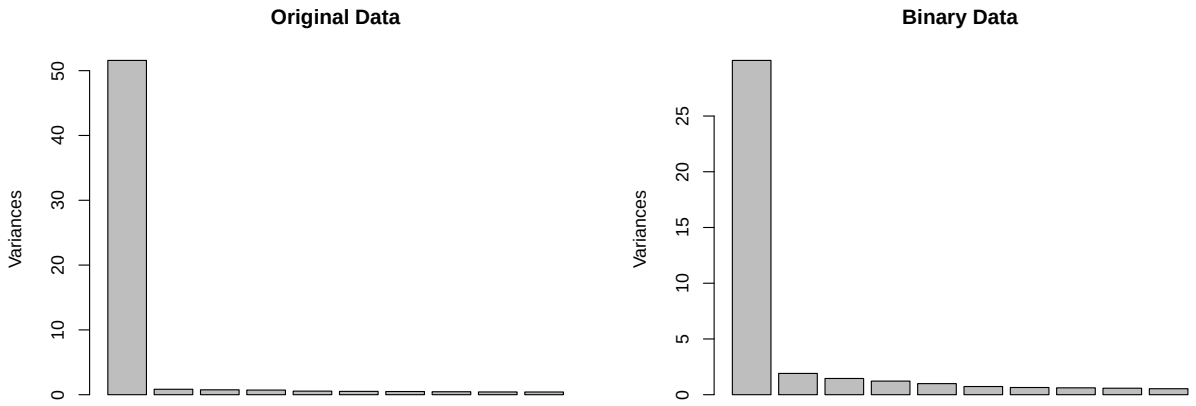

Figure 7: Screeplot for the PCA of the two datasets. On the left, the original categorical data for the answers and on the right, the binary data.

In this section, we would like to use dimension reduction methods and clustering methodology in order to see if we can find groups of people answering similarly. We are mainly interested in two datasets to conduct our analysis: first, the cleaned version of original `lingDataClean` categorical dataset and second, the transformed data that was turned into a 468-dimensional binary dataset.

4.1 PCA

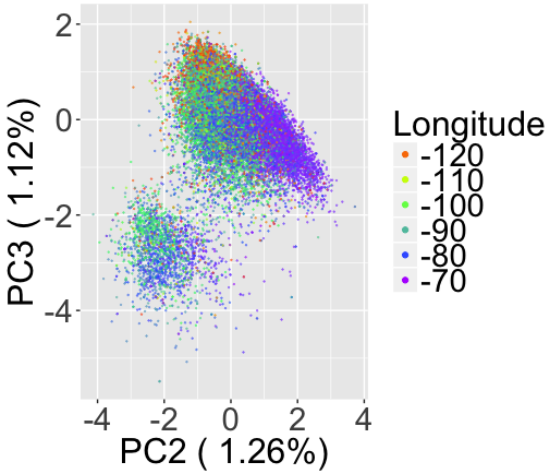

Figure 8: PCA scatterplot for the categorical data. The dots are colored by a rainbow gradient according to their longitude.

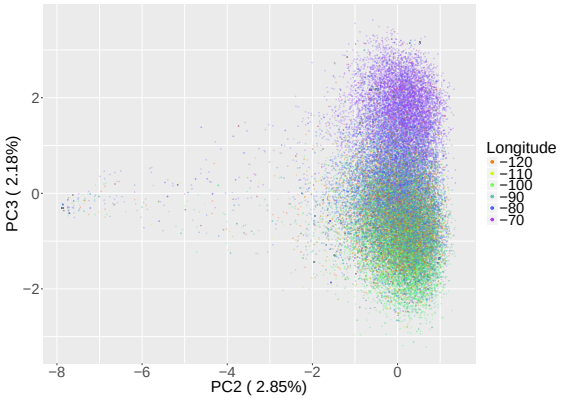

Figure 9: PCA scatterplot for the binary data. The dots are colored by a rainbow gradient according to their longitude.

When looking at the screeplots for both the original linguistics data and the one-hot data as in Figure 7, we find a nice elbow after the first two components, which means that our PCA has been successful in its goal

to reduce dimension. Indeed, we find that for the original data, the first 20 principal components account for 90% of the variance while for the one-hot data, the first 100 principal components account for 88% of the variance. At first sight, it seems like PCA has done a better job in reducing dimension on the original dataset since only around 15% of the original data captures more variance than 21% of the one-hot data. If we project the data on the first two components. In both cases, it is hard to notice a pattern and find clusters. We then try other projections such as PC2 against PC3. If we consider the original categorical, it does not really help (Figure 8). However, if we look at Figure 9, then we notice that PC3 seems to capture the longitude which makes sense geographically.

