Lab 2 - Linguistics Data, Stat 215A, Fall 2021

Anonymous

7th October, 2021

1 Introduction

In this report we analyse linguistic data against the backdrop of the work of Nerbonne and Kretzschmar [2], who build on their own previous work [1]. Given my own inexperience with computational linguistics, even compared to my experience with the Redwood Forest data, I primarily defer to the subject matter experts when it comes to explaining relevant terminology and clarifying the current state of the domain. To suitably place our analysis in the domain context, we first establish definitions: we use Nerbonne and Kretzschmar's definition of dialectology as the study of dialects, while dialectometry is the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography. Notably, while geography is the primary determinant, dialectology studies all kinds of ways languages can vary, be it on the basis of age, gender, social factors, or geographical causes. Dialectology is a well studied and represented subset of variationist linguistics, which is concerned with studying how linguistic variations are correlated in terms of class, age, sex, occupation, etc.

Notably, the need for large-scale computational tools in the context of dialectology stems from the fact that language variation and the corresponding data are complex both geographically and linguistically, resisting naive characterisation attempts. Advancements in computation and techniques developed for dialectology have also benefitted the broader field of linguistic variation, thereby providing insights into historical, social and geographical factors of language use in society. For example Gilliéron showed the correspondence between certain linguistic and cultural divisions in France. Thus by inverting the relationship between culture and language, we may attempt to infer about the connections and interactions between people who speak related varieties of the same language.

In dialectology, there has been an ongoing debate about the existence of so-called dialect areas vs. a dialect continuum. As per Nerbonne and Kretzschmar, a dialect area is an area distinguished from its neighbors by its relatively more limited range of linguistic variation. While there was has been some success in developing consensus about what these areas are, there has been little success in characterising them. In contrast, some linguists suggest the aforementioned dialect continuum, where rather than having clearly demarcated discrete dialect areas, there is a gradual transition from one to the other. Notably this theory has gained prominence since dialectologists have been unable to irrevocably prove the existence of dialect areas and by extension characterise them.

Most non-computational studies have focused on a small number of features, meaning they were ineffective at characterising aggregate levels. Originally there was resistance to aggregate characterisation since there were almost always some contra-indicating tendencies in large scale datasets. While this is still true in general, Seguy [3][4] pioneered dialectometry and took substantial steps in solving this problem by simply counting the number of overlapping features between any two data collection sites. The two primary benefits of this approach were its simplicity and its ability to be applied to most pre-existing dialect atlas projects, whose data were mostly collected by questionnaires with a limited number of possible answers. In Seguy's method, sites that gave the same answer to a given question were counted one point more similar than sites that gave different answers. Crucially, conditioned on a pre-determined set of possible answers or categories, this technique was generalisable to pronunciation and other linguistic features. Thus through this technique of counting differences, Seguy was able to aggregate individual differences over a large amount of material.

While Chambers and Trudgill [5] were initially skeptical of the utility of dialectometry, Goebl [6][7] was able to convince them by elaborating on dialectometrical ideas and systematically demonstrating their potential. Among his numerous contributions, he devised a measure which counted infrequent (and therefore unlikely) matches more heavily and used it to obtain more satisfying results than those using Seguy's naive measure.

In my analysis, I will use computational tools and aggregation to analyse empirical evidence (or lack thereof) of dialect areas and the dialect continuum. Specifically, using clustering and dimensionality reduction techniques, I will examine geo-linguistic data on a binary and continuous scale, investigating whether clusters identified based on certain linguistic data can be used to make predictions about other linguistic peculiarities; in other words, can one's answer to a certain lexical question be used to predict their answer to another? More details are provided in the next section. I will also examine situations where this might not be possible ie. where the geo-linguistic data is not useful in demarcating groups of people. Given that it is not necessary for every facet of language to be different for dialect areas to exist, the aforementioned situation is not a failure of dialect areas, yet we should be aware of the possible limitations associated with using linguistic variation as a tool for demographic analysis.

Furthermore, I am interested in the role urbanisation and the internet may have played in influencing linguistic variation at scale. In other words, I am interested in examining if cities in different parts of the country are similar linguistically to each other even if they differ geographically? On that note, what is the impact of urbanisation on linguistics for regions that are close geographically? While it may not be possible to fully answer this question based on the given data and the scope of this report, I will attempt to touch upon this question.

2 The Data

We begin by qualitatively and quantitatively describing the data. In this report we analyse data from a Dialect Survey conducted by Bert Vaux. To better understand the data, I attempted the survey, available here. The questions and answers are available in the file `question_data.Rdata`. The questions of interest are those numbered 50-121 since they pertain to lexical differences which we will be focusing on, as compared to phonetic differences. To describe the two other datasets, I refer liberally to the provided lab instructions. `lingData` contains the answers to the aforementioned lexical questions for 47,471 respondents across the United States (US). The dataset contains the variables ID, CITY, STATE, ZIP, Q50 - Q121 (a few questions in this range are left out), lat and long. ID is a number identifying the respondent. CITY and STATE were self reported by respondents. The variables starting with Q are the responses to the corresponding question on the website. A value of 0 indicates no response. Nonzero values should correspond to the responses on the website, i.e. a value of 1 should match a response of (a). For the second data set, `lingLocation`, the same categorical responses were turned into binary responses, after which the data were binned into one degree latitude by one degree longitude squares. Within each of these bins, the binary response vectors were summed over individuals. The rows are not normalized, therefore each row can be interpreted as count data for a given square, bearing in mind that different squares can have different numbers of respondents. To concretely clarify the difference in datasets, suppose that persons (i) and (ii) take the questionnaire for two questions. The first question has four answer choices and the second question has three answer choices. If person (i) answered D and B and person (ii) answered A and B, then `lingData` would encode two vectors: (4; 2) and (1; 2). If they lived in the same longitude and latitude box, then it would be encoded in `lingLocation` as one vector: (1; 0; 0; 1; 0; 2; 0).

