Lab 2 - Linguistic Survey, Stat 215A, Fall 2018

Aummul Baneen Manasawala

October 25, 2018

1 Introduction

This report is divided into two parts. The first part deals with the redwood dataset that was collected by the macroscope experiment by Tolle et al. in the "A Macroscope in the Redwoods" paper [1]. We would experiment with the parameters in kernel smoothers therein along with analysing the interplay between temperature and humidity at a given point of time of the day. In the second part we would do Linguistic survey.

2 Macroscope in the Redwood Data Analysis

We would do two types of analysis in this section. First, we would analyze the temperature distribution over the whole dataset and then we would follow it by trying to get insights of the changes in the temperature with respect to humidity at a given time of the day for all the nodes.

2.1 Temperature Distribution

We experiment with different kernels and bandwidth to estimate the true function of the temperature variation in the redwood at all days, nodes and time. The bandwidth of the kernel is a free parameter which exhibits a strong influence on the resulting estimate. To illustrate its effect, we plot and compare different bandwidths as well as kernel shapes. We find that the red curve is undersmoothed since it contains too many spurious data artifacts arising from using a bandwidth of 0.02 which is too small. The green curve is oversmoothed since using the bandwidth h = 5 obscures much of the underlying structure. The blue curve with a bandwidth of h = 1 is considered to be optimally smoothed since its density estimate is close to the true density.

The kernel is the shape of the window function. Several types of kernel functions are commonly used: uniform, triangle, Epanechnikov, quartic (biweight), tricube, triweight, Gaussian, quadratic and cosine. We tried to compare the Gaussian, Rectangular, Epanechniko and Triangular. Since the Gaussian function has infinite support and is smooth, it results in the most smooth plots. However, the rectangular kernels have a very abrupt window. It either gives a uniform weight to all the points that are in the window or doesnot give any weight to points just outside the window. Thus, it results in a very horny density estimates as could be seen in the figure 1.

2.2 Relationship between Humidity and Temperature

We begin our analysis by first subsampling for the data which corresponds to a given point of time in the day. We study the relationship between temperature and humidity using this sample. In order to find out the function that defines their relationship, we choose to fit a loess smoother. The loess smoother function a span argument that tells us what percentage of points are used in predicting x (like bandwidth in density estimation in the previous section of this report). So there's an idea of a window size; it's just that within the window, we give more emphasis to points near the x value. We find the variations in the loess smoother line

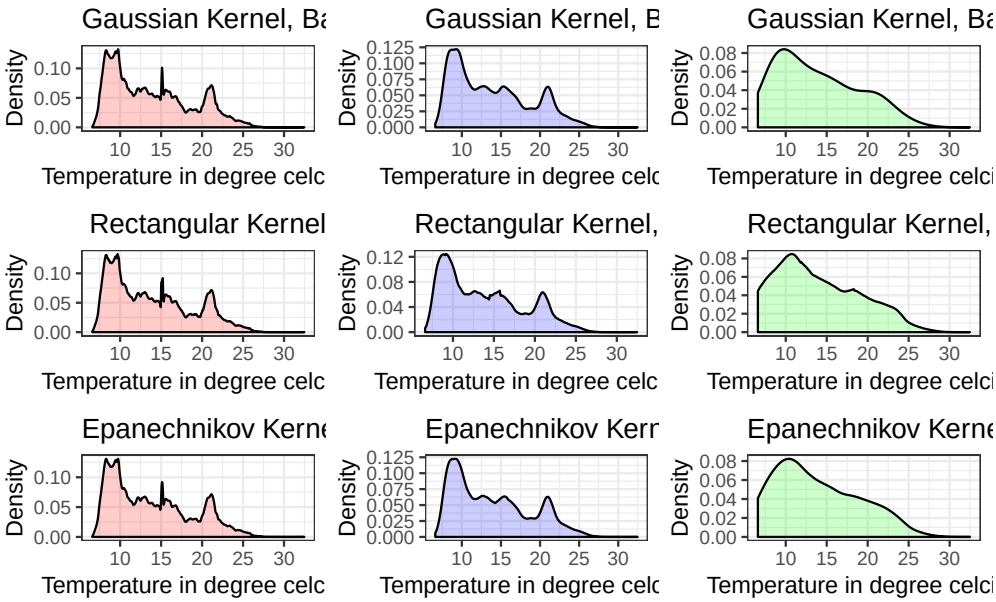


Figure 1: Density Estimate of the temperature in the Redwood with various kernel types and bandwidth

