Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

Hector Roux de Bézieux

October 3, 2017

Part I

Kernel density plots and smoothing

1 Kernel plots for the temperature

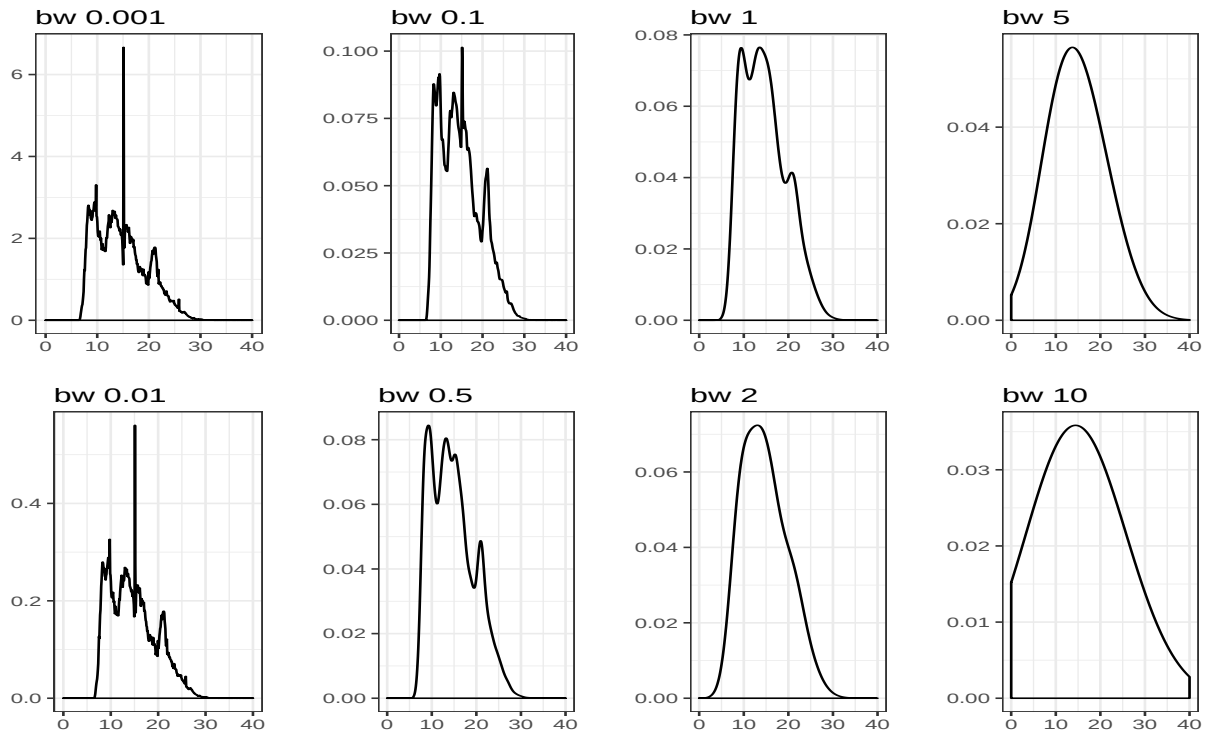


Figure 1: Densities with a Gaussian kernel and various densities

We can clearly see how the plot change as we change the bandwidth. As the bandwidth increases, the curve goes smoother and smoother. A bandwidth of 0.5 is probably the most appropriate here to represent the data (which is is roughly what the automatic setting chooses). Changing the kernels, however, doesn't seem to have any impact on the curve here, probably because of the high number of points being represented for any given value.