We are interested in using this data to reason about dialect areas and more broadly, linguistic similarities and differences between people living throughout the US. To this end, we will examine the distribution of their responses to multiple questions to determine the situations when clustering and/or dimensionality reduction can be combined with known linguistic data to predict lingusictic trends. Put more concretely, can we cluster people in such a manner that the clustering is consistent with both geographical location and linguistic variation based on the questionnaire? Furthermore, can we use these clusters to make any sort of prediction about the response of people to other questions pertaining to linguistic differences, especially lexical differences? If there are certain situations where it is possible to do so but not others, can we determine what the relevant geographical and linguistic factors are? Thus we are interested in assessing both the strengths

and weaknesses of this approach. By doing so, we are better equipped to evaluate the dialect area vs. dailect conintuum debate, and understand the role of geography in linguistic variation. As discussed above, geography is the primary factor studied in dialectology, which in turn is an important subset of variationist linguistics. Deriving from Gilliéron's work, we hope that our analysis will also help one infer about the connections and interactions between people who speak related varieties of the same language.

2.1 Data Cleaning

We first examine which questions we do not have the response data for. For the lexical differences questions (50-121), we are missing data for questions 108, 112, 113, 114, 116. Since we have no external source of data to rely on, we treat the responses to these questions as unknowable and do not consider them in our analysis. Every respondent has at least one question they did not respond to, and therefore we do not automatically disacrd observations with at least one missing response, otherwise there would be no data left to analyse. Furthermore, while `lingData` has 47471 observations, the ID column takes values from 1 to 50064, meaning approximately 2600 observations are missing. As before, we are unable to recover this information, but nontheless we should be aware of this in our analysis. In terms of location, we notice that approximately 1000 observations have missing lat and long values. On examing these observations more closely, we notice that the missing lat and long values come from 200 different zip codes, and the zip codes with the most missing values are not close to each other. Thus we conclude this error is not systematically impacting any particular zip code or state. Given the size of the data set, we simply exclude these obsevration from our analysis. Furthermore for ease of visualisation, we wish to focus on data only originating from the continental US. Thus we exclude the 200 or so observations that correspond to latitude and longitude from Alaska and Hawaii, and leave the geo-linguistic analysis of these states for future researchers.

2.2 Exploratory Data Analysis

To begin, I wished to visualise the geographical distribution of highly symmetric (where all options were voted for almost equally) and highly asymmetric (almost all the votes were for a single option) responses. I looked through the percentages in `all.ans` in order to determine which questions had the properties I was looking for. I found that the responses to Question 52 "Would you say 'where are you at?' to mean 'where are you?' " to be highly symmetric, with all three responses being chosen between 30% and 36% of the time. Thus the responses are visualised in Figure 1. Note that the first response is trimmed for spatial reasons in the graph; it is "I can use"where are you at" in contexts such as asking someone how s/he is coming along on a project, but not in the general sense of "where are you physically located in the world at this moment".

Here we realise that visualisation can be deceiving. Even though the first option was chosen the least number of times, it seems to dominate the map due to its usage around population centres. We should bear this in mind for future geographic visualisations.

On the other end, I now visualise the response to Question 63, "What do you call the drink made with milk and ice cream", which has over 96% of the respondents choosing "milshake/shake" as their answer, in Figure 2.

A similar example is plotted in Figure 3 for Question 81, "When you are cold, and little points of skin begin to come on your arms and legs, you have -", with over 90% of the respondents responding "goose bumps"

As we can see in Figures 2 and 3, it is not appropriate for us to arbitrarily choose 2 questions from those being considered and visualise their geographical distribution and correspondence to one another, because if we did so, we could visualise two questions whose responses seem to have very similar geographical clustering, but in reality this is simply the combination of a population map coupled with questions where almost everyone has the same answer. This is one of the primary criticisms of Chambers and Trudgill regarding early dialectometry. Thus we try to solve this by examining questions whose responses do not fit this category. While exploring the questions, I decided to choose Question 95 "What is 'the City'?" as one of the questions to examine. As someone who first came to the US for college, 'the City' has always meant San Francisco to me because I've been in the Bay Area this entire time, yet that isn't even one of the possible responses. Since