PCA for one hot data

4.2 k-Means

The next step is to perfom clustering in order to identify groups. Here we use k-means using 20 different starting points. We use the first 10 principal components in the two datasets and perform k-means with $k = 4$ since that was the value of k that gave us the most stable clusters (this will be discussed in the next part). Using these clusters, we identify groups. For the original dataset, we can visualize the clusters in the reduced dimension space (Figure 10) and then use these labels to color the dots on the map (Figure 12). We observe that if the separation is clear in the reduced-dimension space, it is less clear on the map. It is rather hard to identify groups from the categorical data. However, if we use the encoded binary data, we see on Figure 13 that there are three major groups: the North-East (red), a big part of the South (especially the South-West) but it also goes up to some part in the Mid-West (purple) and the rest of the map (green). However, the PCA plot is not that great for distinguishing the clusters even on the binary encoded data. It would probably work better on the aggregated data but we will unfortunately not pursue in this direction by lack of time.

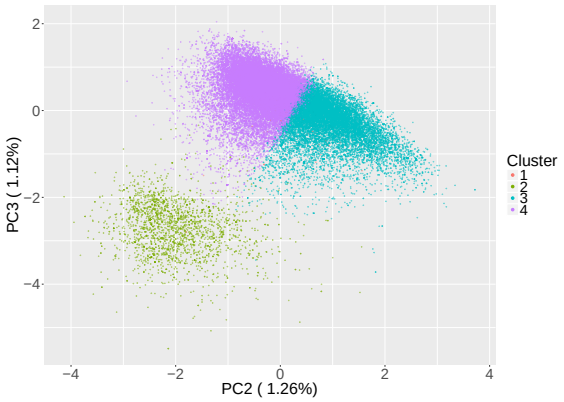

Figure 10: Projection on the second and third principal components of the categorical data. Dots are colored by cluster.

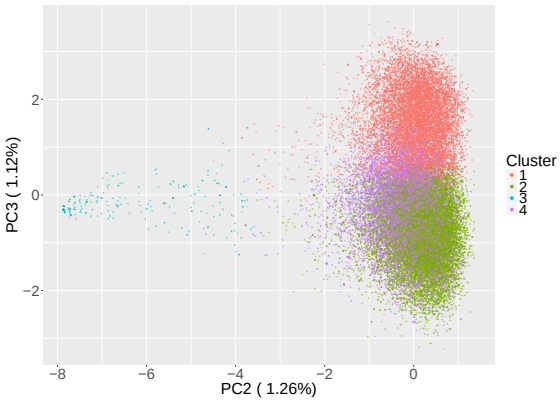

Figure 11: Projection on the second and third principal components of the binary data. Dots are colored by cluster.

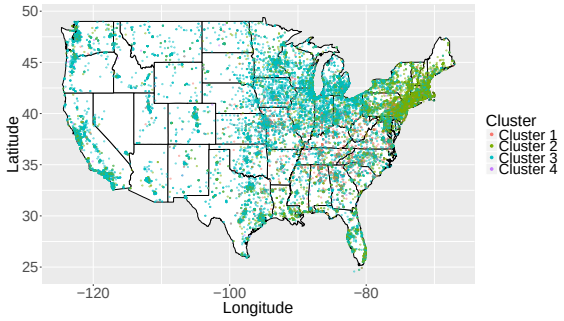

Figure 12: Map plot with dots colored by the clusters found after performing k-means on the first ten principal components of the categorical data.

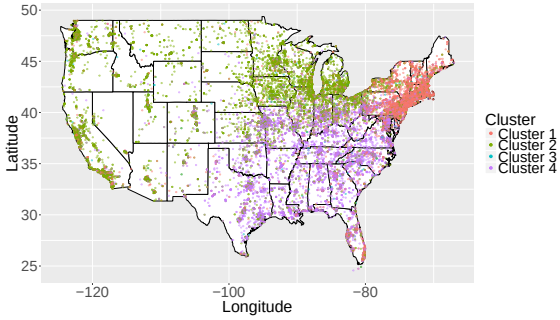

Figure 13: Map plot with dots colored by the clusters found after performing k-means on the first ten principal components of the binary data.