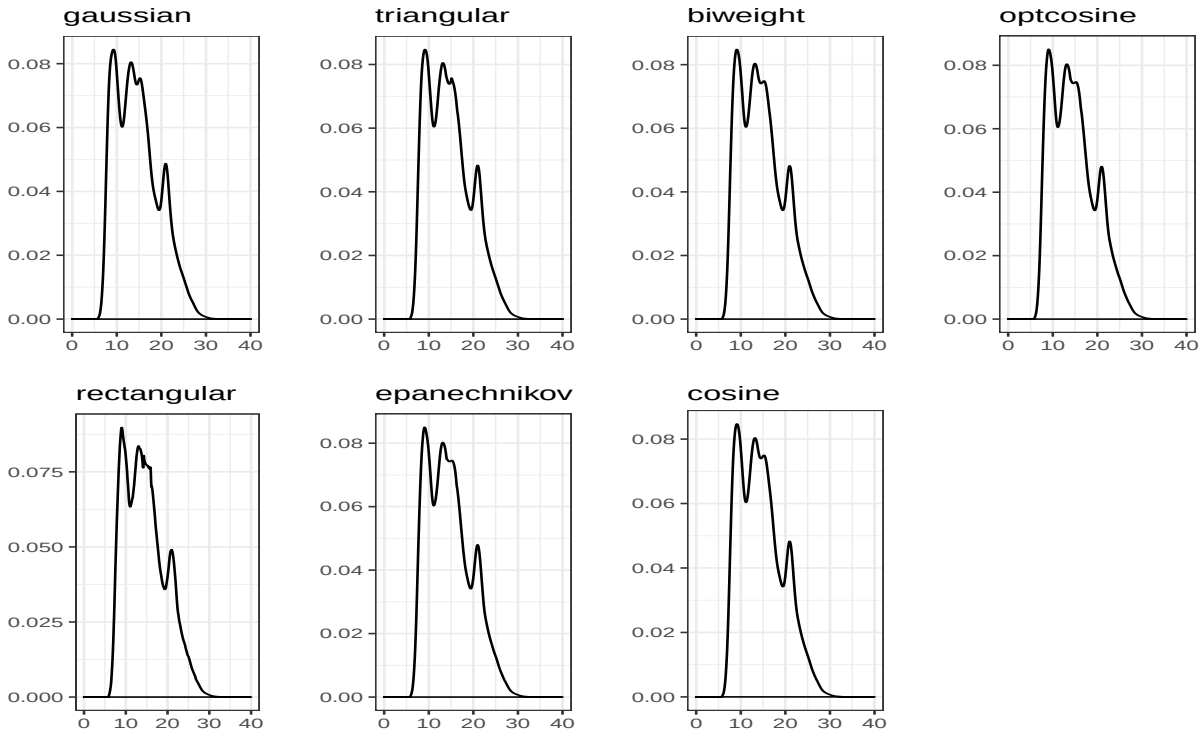
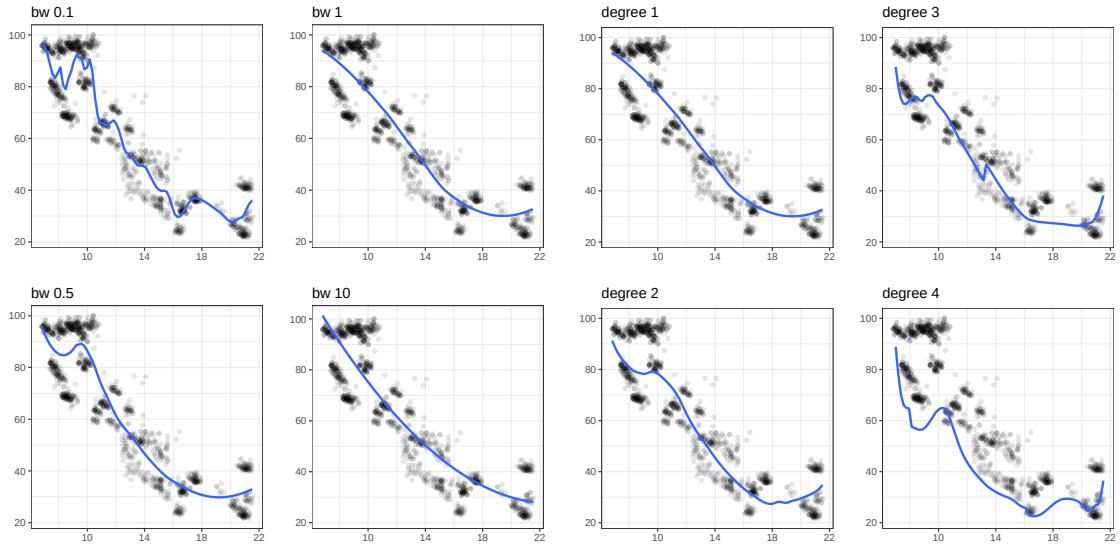


Figure 2: Density with various kernels and a bandwidth of 0.1

2 Temperature versus humidity



(a) Loess fitting with varying spanning values and degree 1

(b) Loess fitting with various polynomials and span 1

In that case, we can clearly see the impact of changing the spanning value (the bandwidth) on the fit. Smaller spanning values mean a closer fit to the data but the fit is also less smooth - and we probably have a lot of bias. On the other hand, a too high spanning value is just a polynomial fit on the points.

When varying polynomial degrees, we have a similar effect. Fitting with a higher degree polynomial allows for a closer fit to the data but the curve is less smooth. On a side note, the loess fit gets extreme when you increase the polynomial degree without specifying a spanning value (not shown). The formula used by R to compute the optimal bandwidth only applies for a linear polynomial local fit.

Part II
Linguistic Data

1 Introduction

The study of dialects has long been of interest to linguists. In particular, the differences of dialects along geographical differences - or dialectometry - helps to explain many variations of local speeches and can inform or validate historical hypothesis on population movement and interactions. While this field has drawn a lot of attention, it has only recently started to use computational methods to better aggregate the vast amount of data collected. This data mainly constitutes of answers to questions asked by an interviewer. Before, the linguist could decide which questions he thought were the most significant and relevant in classifying dialects. Computational methods can avoid that subjective step and point to questions that actually differentiate between different dialects.