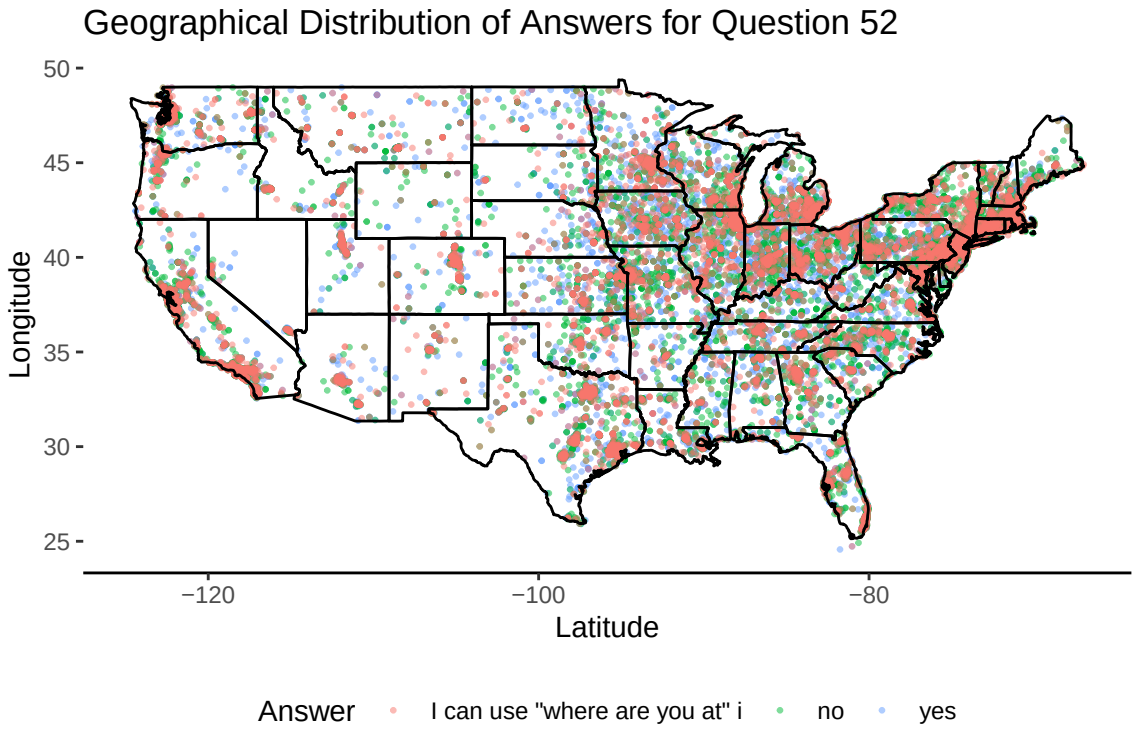

Figure 1: Geographical Distribution of Answers for Question 52

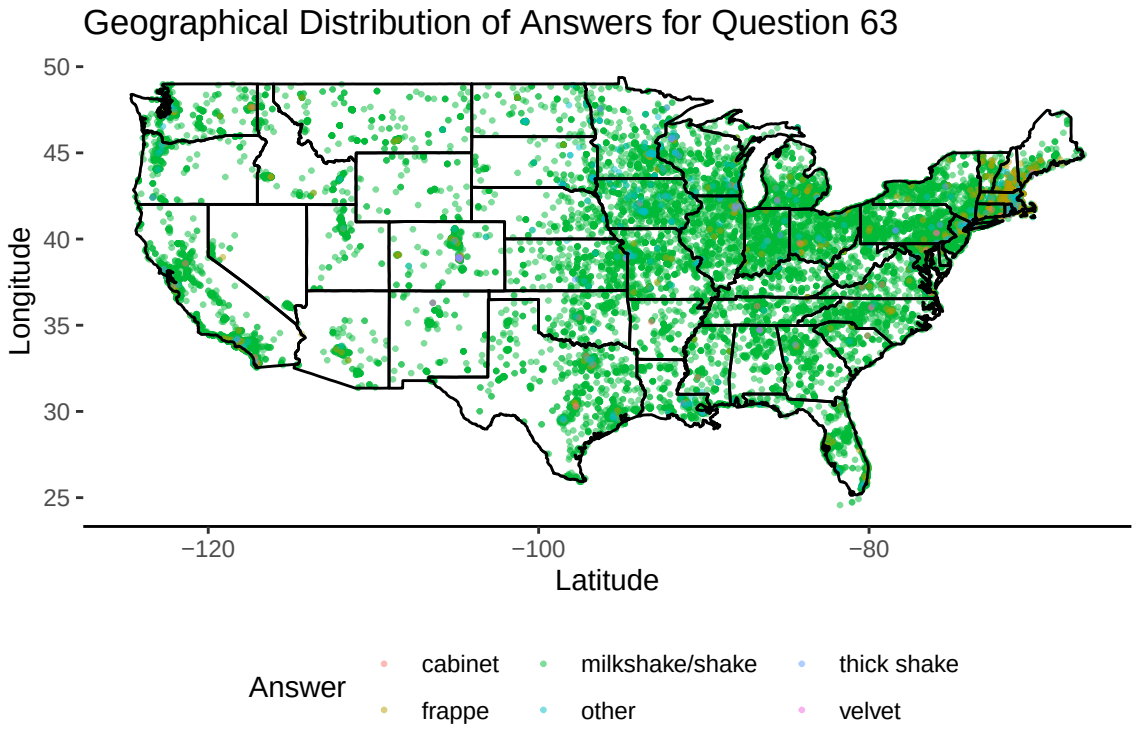

Figure 2: Geographical Distribution of Answers for Question 63

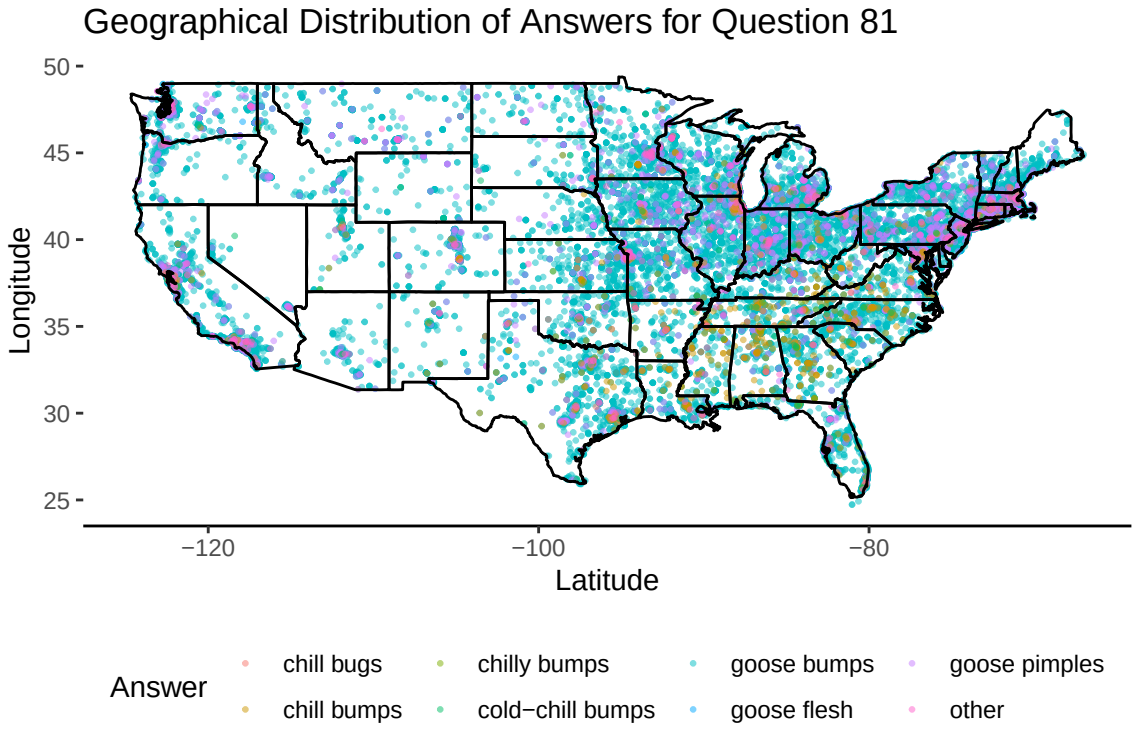

Figure 3: Geographical Distribution of Answers for Question 81

'the City' is a highly localised term, it can be an effective tool at finding dialect areas or examining different parts of the dialect continuum. The distribution of responses is available in Figure 4 below.

