Lab 2 - Linguistics Data, Stat 215A, Fall 2020

Your name

October 09, 2020

1 Introduction

On this lab we will experiment with computational techniques in Dialectometry, in particular clustering and dimension reduction.
Dialectometry is the study of language variation, on this lab our main focus would be to shed some light on variation determined by geography. The nature of language variation across areas is complex and requires careful analysis hence sophisticated computational methods are needed to gain correct insights. Truthful analysis of language variation is particularly useful in understanding historical, social and geographical factors of language use in society. On this lab we will use data from a Dialect Survey conducted by Bert Vaux and our focus would be questions that look at lexical differences. Our main goal is to find clusters of responses that would be interesting in terms of their geographic properties, to that end we will first re-embed our data into lower dimension and then cluster it using off the shelf methods for both tasks.After obtaining the results, bearing in mind the PCS framework, we will preform stability analysis.

2 The Data

We will investigate the Dialects of American English Survey conducted by Bert Vaux. This datasets contains the responses of 47,471 respondents across the United States along with their self-reported geographical information (state, city and zip code). This is a categorical data set, encoded numerically such that each response choice has a number. We also have an additional dataset with responses aggregated across zip codes. The answers to the suvey questions should reflect language variation, combined with the zip code coordinates we can make an attempt in understanding the structure of the joint distribution of the location with the suvey answers.

2.1 Data Cleaning

- The first issue was na values for latitude, longtitude and state variables, which were simply removed, since there are so few of them and the geographical information is crucial for out analysis.

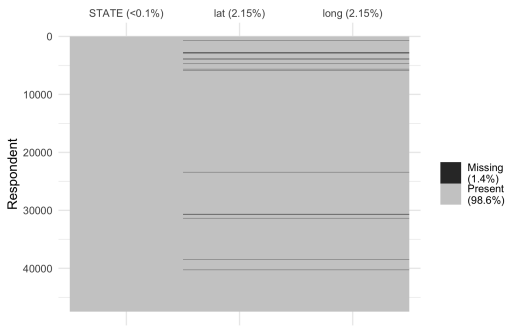

Figure 1: Missing values for each respondent.

- Next we had some longitude values that were too small and they were removed as well. Threshold was -125 data points with smaller values were removed.

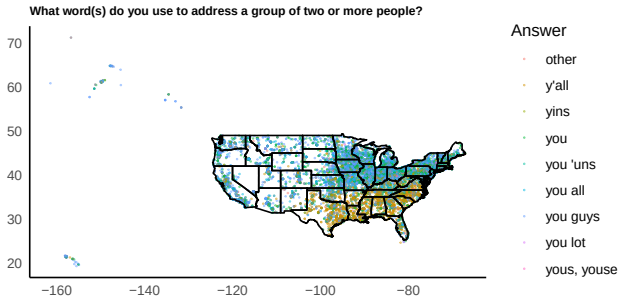

Figure 2: Responces for question 50 geographic show few outliers that were removed

- There were also few typos in the state columns in the data, but since the location data seemed valid and there were very few such responses nothing was done.

2.2 Exploratory Data Analysis

The first thing we wanted to understand is the geographic distribution of the respondents, as expected most responses come from big cities. We can also see that east and north areas have much more responces than west and south.

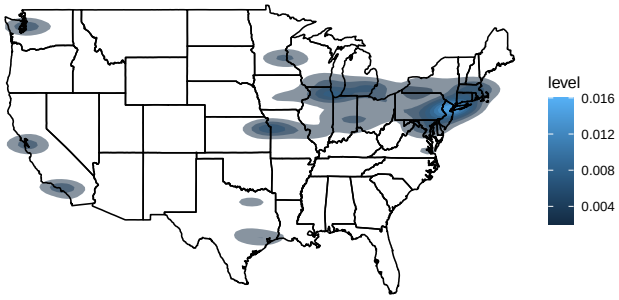

Figure 3: Two dimentional density of responces over the lat and long variables

Next we wanted to look at global statistics and were interested in the general variation of responses, that is since many of the methods are based of metrics similar to variance. The distributions of the first three most frequent answers provides us with an intuition for variance across questions.

