Lab 2 - Linguistics Data, Stat 215A, Fall 2024

October 4, 2024

1 Introduction

The problem of interest in dialectometry, as highlighted by Nerbonne and Kretzschmar[6,7] in their papers, is to understand linguistic variation across geographical regions. Dialectometry is the specific field of linguistics concerned with measuring the differences between dialects, which are primarily determined by geography.

In the context of domain studies, dialectometry aims to quantify and visualize linguistic differences, such as variations in pronunciation, vocabulary, and grammar, across regions. This field is of great importance for a number of applications, including linguistic research, language documentation, and the comprehension of sociolinguistic patterns. The papers by Nerbonne and Kretzschmar illustrate the transition from a purely descriptive approach to dialectology to a more analytical and quantitative methodology made possible by computational techniques.

The aim of this study is to gain insight into the linguistic variation across the United States, with a particular focus on lexical differences. The goal of our analysis is to identify and characterize distinct geographic groups based on linguistic responses, to explore the relationship between linguistic choice and geographic location, and to potentially reveal patterns of language use that correspond to regional, social, or cultural boundaries. Furthermore, we will apply and evaluate various data analysis techniques, such as dimension reduction and clustering, to better understand the structure within the data and to uncover latent groupings within the respondents.

2 The Data

2.1 Data Overview

The dataset under examination was derived from a dialect survey conducted by Bert Vaux and originally available at http://dialect.redlog.net/index.html. This survey documents the lexical variations in English dialects across the United States. The data set comprises responses from 47,471 participants and encompasses the aforementioned questions, numbered 50 to 121, which concentrate on lexical differences as opposed to phonetic differences. The dataset, designated 'lingData.txt', comprises a series of variables, including 'ID', 'CITY', 'STATE', 'ZIP', and a set of variables spanning 'Q050' to 'Q121'. Additionally, it incorporates geographic coordinates, specifically latitude and longitude, for the central point of each respondent's ZIP code. Furthermore, a second dataset, 'lingLocation.txt', categorizes the data into one-degree latitude by one-degree longitude squares and aggregates the binary response vectors within these bins.

The data provides a comprehensive and detailed account of dialectological information, with each response representing a linguistic choice made by participants. The variables commencing with 'Q' correspond to the responses provided on the survey website. A value of 0 indicates no response, while other numbers correspond to the responses given on the website. For example, a value of 1 corresponds to the response 'a'. Additionally, the dataset incorporates geographical data, including city, state, and ZIP code, which is essential for mapping dialect variations onto a geographical landscape. The inclusion of latitude and longitude variables facilitates a more precise geographical referencing of respondents.

In conclusion, the data is highly pertinent to the issue under examination, offering a quantitative perspective on linguistic variation that is anchored in geographical reality. This permits a more profound comprehension of the interconnection between language, geography, and society, a relationship that is pivotal to the field of dialectometry.

2.2 Data Cleaning

The initial step will be to examine the linguistic data for any invalid or missing entries. The analysis yielded 1,020 missing values in the 'long' and 'lat' columns, 540 missing values in the 'CITY' column, and 3 in the 'STATE' column. Rather than deleting these entries, our intention is to employ the corresponding linguistic data to fill in the missing values at a later stage.

Upon calculating the total number of respondents included in the linguistics location dataset, we observed that it was identical to the count in the linguistics dataset. It is noteworthy that the linguistics location dataset provides comprehensive latitude and longitude data for individuals from the same or nearby regions. While the latitude and longitude coordinates in this dataset are recorded in integer form, the linguistics dataset provides these coordinates in decimal form. Accordingly, the latitude and longitude values in the linguistics dataset were rounded to the nearest integer. Following verification of the 'Number of people in cell' in the linguistics location data, which was aligned with the rounded latitude and longitude, it was confirmed that the rounding method was valid.

The next step is to impute the missing latitude and longitude values in the linguistics data set using the location information from the linguistics location dataset. It was found that the absent latitude and longitude coordinates in the linguistics data are marked with latitude $90°$ and longitude $-170°$ in the linguistics location data. This approach is justifiable for imputing missing location information, given that there are only three missing 'STATE' and 540 missing 'CITY' entries. Furthermore, no instances were identified where both the 'STATE' and 'CITY' fields were absent, suggesting that each record contains at least one of these fields. This allows us to estimate the longitude and latitude from the available regional data.

We also impute missing 'STATE' and 'CITY' in the linguistics data based on the provided latitude and longitude coordinates. In the event that an entry lacking information regarding the state or city has the same latitude and longitude as another entry containing the requisite data, it is possible to utilize the known state or city information from the latter entry to impute the former.