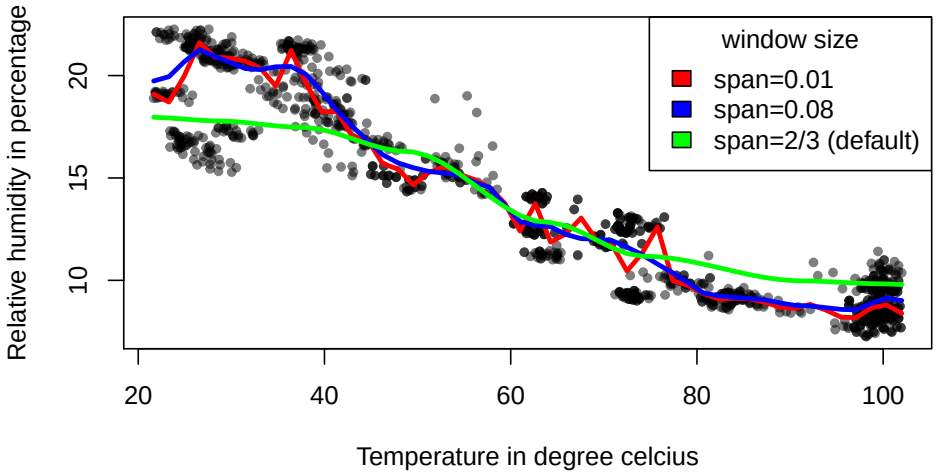


Figure 2: Variation in the loess smoother with changing window span

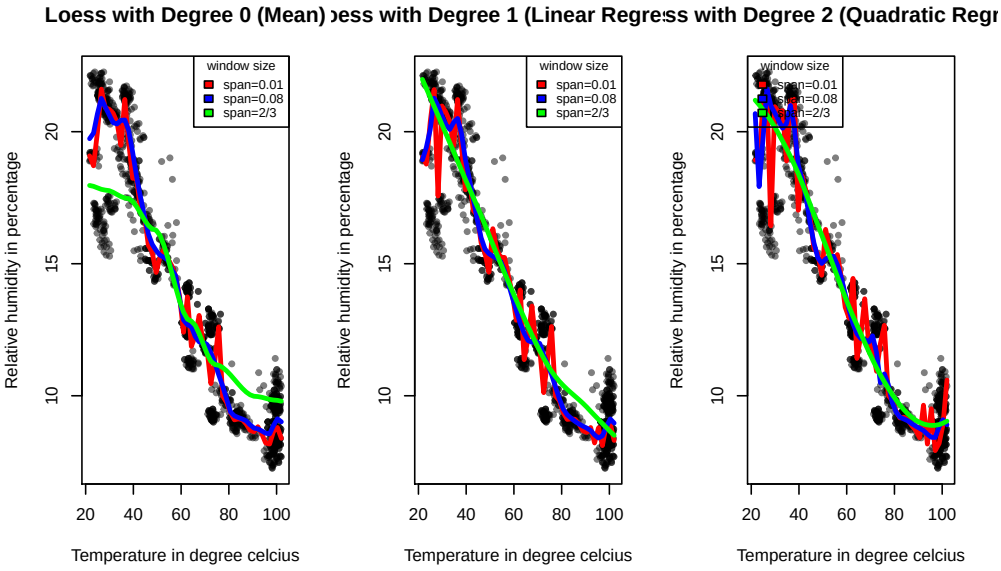


Figure 3: Variation in the loess smoother with different degrees

by variation in the span similar to the variation in the kernel density estimate by variation in the bandwidth. A larger window like the green loess smoother in fig 2 would be biased and would assume that all the points within the window follow the same function. We loose a lot of information about the variability in this case. On the other extreme, if we choose a window that is very narrow like the red loess smoother line in figure 2 , out loess smoother line would cater to the temporary outliers as well and would be very jerky. The blue line would be the optimum in our case.