2 The Data

Here, the subject of interest is dialects in the English-speaking population of the US in 2003. 67 questions are considered, aimed at exploring which word would be use in a specific situation. Answers are to be chosen among a list, with an "other" option available. Questions includes *What do you call the insect that flies around in the summer and has a rear section that glows in the dark?* or *What do you call a point that is purely academic, or that cannot be settled and isn't worth discussing further?*. Answers for the latter questions would be *a moot point*, *a mute point*, *either one of the above*, *I have no idea* and *other*. The position of each individual is known through their zip-code, city and state (self-reported). The aim of this study is to find relationships between location and specific answers to questions

2.1 Data quality and cleaning

First, there are only 67 questions that are of interest rather than all questions between 50 and 121 (some are related to pronunciations) so they are removed. Then some respondents didn't answer some questions some we remove them.

There are also some missing values for State and City, which in turn mean some missing values for longitudes and latitudes coordinates. There are 0 such respondents and they are filtered out as well. All respondents with non-existing state name are also filtered. 2476 points have coordinates that don't map their states. They are also removed. Finally, for plotting purposes, the states of Alaska (AK) and Hawaii (HI) will not be considered.

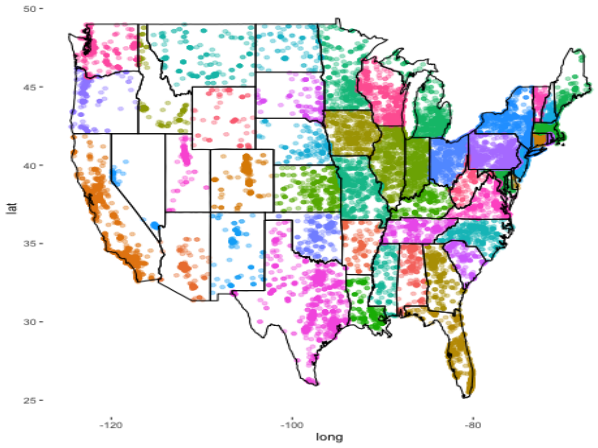
We can plot the resulting map and check that there doesn't seem to be any anomalies anymore in Fig 4a.

2.2 Exploratory Data Analysis

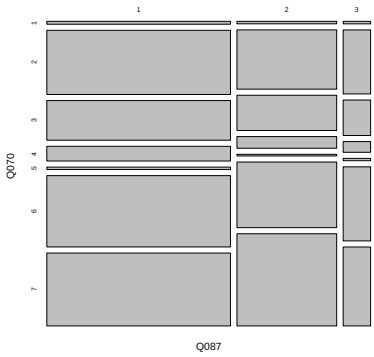
To get a better feel of the data and experiment with visualization, two questions are chosen and their relationship is studied in deeper depth. For this analysis, the focus will be on Q87 *Do you use the term 'bear*

claw' for a kind of pastry? and Q70 *What do/did you call your maternal grandfather??.*

Those two questions are selected and individuals that didn't answer to one or the two questions, so 0 respondents are taken out.



(a) Visualization of all respondents colored by state.



(b) Relative frequencies of each pair of answer

Figure 4

In Fig 4b, a few things are of note. Firstly, all answers are far from being equally likely. For Q118. answers 2,3,6 and 7are far more likely. There is also some relations. People who answered 2 ("no, but I know what it means") at Q87 are more likely to answer with 7 ("other") at Q70.

A $\tilde{\chi}^2$ analysis with $(7-1)\times(3-1)=12$ degrees of freedom allows to check for whether the samples are independent. The $\tilde{\chi}^2$ equals 234.05, the answers to the questions are not independent.

Reducing dimensions is not relevant here and there are already 21 natural clusters based on all possible answers. A few respondents had impossible state names (94 or XX) and they are also removed. Fig 5 gives a visualization of those points on a map of the US.

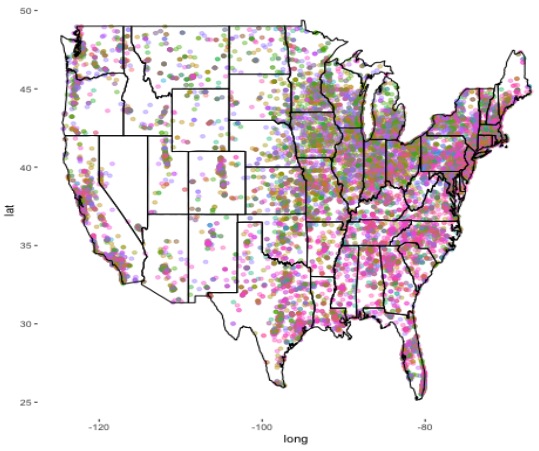
Plotting all the possible combinations of answers, as in Fig. 5a, is not very meaningful for 2 reasons. One is that it is hard to find 21 really distinct colors so some clusters are colored in similar fashions even thought they may not be similar at all. Secondly, the biggest clusters can hide the smallest ones. Fig b-d show some subsets of respondents.

From Fig 5b, we can identify small but strongly localized clusters: one in Pennsylvania (in blue), one around New York (in brown) and one around Boston. If we look at the labels (not shown), we can see that they match rare answers to Q70. Some people from Pennsylvania call their granddad Pap, some from New York and Boston call theirs gramps. The difference between the New York and the Boston clusters are their answer to Q87.