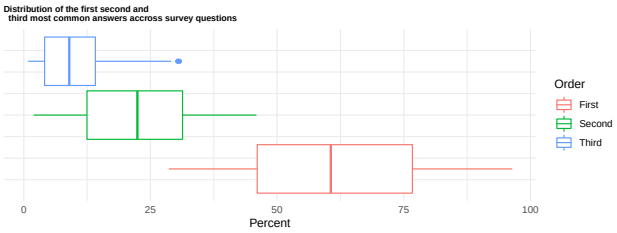

Figure 4: Most frequent responses distribution across all questions

We can see that in many cases a single answer dominates the responses which could be due to the uneven sampling across the states or due to dialects with low variation across the US.

We continue with a detailed look into two particular questions

Table 1: Questions 109 and 120

Question Numner	Question
109	What do you call the paper container in which you might bring home items you bought at the store?
120	What do you say when you want to lay claim to the front seat of a car?

One way to understand if the answers to the questions define any geographical areas is to check how the local distributions of answers to that particular question is different from the global distribution across the US. For each question we look at the two most answered questions, for each state we calculate the the proportion of respondents who submitted each of the answers - those are the observed proportions.
We also calculate the expected proportion under the assumption that the proportion is the same across all states those are the expected proportions under our assumption (in the spirit of FIsher's exact test calculation). We present the log of the observed proportions divided by the expected, which is large when the answer appears more often in a particular state than in the total population.
The reason this metric was chosen as opposed to simply show the observed proportions is that we are interested in finding area that have distinct dialect compared to the entire population hence that normalization.