We could also choose the function that we want to fit for the selected window. The function could be a polynomial of any degree. For the purpose of illustration, I chose to compare three functions with degree zero, one and two respectively. The first function draws a line which is the mean of all the values. The function with degree one would fit a linear regression line with all the points in the given window. Likewise, the function with degree of two would fit a quadratic regression curve for the points in the span and so on. The three case for our temperature and humidity graph is with loess smoothening function of degree zero, one and two could be compared in the figure 3. The graph with fits the mean assumes a constant relationship of the points in the window and if we extend the resultant loess curve, we find that it would not predict the humidity with increasing temperature very well. The rightmost curve in the figure 3 with degree two fits a quadratic regression line for the points in the window and is overly sensitive to the variability. Thus, we see that there is a prominent bias-variance trade off in selecting the degree of the curve to be fitted in the loess function. As we increase the degree, we reduce the biased assumption of the data in the window following a low order relationship but we risk heavily increasing the variace that could negatively affect our predictions as it would cater to noise a lot.

3 The Data

According to Nerbonne et al. in the paper "Introducing Computational Techniques in Dialectometry" individual linguistic features — words, constructions, and pronunciation variants are associated only weakly with geography. Nerbonne and Kretzschmar (2003) focused on the role of computers in dialectometry and used the data from the linguistic survey to explain the language variation. Our data thus consist of questions and answers of people with respect to their location in the US.

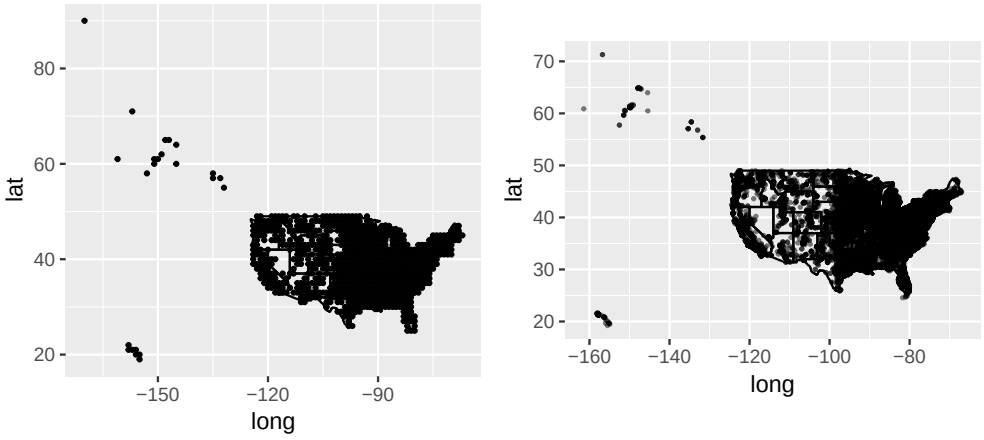


Figure 4: Outliers in the linguistic location data(left) and individual linguistic data(right)

3.1 Data quality and cleaning

We find 4 percent of the data enteries in the individual linguistics data were NAs. We initially hoped to keep the rows and condition them to become functional. Since, if the entries were not entirely NAs, we could gather some information out of it. But we found that all those rows had the same profile of all the enteries as NAs and thus they were completely useless. If we would atleast have the latitude and longitude nformation of those rows, we would have substituted the NAs with the mean of the values of the nearby latitude and longitude regions. However, in the absence of any such data, we found it suitable to discard these rows from the analysis.

We find that the data for some questions from question number 50 to 121 is missing. We acknowledge that we only have data for 67 questions as opposed to 72 that we expected. For the next step of data cleaning, we convert the data types of the option number selected for each question as factors to ease our analysis. Since this survey is for the United States of America, we also removed all the values that are outside the US. We can use this figure to see that some of the location latitude and longitude are wrongly mentioned which makes the data points fall outside the US in both the dataset the aggregared and individual as can be seen from figure 4.

3.2 Exploratory Data Analysis

To get a better picture of the variations in the data, we start with narrowing down to two questions and delving deeper into those. We would investigate their relationship to each other and geography. The question that is personally very interesting to me is how people call their maternal and paternal grandfather. I don't have either of them now but I have good memories of them. In their honor and respect, I would like to explore the variations in the way these two prominent person in every one's life are being called.

The question number 71 is about what do you call your paternal grandfather. The 5 options were gramps, grandpa, grampa, pap and other. Sadly we would not be able to know how most people address their paternal grandpa because about 37.3% people have voted for other. This mysterious and obscure names are generally used by the people in the southeast and the northeast. Most people in the midwest and west address use standard conventional "grandpa" and "grampa" and very few people less than 2% use pap and gramps to call their paternal grandfathers.