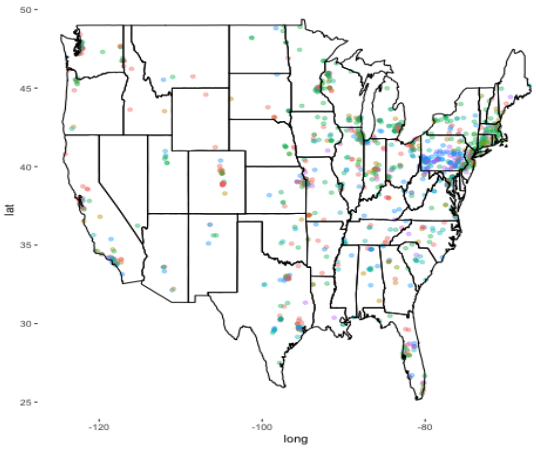
From Fig 5c, we can see the emergence of two regions in the east of the US, even though it is hard to be sure. Fig 5d marks the distinction more clearly by focusing on the 4 biggest clusters. East of -100 long, the US are divided along a line South-West to North-East. The main divider is again Q70. South of the Line, respondents use a different name for their granddad in private and in public whereas north of the Line, they use the same. The differences inside those 2 regions are either due to the name used (for the northern part) or to Q87. West of -100 long, it is hard to sport any particular trends.

Picking two questions to study in depth can be useful to get a better sense of the data. With only two

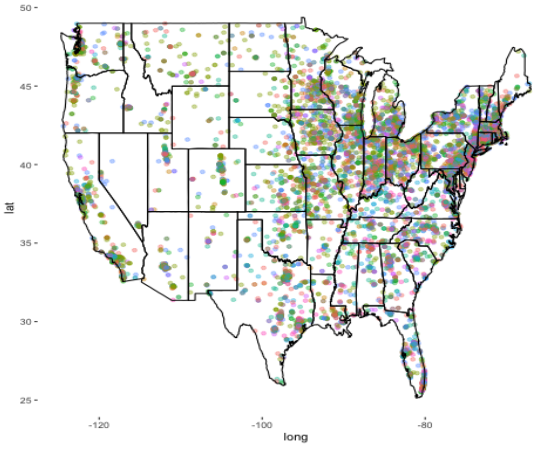
questions, we can already see the emergence of regional trends. However, as could be expected, the most meaningful differences are based on the answers to the question with the most choices (here Q70).



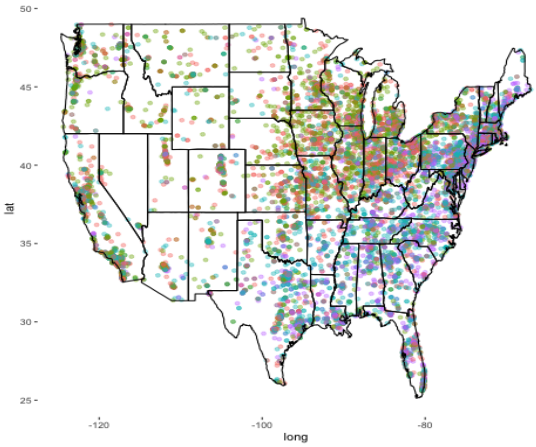
(a) Visualization of all respondents colored by their answers to Q70 and Q87