There is a fourth cluster that is not concentrated in a specific geographic area and might correspond to the outliers that do not belong in a specific region for which they speak the typical dialect. That might be international people or Americans that lived in many different areas of the United states and speak a mixture of different dialects and not just one in particular.

5 Stability of Findings to Perturbation

We now want to assess the robustness of the finding. We found 4 clusters and would like to know how stable they are to perturbation. A way to do this is to do it by subsampling: we split the data into 4 random independent subsets and fit the k-means algorithm on each of the fold separately for different values of k. If the clustering method is stable, we should expect to observe some consistency between the four folds. We then perform this for k between 2 and 6 and find that the one that gives the best stability is when $k = 4$ (Figure 14).

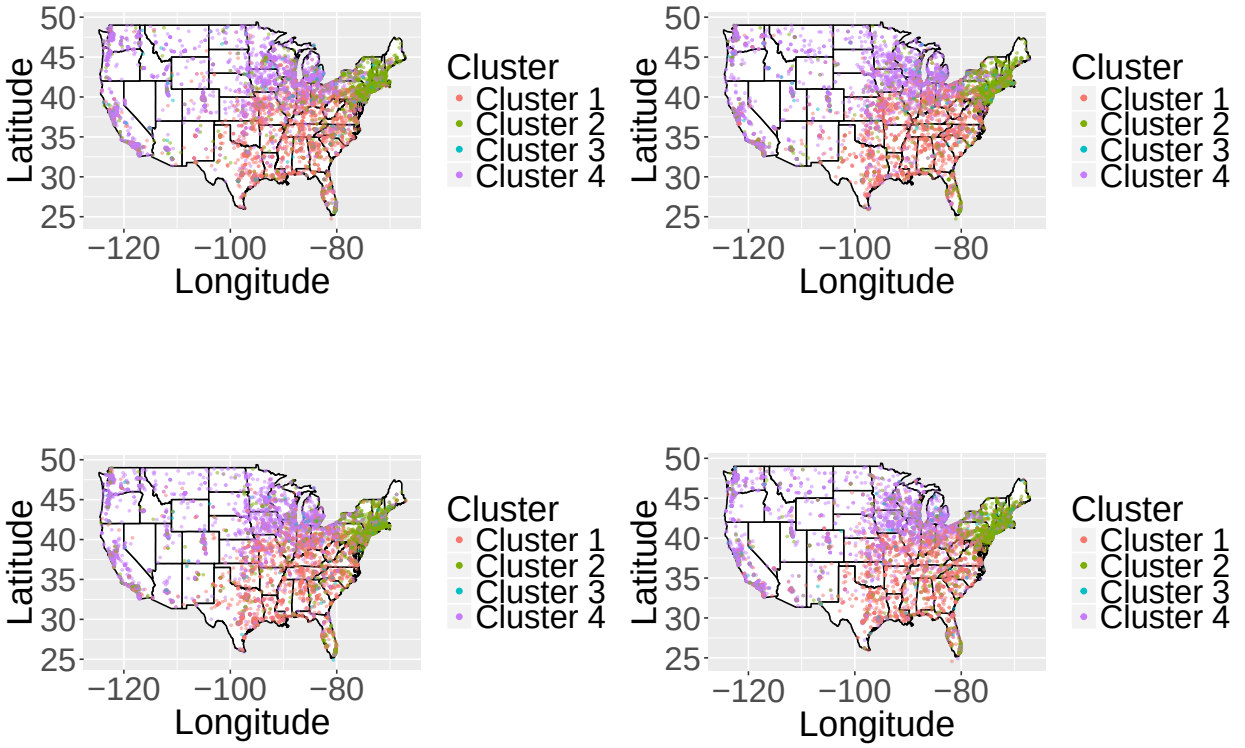

Figure 14: The k-means algorithm with $k = 4$ was applied to each of the 4 subsets separately using twenty different starting random subsets ($n_{start} = 20$). The labels of the four clusterings were renumbered in order to make the color of similar geographical regions match.