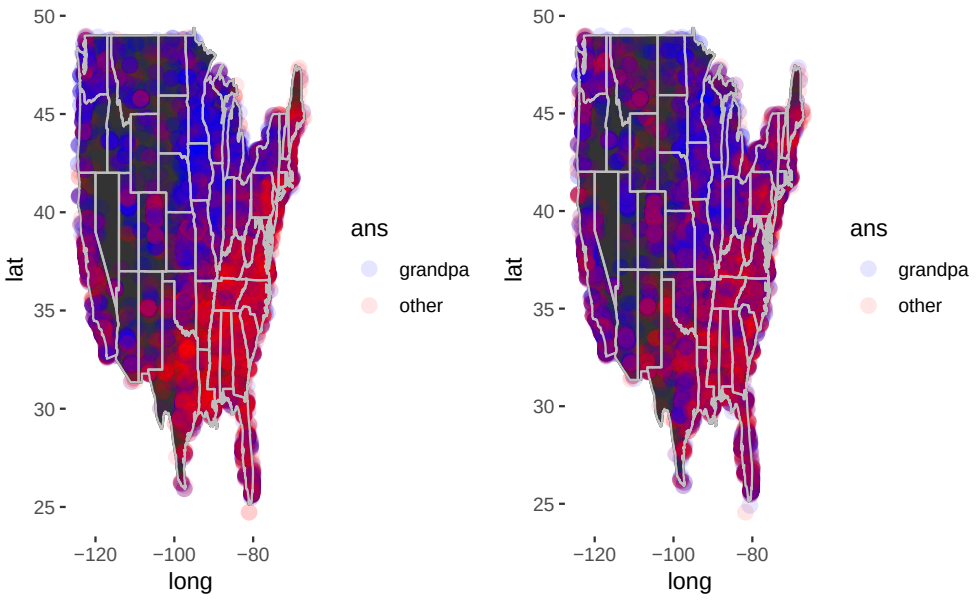


Figure 5: Geographical responses to call maternal(left) and paternal(right) grandfathers respectively

Comparatively, when we see the distribution of the names used for maternal grandfather, the seven possibilities are gramps, grandpa, grampa, grandad, pap, spell it as grandpa but pronounce it as grampa and other. Very similar to the case of the paternal grandfather, the maternal grandfathers are also mostly called by names that we could not gather by the survey as people chose the 'other' option. I wish our survey could account for those names, it would be interesting to find out the eccentric names people use to address their maternal grandfathers. These are the same maverick people of southeast and northeast USA that constitute 32% of the total. Rest people spell it as grandpa but pronounce it as grampa or grandpa. They are also the ones mainly from midwest and western region.

In all, we can visually correlate from figure 5 that the two questions are highly correlated. Answer from someone for one question very strongly predicts the anwer for the second. For example, if a person who address his paternal grandfather as grandpa would most likely also call his maternal grandfather as grandpa. In our case, the answer has also high geographical correlation. For example, if the person is from southeast it would be a fair prediction that he would not call his/her maternal as well as paternal grandfather as grandpa but with some other customized name. Likewise, a person from midwest is highly likely to address his grand father as grandpa. So, they are geographically related.

The correlation between the answers to the above questions of calling maternal and paternal grandfather as the same was a bit obvious. To further bolster our findings and in order to explore more such correlations in the geographies as well as interdependencies in the questions, we considered two other interesting questions. The first one is what term people use for carbonated drinks. By plotting out the people on the graph of US, we found that north-eastern people generally call the carbonated drinks as "soda". The midwestern and western people use the term "pop" while the south eastern people address by the term "coke". Interestingly, when we changed the question to what people of US call a miniature lobster that is found in stream and lake, we found the people to be categorized with respect to same geographical buckets as in the question of carbonated drinks. The people in north east become a group that calls it "crayfish". Likewise people in midwest generally call it "crawdad" and the people of southeast call it "crawfish". We can visualize the strong geographical dependence as well as correlation between these two question which can help us predict the answer of one from the other in figure 6.