(b) Subset of all respondents for rare answers



(c) Subset of all respondents for intermediate answers



(d) Subset of all respondents for common answers

Figure 5: Visualizations of respondents based on answers to Q70 and Q87

3 Dimension reduction methods

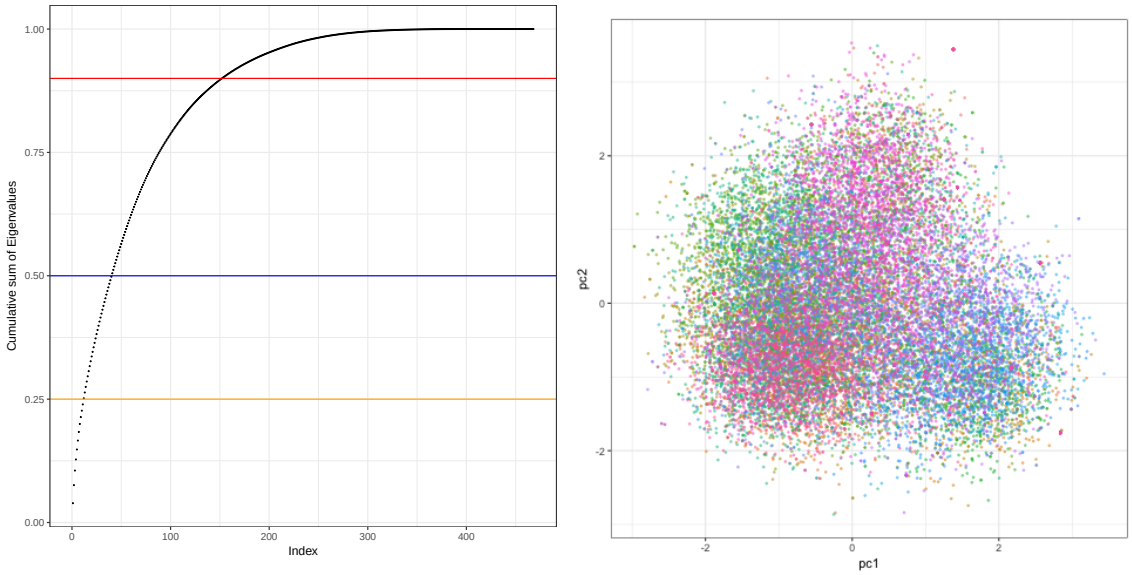
3.1 Binary format

To reduce the dimensions of the data set, it needs to be put in binary format. Then, we sampled 70% of the sample from each state and keep the rest as a test set. Sampling among state ensure that the proportion of respondent by states are conserved. Otherwise, since some states have few respondents, our test might be to sensitive to the sampling.

3.2 PCA

Afterward, Principal Component Analysis is run. As we can see in Figure 6a, the first 155 PC explain 90% of the variance in the data, the first 42 PC explain 50% of the variance, the first 13 explain 25%. The most important answers based on the first 42 PCs are 72.1 and 72.5. Question 72 is *What do you call the big clumps of dust that gather under furniture and in corners?* and the answers are *dust bunnies*, *dust kittens*, *dust mice*, *kitties*, *dust balls* and *other*.

The aim is to see if the geographical distances can be recovered from the distances computed with answers. Here, the implicit distance chosen when using PCA is 1-correlation. Therefore, the points are plotted according to their 2 first PCs in figure 6b. Since there are some many points, it is hard to see much, so mean values are considered.



(a) Cumulative sums of the PCs, with cutoffs at 0.25, 0.5 and 0.9

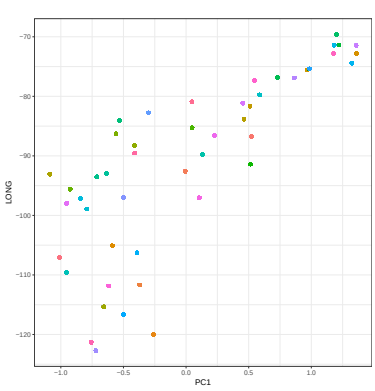
(b) Plotting of the 2 first PCs for all points

Figure 6: Principal Component Analysis

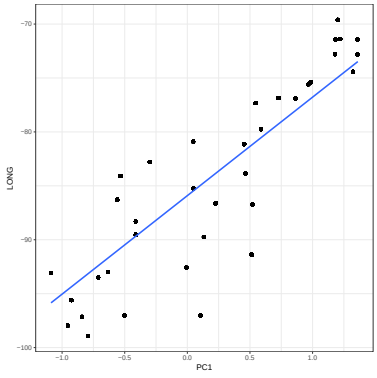
After grouping points per states, the average PC1 is plotted against the average longitude (Fig 7a) and latitude (not shown) to see if there is any correlation. As can be seen on Figure 7a, the mean of first PC seems really linked to the mean longitude for longitudes greater than -100, and excluding Florida. This can be understood since those states are much more recent and may not have time to develop local dialects. The origin of the respondents is more relevant than there location. Filtering out those states lead to Fig 7b and 7c. There seems to be a clear relation between PCs and geographical coordinates.

It is therefore possible to fit linear regressions on PC1 and PC2, based on longitude and latitudes. This can be used to transform the us map and plot it on top of the points, on Fig 8. However, most points actually overflow the map. Therefore, to get a better sense of whether points really cluster by location, only a subset of the points corresponding to well-known regions are plotted, and colored by region. The function *Region* in the code gives a More precise list of which state is in which region.

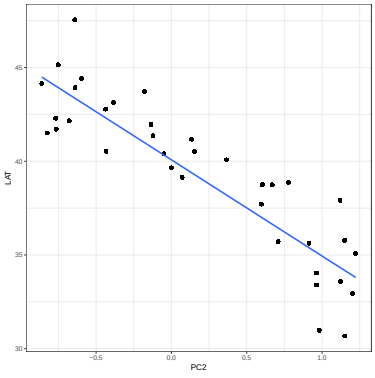
On Fig 9, the relative positions of the regions are well-respected and the Boundaries are quite clear, even if some points cluster with the wrong labels. What is also apparent is that, even though all labelled clusters cover their assigned states, most of the points are outside the map. This can be expected since the distance between dialect probably don't evolve linearly with physical distances, but probably with varying paces depending on the place, the population density, and soon and so forth. However, being able to recover the



(a) Mean longitude versus mean PC1, per state



(b) Mean longitude versus mean PC1, per state east of-100 long, excluding Florida



(c) Mean latitude versus mean PC2, per state east of -100 long, excluding Florida