Furthermore, an examination of the number of questions and corresponding answers provided in the question data revealed a total of 485 answers. However, the linguistics data set only contained 468 answers. This discrepancy arises because the linguistics data omitted questions Q108, Q112, Q113, Q114, and Q116. According to the third task (after encoding, we will obtain p=468), we will not impute these missing questions with the information provided in the linguistics location data.

Ultimately, the linguistics location data was normalized to ensure consistency in the number of individuals represented in each cell. This step is essential for ensuring the comparability and analyzability of the data, as the data points represent varying numbers of individuals within each cell. Normalization is therefore necessary to facilitate a fair comparison and analysis.

2.3 Exploratory Data Analysis

In this section, We randomly choose two questions to look into:

- **Question 50**: What word(s) do you use to address a group of two or more people?
  This question provides insights into regional variations in addressing a group, which can be influenced by local culture or dialect.
- **Question 53**: Modals are words like "can", "could", "might", "ought to", and so on. Can you use more than one modal at a time? (e.g., "I might could do that" to mean "I might be able to do that"; or "I used to could do that" to mean "I used to be able to do that")
  This question explores the use of multiple modals, a linguistic feature that varies across different English dialects, providing a glimpse into geographical language habits.

In Figure 1(a), we have plotted the distribution of linguistic survey responses for Q050 and Q053 in North America. Due to the rounding of longitude and latitude, some points overlap, which may facilitate interpretation. The diverse coloration of the points signifies the disparate responses to Q050, whereas the magnitude of the points correlates with the prevalence of answers to Q053. With regard to question 50, respondents in the southeastern region exhibited a clear preference for the ninth answer, "y'all". This selection is indicative of a regional dialect feature. The use of "y'all" as a plural second-person pronoun is a well-documented

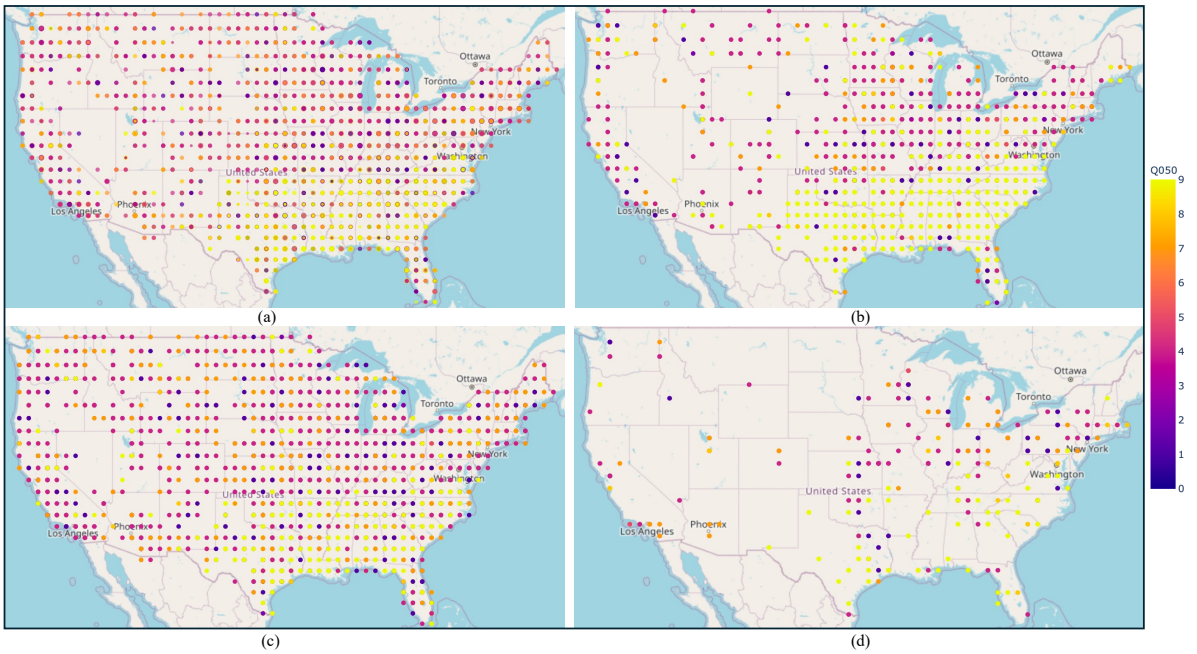

Figure 1: (a) North American Survey Responses for Q050 and Q053 (Excluding Hawaii and Alaska): Points are color-coded based on responses to Q050, with variations in size indicating the prevalence of answers to Q053. (b) Geographic Distribution of Q050 Responses (Q053 Answer = 1) in North America. (c) Geographic Distribution of Q050 Responses (Q053 Answer = 2) in North America. (d) Geographic Distribution of Q050 Responses (Q053 Answer = 3) in North America.

feature of Southern American English[5], shaped by the historical and cultural evolution of the region.