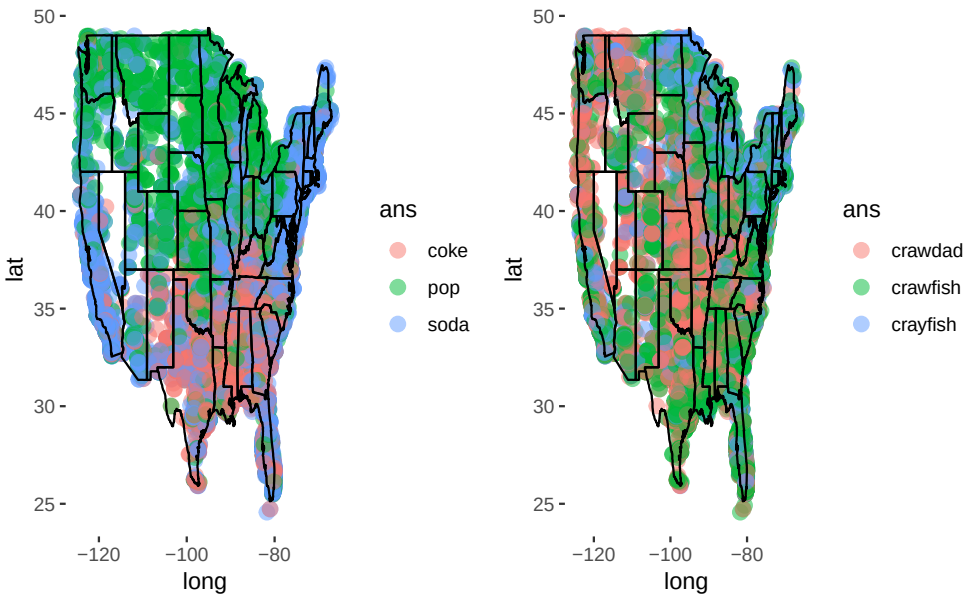


Figure 6: Geographical responses to address carbonated drinks(left) and miniature lobsters(right)

4 Dimension reduction methods

Because of the curse of dimentionality, we chose to project our data in lower dimensions before clustering for more efficient computation as well as better results by reorienting to the dimensions that explain the most variability in the data. We started with conditioning our data. Since each data point represents the number of people in the given block that answered a particular option of the question, we would like to replace the number by the percentage people that chose a particular answer. We do that in order to avoid catering to the answers of places where the number of people who responded to the survey was high. We don't need to normalize the columns because we are comparing similar scaled proportion of people who opted for an option in the survey.

Moving forward, in order to choose the number of dimensions to reduce, we analysed the scree plot and identified an elbow at dimension number 63 as shown in the figure 7. The amount of variability explained by the dimensions higher than where the elbow is contributes very less to the overall variability. Therefore, we go with reducing our data in 63 dimensions which covers about 80% of the total variability.

After getting the data in a reduced dimensions, we try to find some structure/categorization in the data. In order to do the division, we use the clustering technique of k means. In order to figure out the number of clusters in the data, we use a elbow curve. For our data, the elbow curve shown in the figure 9 gives a elbow at the number of clusters as 3. Therefore, we select 3 clusters to group our data.

After obtaining the 3 clusters in the data, we try to visualize the data in 2 dimensions of some of the prominent principal components. From figure 10, we find that the clustering that we obtained through k means divide the data in a way that makes sense visually. In order to find out whether, this clustering and division is related to the geography, we plot the points on the US graph and colour them by the color in figure 11. Viola! the k means clustering divison is in complete consistency with the geographical divisions on which the answers of the previous questions divided the nation. The three k means group can be very cleanly used to divide the three main geographical regions of the US, namely : north easterm, south eastern and midwest including some areas of the west. There is a complete continuum except for some regions of the west which are not very populated and are a bit remote. Because of this, there are not many people from that region who participated in the survey. Therefore, there is a break in the continuum.