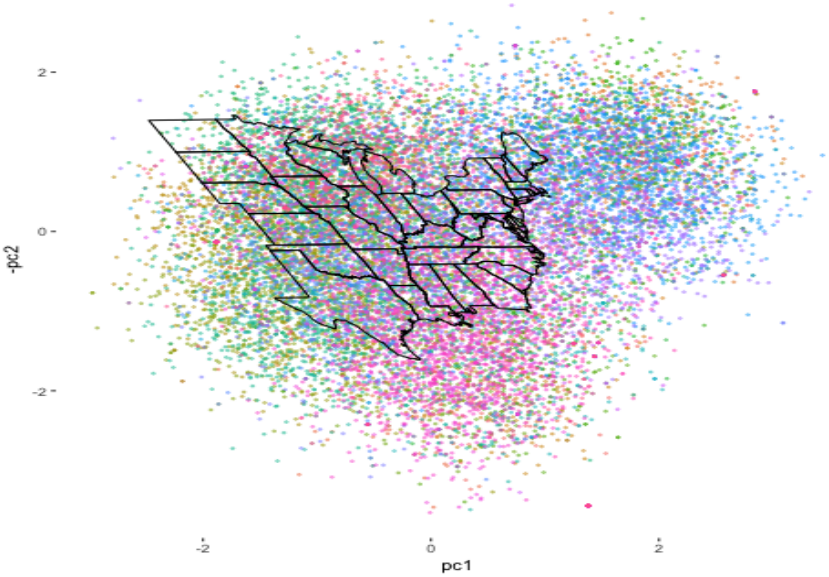


Figure 8: Plotting of the 2 first PCs for all points, with the transformed map added on top

relative positions from the PCA and relatively fit the map can still be satisfactory.

It can also be expected that, even if we get the right relative positions, there won't be a perfect match with the distance. Also, only the first 2 PCs are used and they represent only a small fraction of the variance. Finally, using correlation as a distance might not be very relevant.

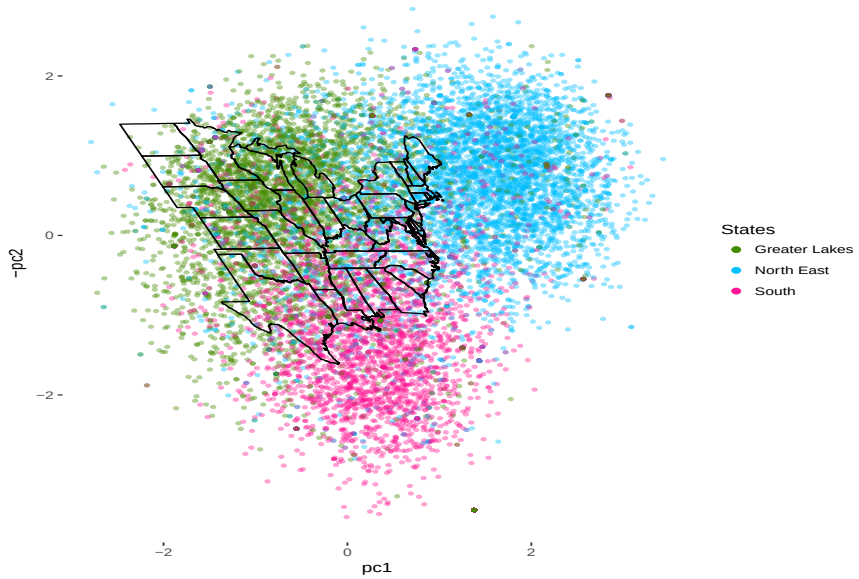


Figure 9: Plotting of the 2 first PCs, for 3 states

3.3 Multi-dimensional scaling

On the binary matrix, an euclidean distance is just the square root of the sums of dissimilar answers minus the average number of answers Using multi-dimensional scaling, this distance matrix can be projected on two dimensions.

Computing a distance matrix on the full training set is not computationally feasible (it would be a 25415 $\times$ 25415 matrix so it weight more than 10GB) so only a subset of the points is considered. Sampling is done by state so that the the repartition of the data remain identical.