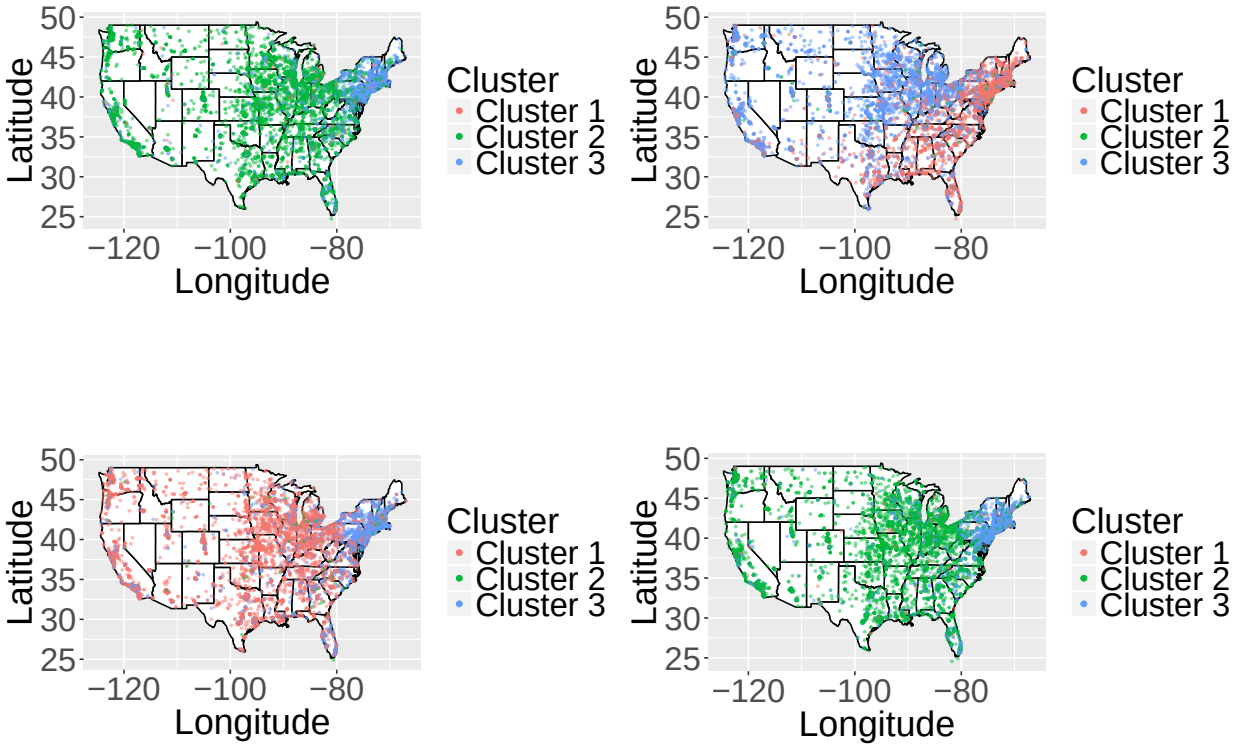

Figure 15: The k-means algorithm with $k = 3$ was applied to each of the 4 subsets separately using twenty different starting random subsets ($n_{start} = 20$). The labels of the four clusterings were renumbered in order to make the color of similar geographical regions match.

We also give in Figure 15 the same plot as the one in Figure 14 to show that k-means with $k = 3$ is less stable. The upper right plot shows that Florida is included in the same cluster that covers the North-East while this is not the case in the other three. Repeating the process for other values between 2 and 6 did not improve the stability and more than 6 clusters would become harder to interpret.

Finally, another interesting point is that the two states for which it is not really clear whether they belong to a cluster or not are California and Florida. This makes sense with intuition as they are probably the ones for which the proportion of native people is the lowest as people from everywhere in the United States (and even from everywhere in the world) move over there.

6 Conclusion and Discussion

Although we expected it, we found that dialects are related to geography. We were able to find meaningful clusers by first performing dimensionality reduction to our data then clustering them using k-means where we chose a number of clusters that guarantees stability. The clusters were rather stable as we have seen and did not contradict the exploratory data analysis we performed earlier. Except the geographical zones, we were able to detect another cluster which intuitively makes sense as not all the participants were born, raised and spent their entire lifetime in the same area.

In further investigations, it would be interesting to explore multidimensional scaling on the binary encoded data.

Informal Collaborators

Some discussions with Yutong Wang, David Chen and Philippe Boileau.

References

Nerbonne, John. “Introducing Computational Techniques in Dialectometry.” Computers and the Humanities, vol. 37, no. 3, Computational Methods in Dialectometry, 1 Aug. 2003, pp. 245âĂŞ255.

Nerbonne, John. “Progress in Dialectometry: Toward Explanation.” Literary and Linguistic Computing, vol. 21, no. 4, 8 Sep. 2006, pp.387-397