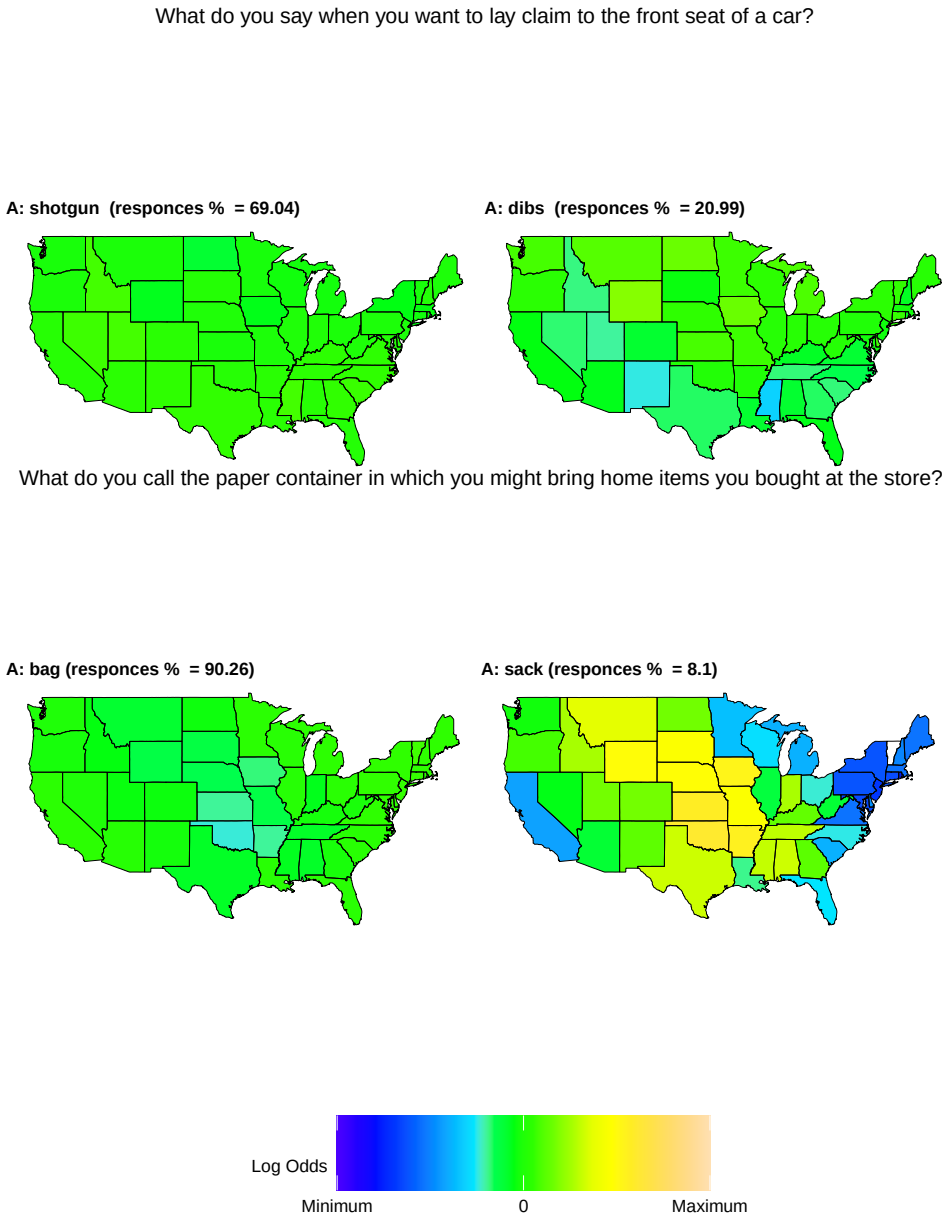

Figure 5: Log Odds ratio of the observed response in each state and the expected values under the assumption that all states are identical it terms of answers distribution

We can that question 109 second response seem to define distinct geographical areas, while the first response does not. Question 120 first two responses doesn't seem to define any geographical area with the distribution similar across states.

As far as prediction go the most naive thing is to first look at contingency tables

Table 2: Joint contingency table for questions 109 (rows) and 120 (columns)

	no answer	dibs	shotgun	hosey	high hosey	I have no term for this	other
no answer	1290	16	54	0	3	13	6
bag	87	8315	27089	193	40	2528	951
sack	14	1706	3319	17	3	564	142
poke	0	24	58	6	6	10	6
other	4	251	518	0	1	164	73

From the table above it is clear that without additional information those variables and not useful predictors for each other as given any answer to one question the best we can do is predict the most common answer to the other (bag or shotgun).

3 Dimension reduction methods

I will explore three methods for dimension reduction PCA, t-SNE and auto encoders. The reason for choosing those methods is that I wanted to find methods as different from one another as possible, PCA is a linear and parametric methods, t-SNE is non linear and non parametric and auto encoder is non linear and parametric. Additionally using auto encoders allows great flexibility in defining the objective function of the optimization problem, as opposed to PCA and t-SNE where I am restricted optimizing euclidean distance and KL-divergence respectively. After obtaining the embedding I will use k-means to cluster each of those methods.

- PCA

The first, most obvious choice for dimension reduction is PCA, that we can simply apply to the full dataset (with the one hot encoding). The reason that the one hot (binary) encoding is important is that PCA is the solution of an optimization problem based on euclidean distances. Those distances would change if we were to permutate the ordering of the answers which is not a desired property of an algorithm since the order of the answers is not relevant to our domain question.

- t-SNE

The idea behind t-SNE to find an embedding of the data that the distribution of distances on that embedding is similar to the distribution of distance on the original data, the distance we used was the euclidean distance and for that reason, the data that was used is the one hot encoding (same as PCA). Unlike PCA our embedding here is not restricted to linear projection of the original data and if fact we are not restricted in any way. On the other hand the optimization problem here doesn't have a closed form solution and is done using gradient decent which take ling than PCA. I was interested to see if the non parametric model would be able to detect gentle patterns that were not feasible using PCA.

- Auto encoders

Auto encoders is a self-supervised method in which a feed forward deep neural network is trained to predict it's own input, one of the layers (called the bottleneck layer) is of lower dimension and represents a compression of the data. We train the network and then use the evaluation of the input on the bottleneck layers as our new embedding. We used a feed-forward network with tanh activation function on each layer, the dimension of the bottleneck was 10 and the dimension of all the other hidden layers is 100. The network was trained on

the entire dataset with Adam optimizer for 20 epochs, the loss was mean squared error. This is a parametric method, which is not linear since we use non-linear activation this model is much more expressive than PCA but not as expressive as t-SNE so in that respect it'll be interesting to see how they compare.