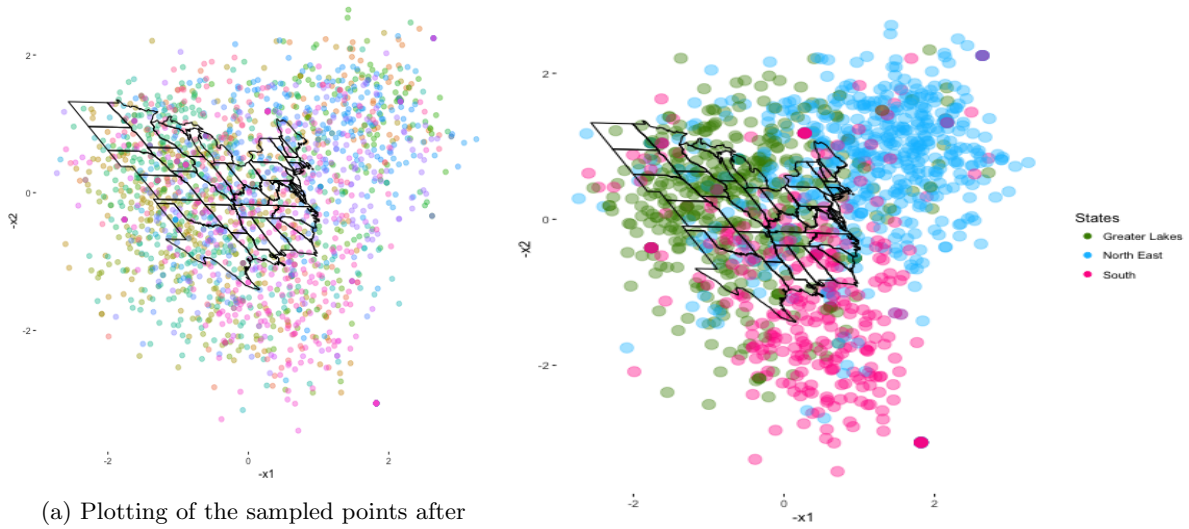
This time, there also is a relationship between the two first dimensions of the MDS, and the longitude and latitudes of the points, on average, for the states east of -100 long excluding Florida (not shown). Therefore, as in the PCA analysis, the US map can be transformed and fitted to the data points which underwent multi-dimensional scaling. This is what is present in Fig. 10a and 10b, which are the equivalents of Fig 8 and 9. The points also overflow the fitted map but we can recover the geographical clusters.

3.4 Comparison of the two techniques

It is hard to know whether the two methods presented are very different since they are not performed on the same data set. Therefore, we only select the points that where used for both methods.

If the results from the previous PCA are used, the Multi-Dimensional Scaling method and the PCA are nearly as efficient: the average silhouette length is 0.138 for MDS, while it is 0.143 for PCA.

If we run the PCA and the MDS specifically on the selected points, the average silhouette length is exactly the same! and is worth 0.158. It is something that could have been expected. The PCA uses 1 minus correlation as a distance, while the MDS was run with euclidean distances. However, in the case of the binary data, it amounts to the same: the first distance counts the number of similar answers while the second counts the number of different answers. After scaling, those distances are therefore identical. Therefore, the two methods will give the same results in that specific case.



(a) Plotting of the sampled points after multi-dimensional scaling, with the transformed map added on top

(b) Subset of (a) with a selection of states

Figure 10: Multi-dimensional scaling

3.5 Clustering

Running a clustering method on the whole training is not computationally feasible, since this also requires computing a distance matrix. Therefore, a sample of the respondents needs to be considered for clustering. Furthermore, running a spectral clustering method (PCA + k-mean) on all states yield unsatisfying results: the clusters are not stable. Therefore, the clustering will focus on the subset of states identified in the previous steps for clustering.

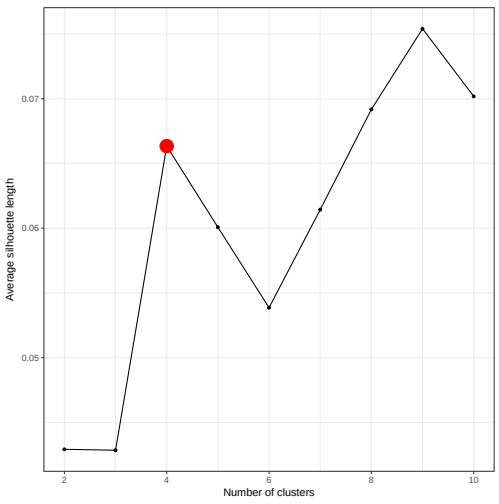


Figure 11: Average silhouette width for various number of cluster

Using pam as a K-mean algorithm after computing a distance matrix using the first 42 PCs (that explain 42% of the variance), the average silhouette width can be computed and plotted for various number of clusters. This can help to determine which k to choose. On Fig 11, k = 4 or k = 9 can be picked as clusters. Using Ockam razor, k = 3 is selected.

On Fig 12, the clusters learned from the data do seem to partially match the geographic labels. The table of frequencies is:

	Greater Lakes	North East	South
1	267	175	68
2	111	117	199
3	58	233	27
4	6	13	18

It is clear that clusters are linked to geography but the relation is not at all straightforward. Working with more data points (and a more powerful computer) may help to better match geography and dialects.