The mathematical model behind this clustering was made of vectors that represented the percentage people

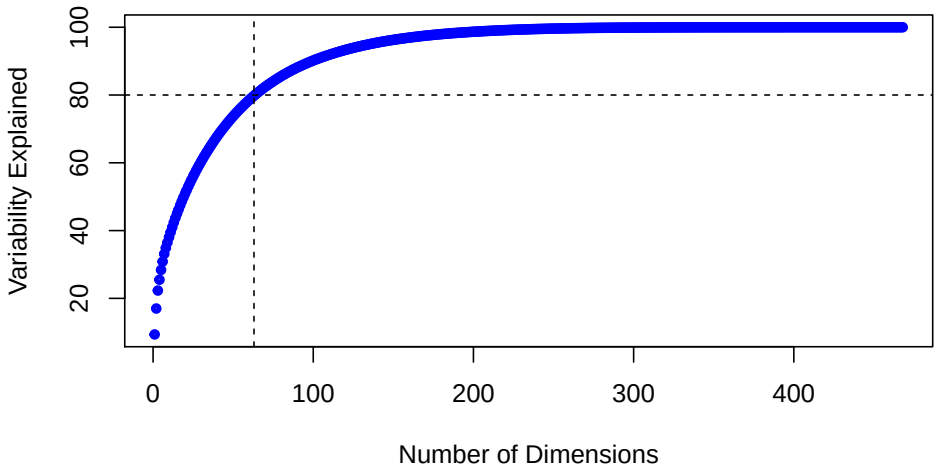


Figure 7: Scree Plot for PCA

Figure 8: Visualizing the data distribution in top 6 PCA components

Figure 9: Elbow curve for k means

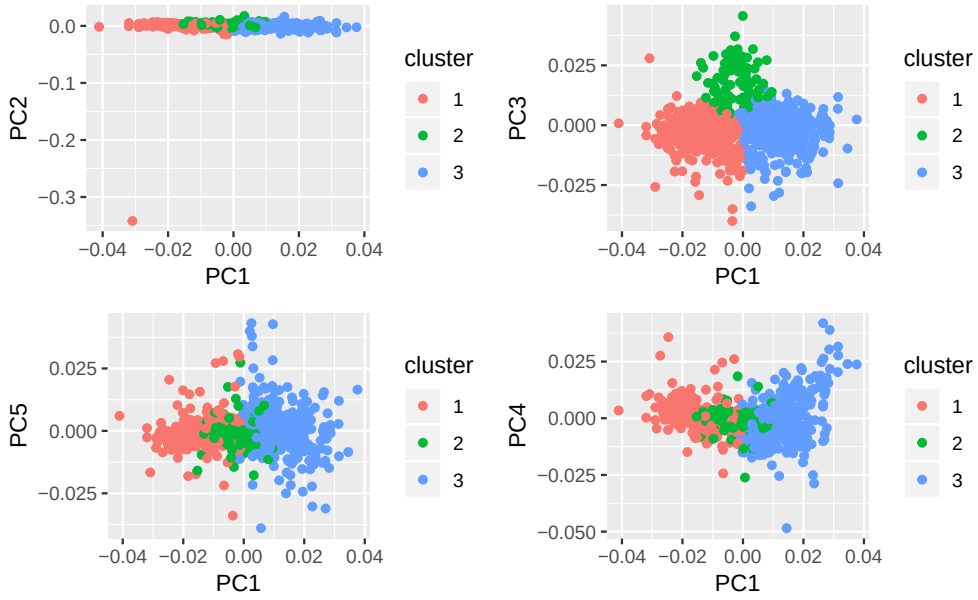


Figure 10: Viewing clustered data in some of the prominent dimensions obtained from PCA

n that small area that chose a certain option of a particular question. This makes sense for these clusters as each cluster is a set of points which is distinct to points in the other clusters in the basic characterization of the proportion of answers opted in the small region. The underlying dividing feature for these groups as we found from figure 11 is thus the geography. Using geography and correlation of one question to another, very thoughtlful and informed predictions can be made.

5 Stability of findings to perturbation

Now, that we have got the data into nice three clusters that represent the three prominent regions of the US, we must question ourselves on whether our findings are stable enough to withstand the noise and perturbations or not. Stability is one of the most prominent feature that validates the correctness and applicability of our findings and results. Therefore, to check our stability, I would sample with replacement the data to change the starting points and test if the result of the interesting finding about the geographical dependencies and intercorrelation of the two questions still hold. We can see in the figure 6 that there is no change in the high level overall division of the regions based on the answers to the two questions after resampling. Thus, we can say