To get an intuition as to how the methods differ in results, we use two dimensional visualization of the embedding obtain by each of the methods:

- PCA - we show a density plot of the first two principal components
- t-SNE - we choose to reduce the data to a two dimensional sub-space, so we simply show the density of this embedding
- Auto encoder - we use PCA on the Auto encoder embedding (10 dimensional) and show the density of the first two principal components.

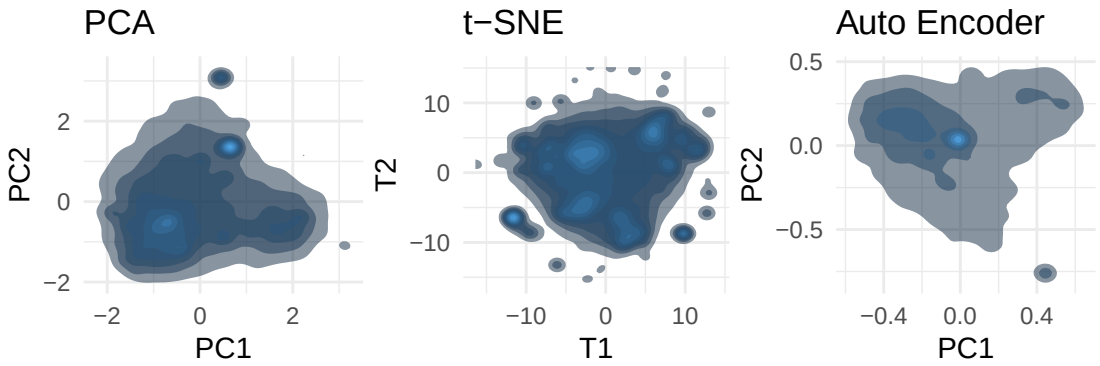

Figure 6: Visualization of the data embedding for the three dimention reduction methods: PCA, t-SNE and Auto encoder

Looking at those density plots we note that:

- the data doesn't seem to have any distinct clusters, which I don't find surprising because it was explicitly stated so in the 2003 paper "The earliest works in dialectology showed that language variation is complex both geographically and linguistically and cannot be reduced to simple characterizations.".
- The t-SNE embedding seem to have more clusters that Auto encoder and PCA (the density have more peaks).

3.1 Clustering

The next step is to cluster the data for the three embeddings we obtained.