In Figures 1(b), 1(c), and 1(d), we present the geographic distribution of Q050 responses for respondents who answered 1, 2, and 3 to Q053, respectively. In contrast, respondents in the northeastern part of the United States demonstrated a proclivity for the first, fourth, and seventh responses, respectively. The observed variation is likely attributable to the influence of diverse dialects and linguistic traditions within the region. For example, the term "you guys"[8] is frequently employed in informal discourse across numerous regions of the United States, particularly in the northern and eastern areas. The selection of "you" in preference to a more specific plural pronoun may be indicative of the general tendency in Standard American English[9] to utilize singular pronouns in informal contexts.

In the subsequent analysis, we intend to investigate this relationship further by employing principal component analysis (PCA) to identify significant responses to specific questions and using clustering methods to segment the population based on their responses.

3 Dimension Reduction

In this section, we employed both the Sklearn PCA and the SVD PCA algorithms on our dataset. We selected SVD PCA because, in binary data, zero values serve as meaningful indicators of the absence of a feature. It is possible that centering the data may not be an appropriate approach, as it could potentially distort the meaningful zeros. Notwithstanding this preference, we conducted a comparative analysis of the two methods, examining their respective runtimes and the variance they explained, while utilizing an identical number of principal components.

First, the categorical data was transformed into binary vectors. Given that all variables have been encoded in binary form, rescaling is an unnecessary procedure. Rescaling is a more pertinent procedure for variables that exhibit a normal distribution or are not binary, as scaling adjusts the range of values rather than their presence or absence.

Figure 2 depicts the first two principal components plotted against latitude and longitude. It is unfortunate that no clear clustering is evident in the PCA plot, which may be attributed to a number of factors[4]. First, the size of the dataset is a significant factor. The dataset is of a considerable size, and the initial two principal components may only account for a minor proportion of the overall variance. The inclusion of additional principal components may facilitate the identification of more insightful structures. Secondly, the 'curse of dimensionality' may also be a contributing factor. In high-dimensional datasets, the proximity of data points within the same cluster may be obscured, making it challenging to discern distinct groupings. Thrid, the data exhibits a lack of linearity. PCA is particularly adept at uncovering linear relationships. In the event that the relationships between data points are either weak or non-linear, it is possible that PCA will be unable to effectively capture these patterns.

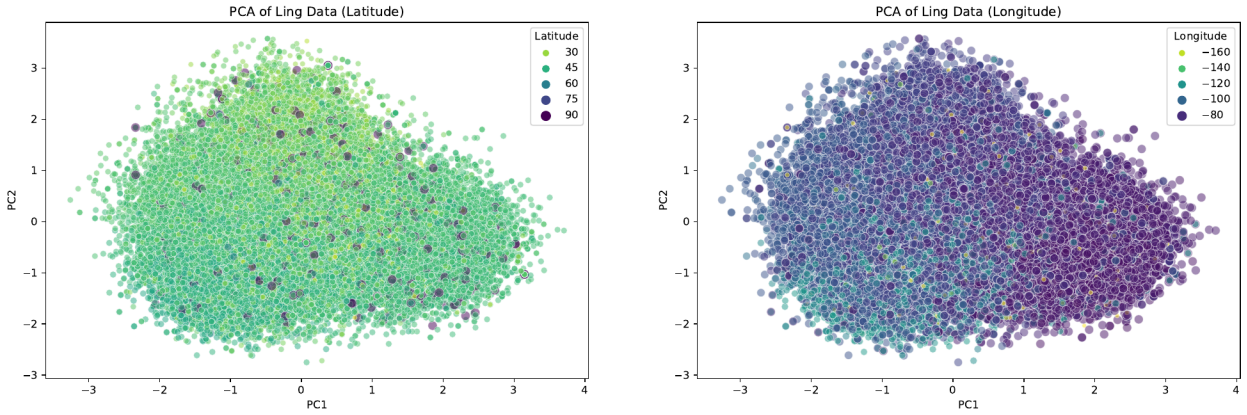

Figure 2: Sklearn (Mean-centered) PCA of Linguistics data. Left: PCA with respect to latitude. Right: PCA with respect to longitude.

In Figure 3, clustering remains unclear. However, one cluster appears more concentrated, and there are intriguing potential outliers. It seems probable that these outliers are data points from Alaska and Hawaii, given the considerable geographical distances from the mainland and the potential for dialect differences to result in distinct responses.