We note that even though a majority of people voted for "New York City" (47%), the map appears to be dominated by "Other" (42%) because people from around the country have a different 'the City' in mind depending on where they live. For comparison to Question 95, I chose Question 85, "What is the thing that women use to tie their hair?" since the geographical distribution looked similar enough to me to make it a suitable choice to experiment with, without falling into the pitfalls associated with Figures and 3. The distribution of the responses is available in Figure 5.

We notice a similar structure in the responses to Question 85 as in the case of 95, suggesting this is a suitable pair of questions to compare.

3 Dimensionality Reduction

We now perform dimensionality reduction, exxperimenting with PCA and then NMF. To perform PCA, we first one hot encode the responses to the questions. The justification for doing so is that in the case without one hot encoding, if person (i) answered A and B to questions 1 and 2, and person (ii) answered B and D to questions 1 and 2, then the encoded vectors are (1; 2) and (2: 4), respectively. Interpreted as vectors in feature space, this implies that person (ii) is somehow twice the value of person (i), which is undesirable since we are dealing with categorical data. Since the response data and the lat and long data are on different scales, it is necessary to scale the variables. Furthermore since the relevance of lat and long is more in the relationship to each other rather than the magnitude, we center those two variables only, while leaving the one hot encoded responses uncentered for interpretability purposes, and also to leverage the sparsity of the feature matrix. The results are visualised in Figures 6 and 7.

As we see, the plot is effective at distinguishing the no responses from the responses. Furthermore, we see roughly three groups along the second PC in both Figures, suggesting that we may experiment with our clustering algorithms for 4-7 clusters.