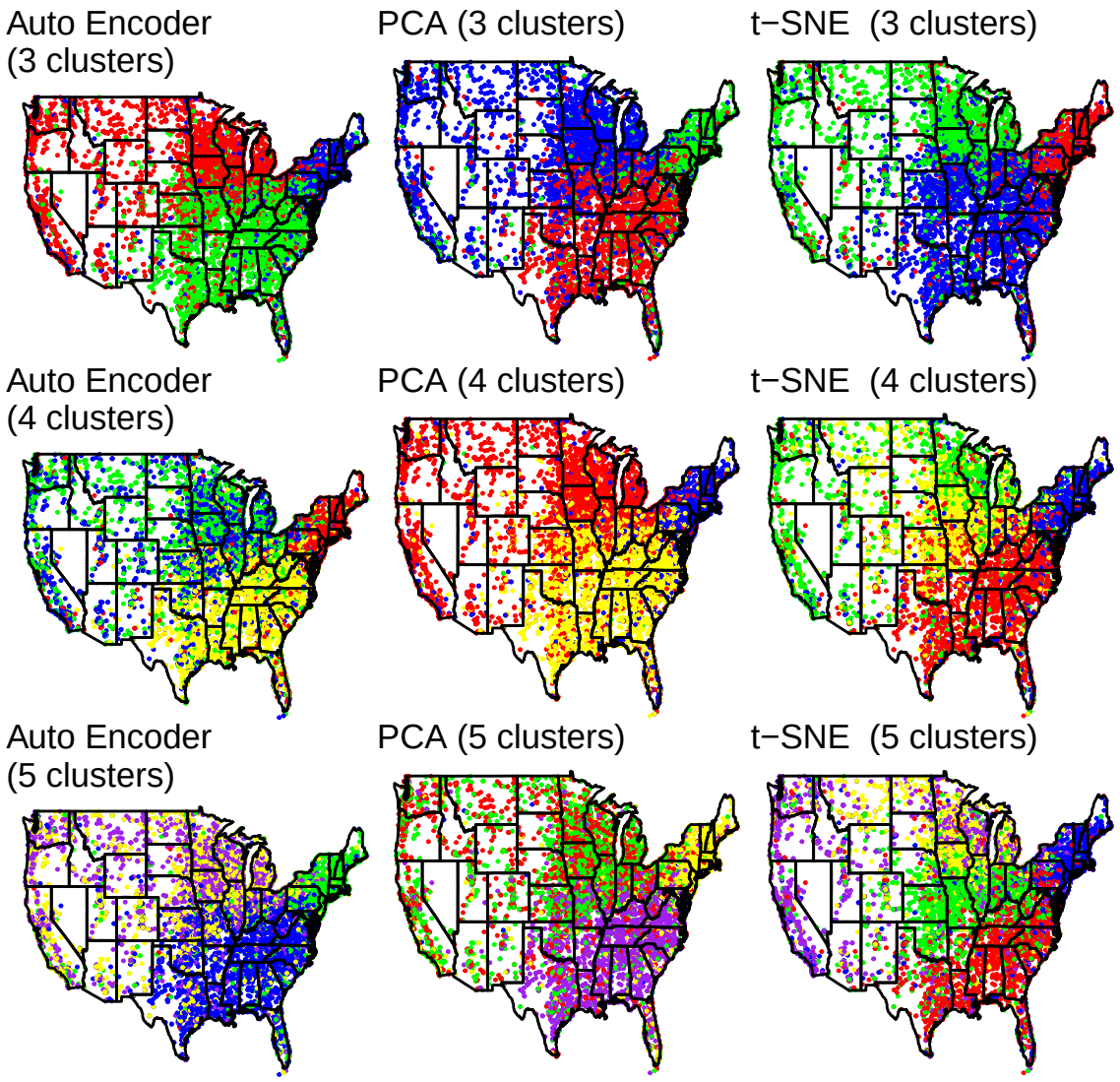

Figure 7: Clusters produced by applying k-means (for k=3,4) to each of the reduced dimention datasets, different colors indicate different clusters

Few interesting things to note:

- Looking at the different clusters we can note that for the choice of 3 clusters all methods produced similar results, which makes a stronger claim for us to believe those areas have a common dialect.
- PCA doesn't seem to have a forth or fifth cluster in the data which can be explained by the linearity limitation.
- t-SNE seem to capture a forth and fifth clusters whereas the auto encoder seemed to learn distances that are not geographically meaningful. It would be interesting to see if changing the loss function of the auto encoder may produce similar results as t-SNE in terms of clustering.
- We focus on 3,4 and 5 clusters since adding more clusters did not add any clusters that make sense geographically, and using less is not very interesting.

Shifting back to the domain problem we would like to be able to identify the questions that separate the clusters, as they indicate the language variation. To the end we look at the three clusters formed by t-SNE and k-means and pick a few questions for which the most common answer is not the same in all clusters. Those questions are probably the ones that separate the clusters.

Table 3: Question for which the most common answer is different among the clusters

	cluster 1	cluster 2	cluster 3
What word(s) do you use to address a group of two or more people?	you guys	`y'all`	you guys
Would you say 'where are you at?' to mean 'where are you?'	`no`	`yes`	`yes`
Which of these terms do you prefer for a sale of unwanted items on your porch, in your yard, etc.?	yard sale	garage sale	garage sale
What do you call the insect that flies around in the summer and has a rear section that glows in the dark?	I use lightning bug and firefly interchangeably	lightning bug	`firefly`
What do you call the miniature lobster that one finds in lakes and streams for example (a crustacean of the family Astacidae)?	crayfish	crawfish	crawfish
What do/did you call your maternal grandfather?	`other` (including if you use a different term to address him directly than you do when speaking about him to a third party)	`other` (including if you use a different term to address him directly than you do when speaking about him to a third party)	I spell it 'grandpa' but pronounce it as 'grampa'
paternal grandfather?	`other`	`other`	`grampa`