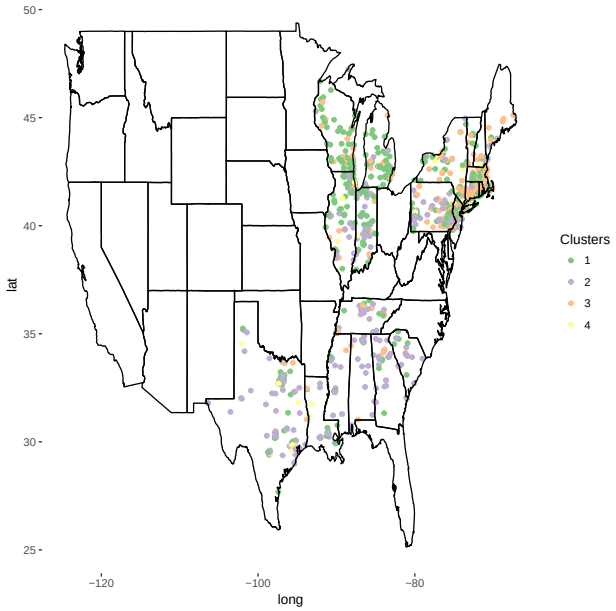


Figure 12: Average silhouette width for various number of cluster

4 Stability of findings to perturbation

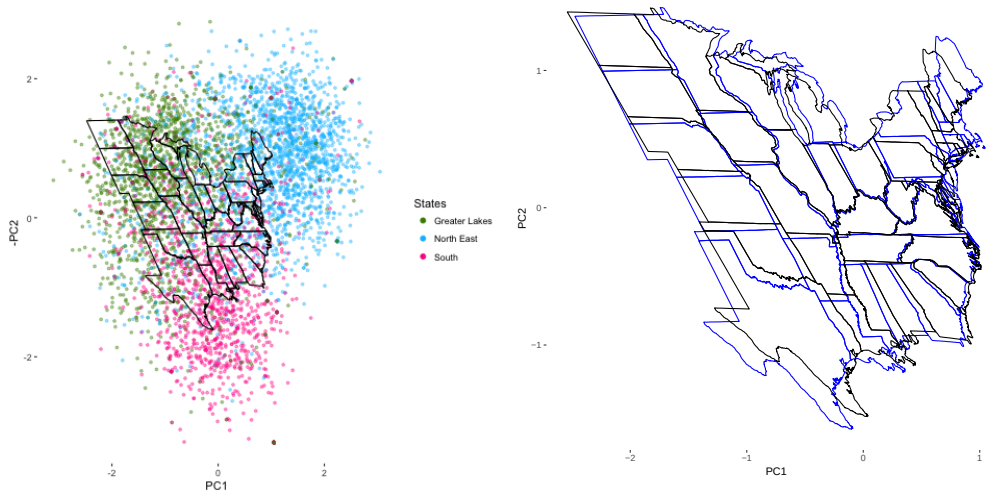
4.1 PCA

Several of the findings relies on random draws so it is easy to test for perturbations. The first perturbation to be studied is the PCA where the test set comes into play in two manners. First, the test set is projected along the 2 first PCs found before and is plotted along with the fitted map to check for consistencies (Fig 13a). Then, a PCA is run on the test set and another map is fitted so it is possible to compare the 2 maps (Fig 13b). The 2 figures show that our results are consistent between the training and the test set and are therefore quite resilient to perturbations.

This also validates the sampling method with grouping by stats before uniform random sampling, instead of straight uniform random sampling.

4.2 Clustering

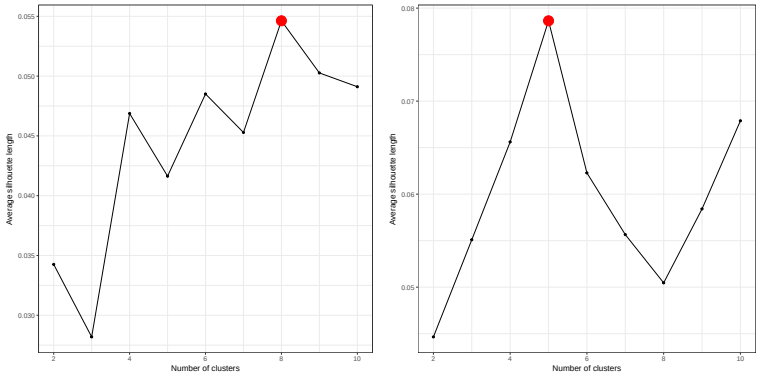
Another part where the findings depend on the sampling is the clustering since the number of clusters and the cluster themselves depend on the chosen sample. Picking different samples and plotting the average silhouette width for various number of clusters lead to Fig 14. Fig 14a - and even more 14b - lead to choices of different k and therfore different clusters (not plotted). As was expected from the preliminary work done on a sample from the whole dataset, the clusters are very unstable and sensitive to data perturbations.



(a) Plotting of the test points after projection

(b) Fiited map for the training (blue) and the test (black) set

Figure 13: PCA robustness



(a) Sample 1

(b) Sample 2

Figure 14: Average silhouette width for various number of cluster

5 Conclusion

Overall, it is quite possible to recover a good sense of the relative geographic positions of the points from their answers to the questions for some parts of the US with older histories. This classification is quite robust to perturbations of the data.

However, trying to define clusters is not possible in any stable manner. Spectral clustering is probably not suited to the task and a softer clustering technique that assigns probabilities of belonging to a cluster might be mor relevant.

The distance metric used here may also lack some precision. Here, every question is weighted the same. On the other hand, domain knowledge might indicate that some questions are more relevant thatn others. The fact that respondents could answer "other' (and that many did) is also a problem. A metric that doesn't take those answers into consideration might be more powerful as identifying clusters.