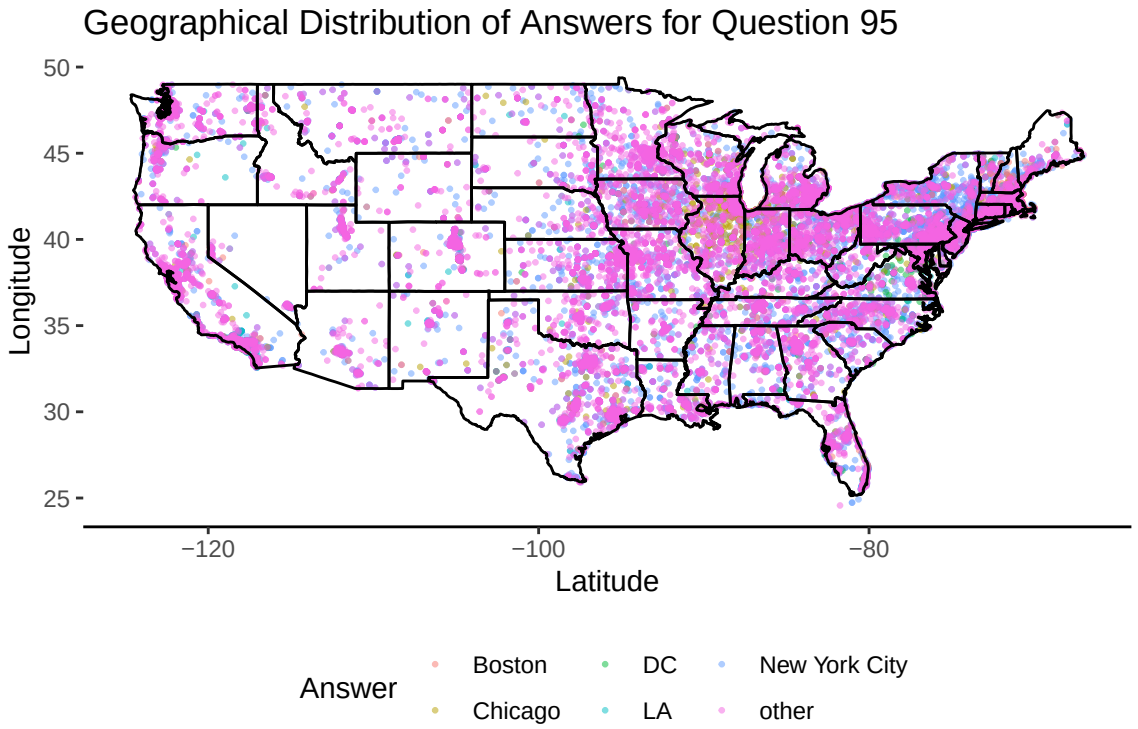

Figure 4: Geographical Distribution of Answers for Question 95

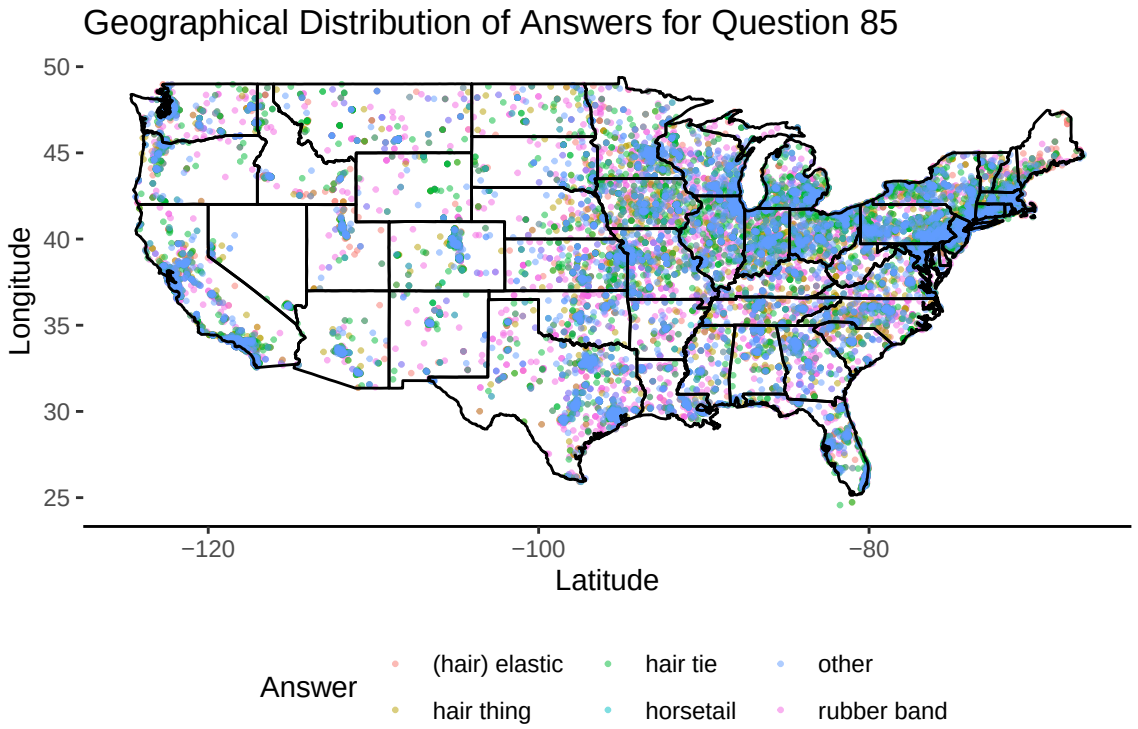

Figure 5: Geographical Distribution of Answers for Question 85

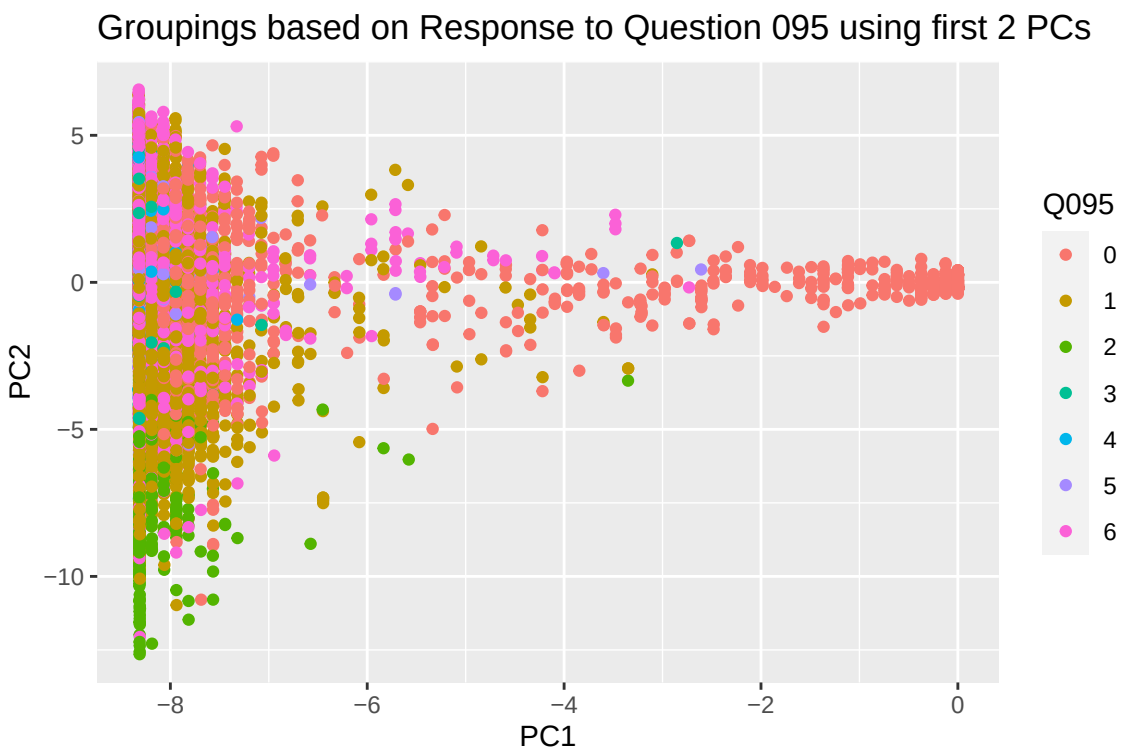

Figure 6: Groupings based on response to Question 95 using first 2 PCs

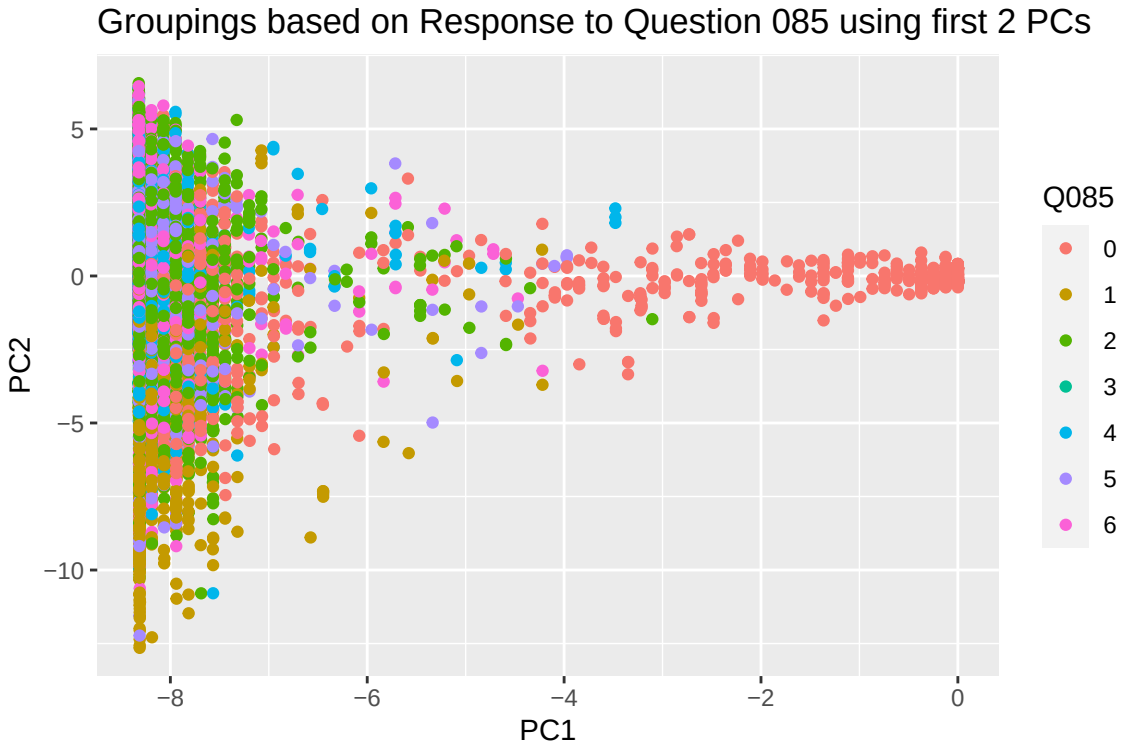

Figure 7: Groupings based on response to Question 85 using first 2 PCs

I also attempted to project the data long the third principal component. Figure 8 helps us visualise the different underlying structure in the data in a higher dimension.

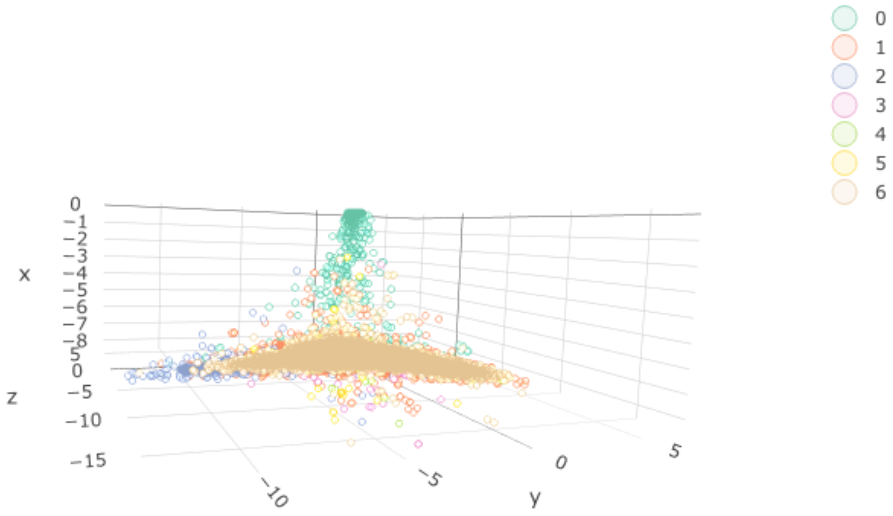

Figure 8: Projecting data along first three prinicpal components. Colour coded based on responses to Question 95

Unfortunately NMF did not prove to be an effective tool at dimensionality reduction. I experimented with two-dimensional projections for various values of k with none proving to be effective. Figure 9 is provided for reference.

4 Clustering

Having experimented with different dimensionality reduction algorithms, we now consider clustering algorithms. I decided to project the data along the first 100 PCs and then use k-means and NMF for clustering. I found that for k = 5, k-means tended to return clusters that were consistent with the geographical clusters. An example is provided in Figure 10. Similarly, setting k = 5 for NMF also resulted in suitable clustering. The results for NMF clustering are available in Figure 11.

As is expecteted of k-means, the clusters are clearly distinguishable from each other to the point where the clusters are far more clearly demarcated than reality. In comparison, NMF shows clusters that are fairly consistent with the structure observed in Figures 6 and 7, and is worth considering using for future tasks.

5 Stability

To test the stability of the results of my clustering algorithms, I bootstrapped the `ling_data`, and then performed the same steps behind PCA as mentioned above, after which I used NMF with k = 5 on the first 100 PCs as before. The clusters are visualised in Figure 12, and in my opinion are similar enough to suggest a reasonable stability and robustness to data perturbation.