We can also repeat the log odd analysis for some of those questions

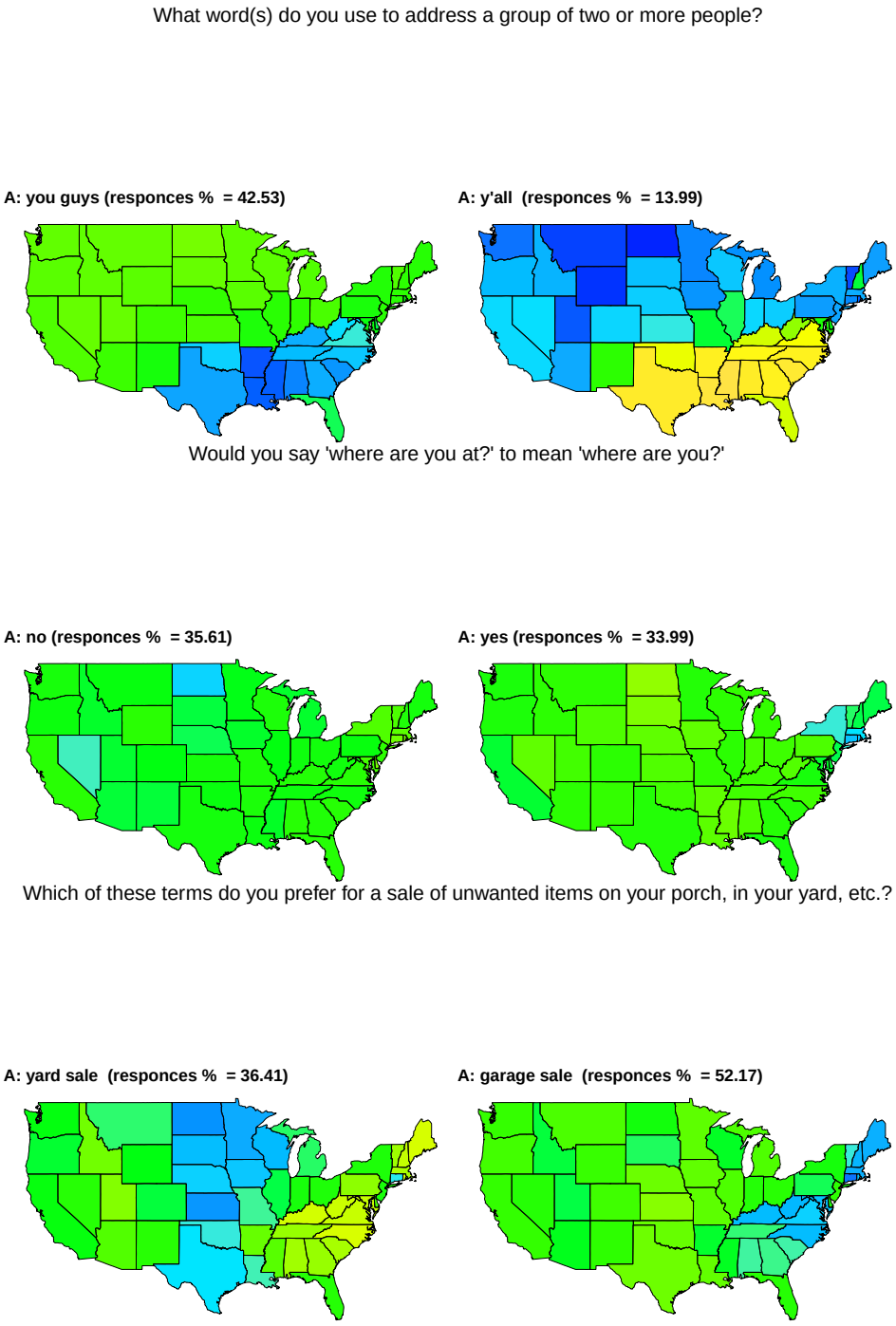

Figure 8: Log odds ratio for three of our suspected seperating questions

- The first question stands out in terms of geographical separation, we see that "y'all" is much more common in the south as opposed to other areas, "you guys" displays the opposite behavior, it is important to note that less than 50% of the respondents submitted one of the two answers so we cannot know who the rest of the respondents fit in terms of this geographical separation.
- The second question does not seem to have an interesting distribution, it may be noise or contribution through interaction with other questions.

- The third question has a more balanced distribution than the first, on a larger percent of the respondents hence it does seem to hold valuable information for clustering that is also geographically meaningful.

4 Stability of findings to perturbation

On this section we'll focus on the t-SNE based clustering with five clusters and try to understand if this finding is stable or simply luck. To that end we'll run t-SNE two more times bootstrapping our dataset. For each of those runs we'll use different starting point for gradient decent as well as k-means and we'll see how the results compare to the original ones (with new k-mean initialization).

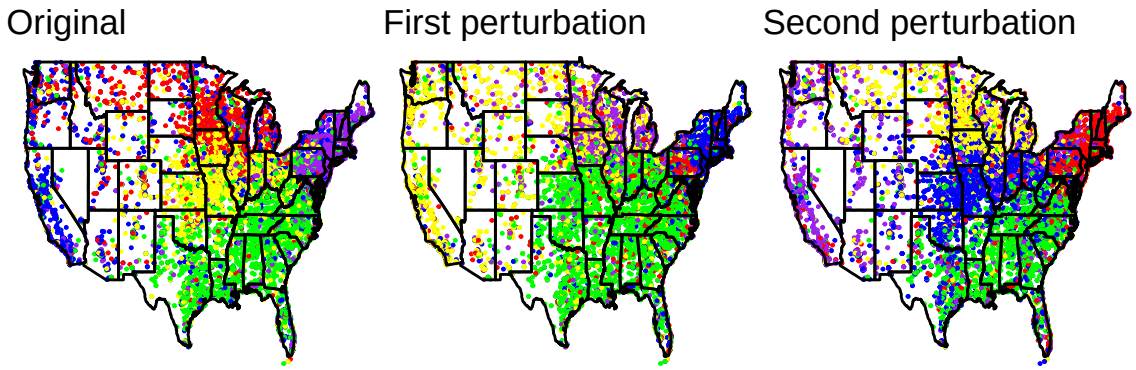

Figure 9: Five clusters obtained by using t-SNE and k-means for our original data and two bootstrap datasets

We can see that in one of the perturbations we performed we obtain similar, geographically meaningful, clusters and that in the other we "lost" one of the clusters. We can also see that changing the k-means initialization did not change the resulting cluster significantly.
I think that doing two perturbations is not enough to reach a conclusion, my intuition is that if we were to perform many more we would see meaningful clusters in many cases. As my computer is not suitable for running so many calculation I can not know this for sure so it remains a speculation.

5 Conclusion

As far as I understand the clusters we were able to find a common sense to most Americans so in that respect we didn't shed any light on new discoveries. I think what we can take forward is that clustering methods are able to extract similarities from linguistic data sets and perhaps can be useful in future research, for example if two areas appear consistently over many perturbations on the same cluster it may be interesting to understand why.

The reality check I did, as non-American, was to show the results to a few people in my class and ask them if that makes sense to them, I think the next step would be to consult experts in the field for interpretation of results and understand if they are aligned with what they know.

Given more time I would probably play some more with Auto encoder and see if other loss functions can produce interesting cluster as well as player with hyper=parameters (number of layers, dimension, learning rate, etc. . .). I would also so a more through stability check on the t-SNE results, as well as see what happens if we play with the hyper-parameters of this model (mainly the dimension we project to).

6 Academic Integrity Statement

- I state that is work is my own and that I didn't copy or cheat in any way. I value academic integrity highly and I think that only benefits me as a students to work hard and use this lab to learn and develop as a scientist