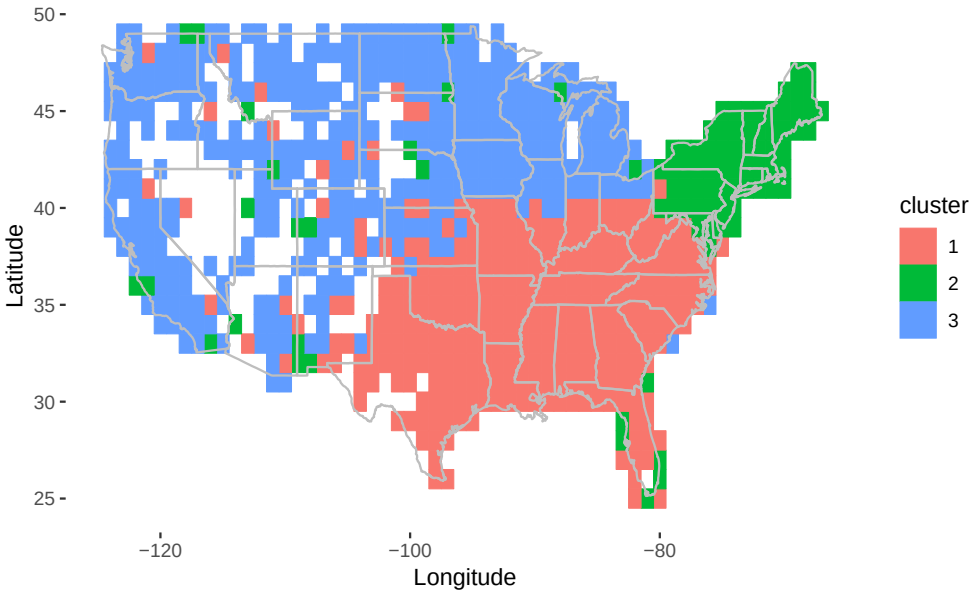
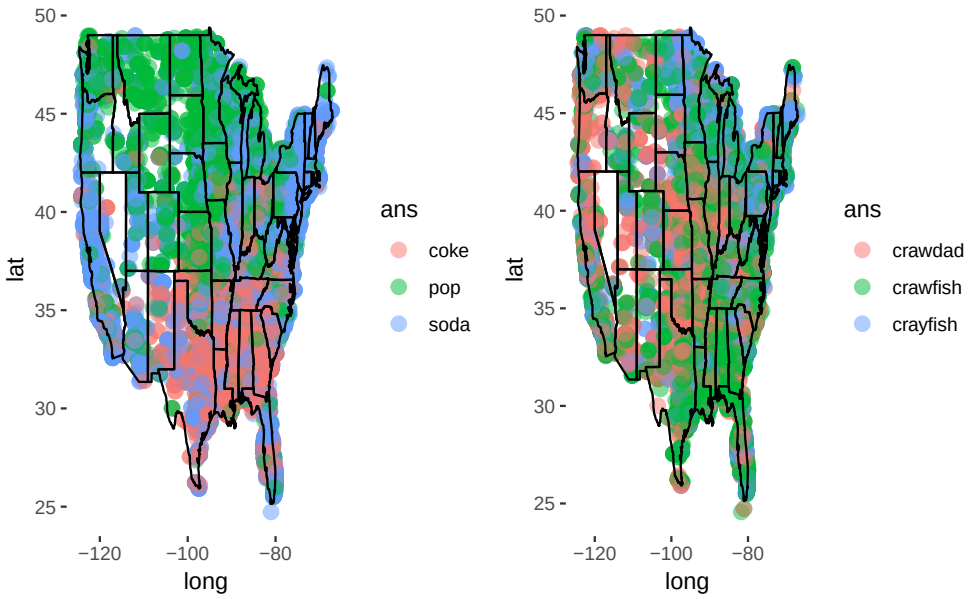


Figure 11: Visualizing k means clustered data on the US map



that our finding is robust and stable.

6 Conclusion

We found that there is a strong correlation between the geographical location and the linguistic terms used by the poeple of the US. This hints towards the fact that due to physical proximity, humans tend to learn each others dialects. If there are more than one options the terms used by the most influencial people tend to become mainstream and dominate the vocabulary of the region. The regional dependency of vocabulary also hints towards collective mindset of the people who interact in trade, knowlege transfer and other activities of life.

References

[1] Gilman Tolle, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, Todd Dawson, Phil Buonadonna, David Gay, et al. A macroscope in the redwoods. In *Proceedings of the 3rd international conference on Embedded networked sensor systems*, pages 51–63. ACM, 2005.