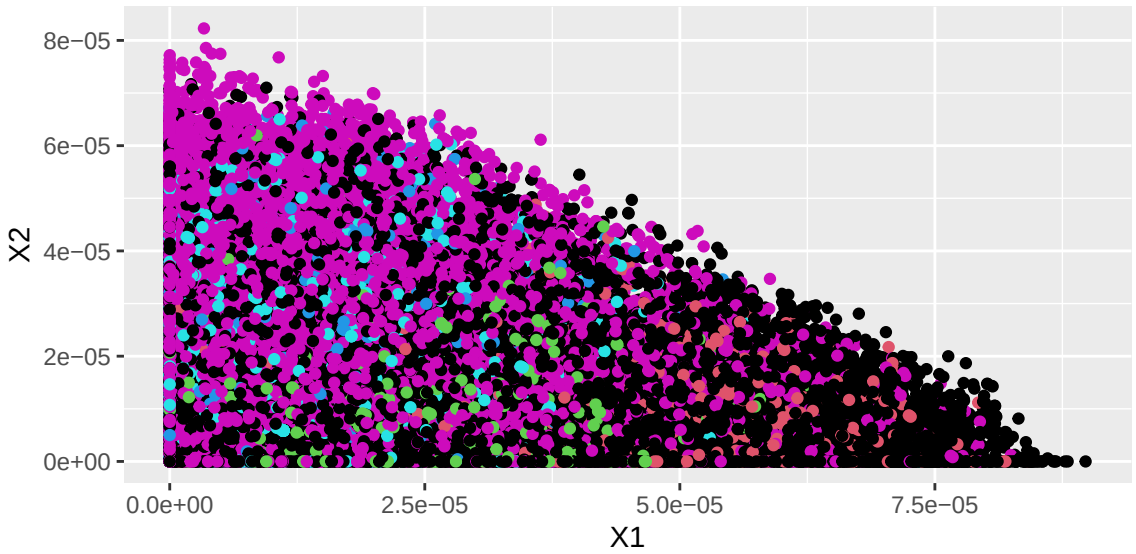

Figure 9: Attempted dimensionality reduction using NMF

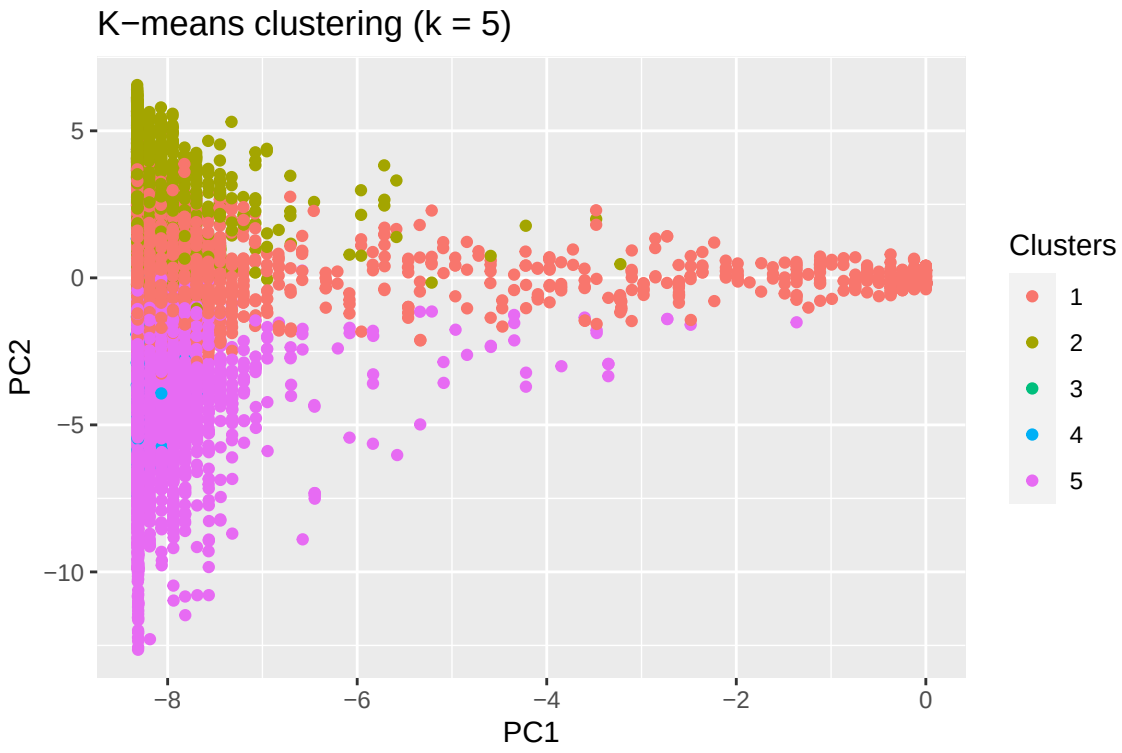

Figure 10: K-means clustering (k = 5) along first 100 PCs, plotted along first two PCs

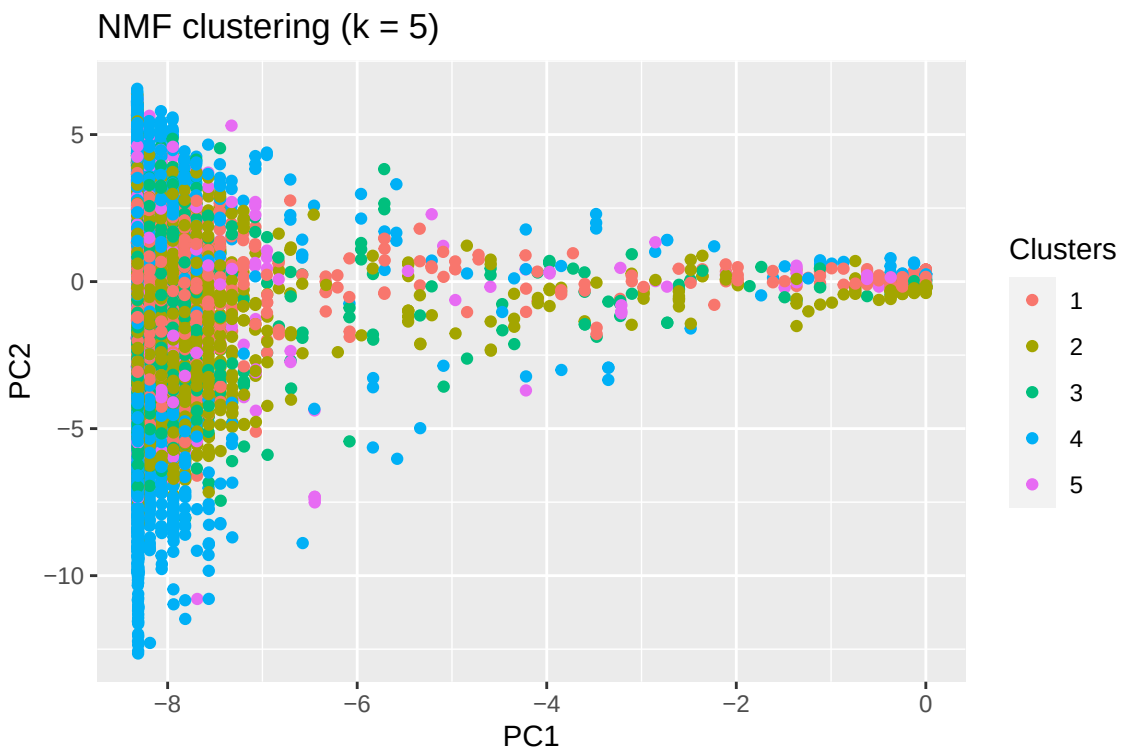

Figure 11: NMF clustering (k = 5) along first 100 PCs, plotted along first two PCs

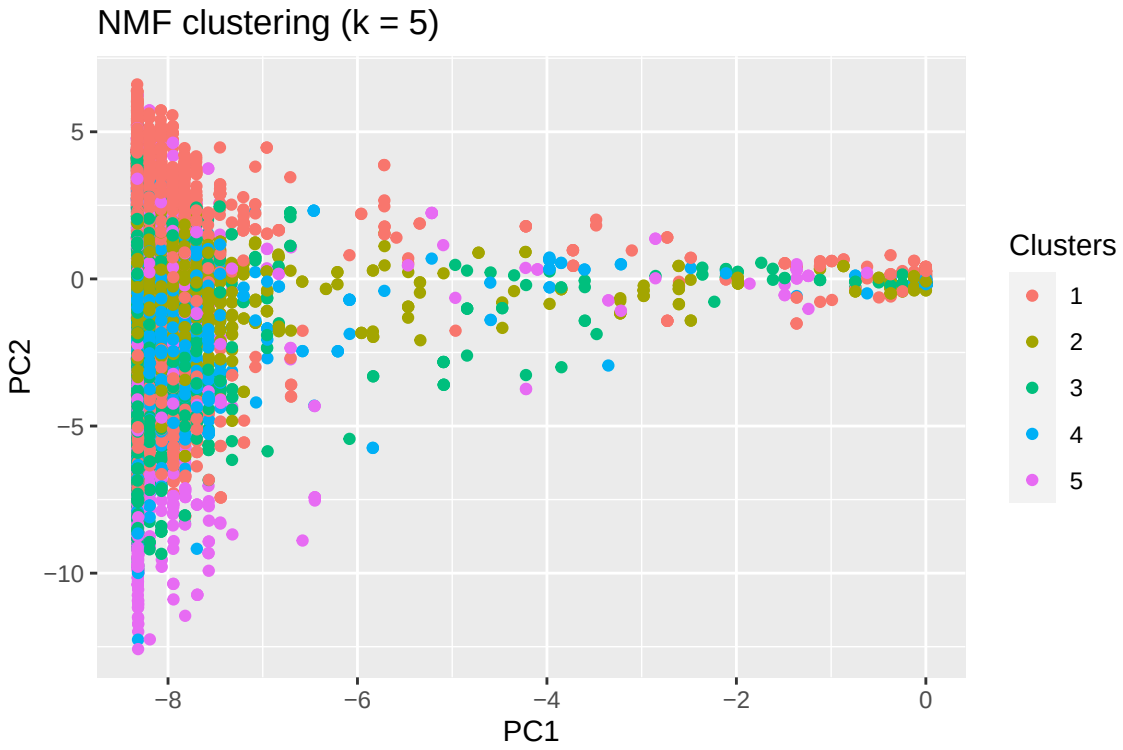

Figure 12: NMF clustering (k = 5) along first 100 PCs, plotted along first two PCs for bootstrapped data

6 Conclusion

We now contextualise our analysis through the lens of the three realms of data science: data, future data, and algorithms and analysis. In terms of using this data for future decision making, I believe it is reliable assuming that people were honest when filling the survey out, which is an assumption I am comfortable making. Notably while language inherently evolves over time, the analysis does not suggest an evolution so soon and so quick that we cannot rely on the data we currently have. An additional assumption being made is that there will not be massive immigration and emmigration trends such that the language changes drastically due to demographic changes. The clusters I found ie. the results of my algorithms and analysis have some utlitity in terms of aggregation and grouping, but could probably do with more advanced methodology before being used for decision making about dialect areas in the future. If I had more time, I would experiment more with the `ling_location` data, and also with Goebl's measure that weighted matches between infrequently used words more highly. Future analysis would also include examining data from Alaska and Hawaii. I also would have liked to be able to conduct a reality check by examining county level data and seeing how clustering at the county level would compare to the clusters I found using `ling_data` which is more granular data.

7 Academic Integrity Statement

I hereby certify that all work in this class, for this assignment and all others is my own. Whenever I consult my peers or collaborate with them, I will cite them. I will also cite all relevant resources used in my work. This is out of respect for myself, this class and its staff, the statistics department, UC Berkeley, and the academic enterprise as a whole.

8 Bibliography

1. Nerbonne, J. and Kretzschmar, W. (2003). Introducing computational methods in dialectometry. Computers and the Humanities, 37(3):245–55. Special issue on computational methods in dialectometry, Nerbonne, J. and Kretzschmar, W.I.,Jr. (eds).
2. John Nerbonne, William Kretzschmar, Jr, Progress in Dialectometry: Toward Explanation, Literary and Linguistic Computing, Volume 21, Issue 4, November 2006, Pages 387–397.
3. Se´guy, J. (1971). La relation entre la distance spatiale et la distance lexicale. Revue de Linguistique Romane, 35(138): 335–57.
4. Se´guy, J. (1973). La dialectometrie dans l'Atlas linguistique de Gascogne. Revue de Linguistique Romane, 37(145): 1–24.
5. Chambers, J. and P. Trudgill. (1998 [11980]). Dialectology. Cambridge: Cambridge University Press.
6. Goebl, H. (1982). Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Wien: O¨ sterreichische Akademie der Wissenschaften.
7. Goebl, H. (1984). Dialektometrische Studien: Anhand italoromanischer, ra¨toromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. Vol. 3 Tu¨bingen: Max Niemeyer.