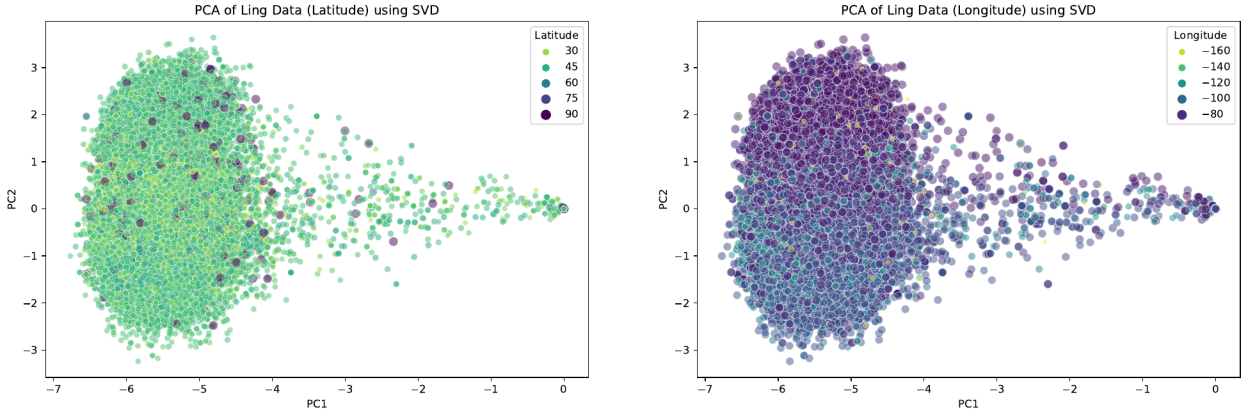

Figure 3: SVD PCA (Not Mean-centered) of Linguistics data. Left: SVD PCA with respect to latitude. Right: SVD PCA with respect to longitude.

Table 1 provides a summary of the comparison. The results demonstrate that SVD PCA necessitates only 47 principal components to account for 75% of the variance, in contrast to the 90 components required by Sklearn PCA. Moreover, when the variance explained with five principal components was considered, SVD PCA achieved a significantly higher variance of 52%, in comparison to Sklearn PCA's 15%. With regard to

Table 1: Comparison of Sklearn (Mean-centered) PCA and SVD (Not Mean-centered) PCA

5 PCs Given	Sklearn PCA	SVD PCA
Running Time (Seconds)	0.7552	**0.5093**
Variance Explained	15%	**52%**
75% Variance Explained	Sklearn PCA	SVD PCA
Number of PCs	90	**47**

Table 2: Loadings of the First Two Principal Components

PC1 Answer	PC1 Loading	PC2 Answer	PC2 Loading
Q063_1	-0.17	**Q073_1**	0.27
Q067_1	-0.17	**Q073_6**	-0.24
Q093_2	-0.17	Q105_1	0.20
Q054_2	-0.16	**Q080_1**	0.19
Q081_1	-0.16	**Q080_8**	-0.18
Q055_2	-0.16	Q105_2	-0.17
Q053_2	-0.16	Q056_2	0.17
Q115_1	-0.16	Q106_1	-0.16
Q109_1	-0.15	Q056_1	-0.16
Q104_1	-0.15	Q103_3	-0.15

runtime, SVD PCA exhibited superior performance, requiring only 0.5093 seconds, in comparison to Sklearn PCA, which required 0.7552 seconds. These findings validate our decision to utilize SVD PCA for subsequent analysis, as it demonstrates superior efficiency and performance with binary data.

Moreover, we investigated the loadings of the first two principal components in Table 2 and categorized the questions into four thematic classes: 1- **Language Usage**: Q50-Q53 (use of specific phrases and linguistic constructs), Q54-Q57 (use of "anymore"), Q58-Q57, Q97, Q98 (terminology for activities, objects, or creatures); these questions explore regional language use and may reveal dialects or colloquialisms. 2- **Cultural and Regional Expressions**: Q68-Q71 (nicknames for grandparents), Q72-Q81 (terms for everyday objects or food), Q82-Q90 (expressions with regional variations); these questions reflect cultural practices and regional language differences. 3- **Social and Behavioral Norms**: Q91-Q96, Q99-Q101, Q120, Q121 (terms related to social interactions), Q102-Q111, Q118, Q119 (terminology related to social practices); these questions could indicate social customs. 4- **Phonological and Pronunciation Variations**: Q112-Q117 (pronunciation of specific words), indicative of dialectal pronunciation differences.

As illustrated in Table 2, the initial principal component predominantly captures variations in language usage, as evidenced by its negative loadings. This indicates that as the score of the first principal component increases, the influence of the Language Usage questions decreases, suggesting a potential contrast or inverse relationship within this dimension. In contrast, the second principal component is associated with cultural and regional expressions. While these principal components appear to encapsulate dialectal information, their comprehensive coverage of diverse aspects of the data renders them unsuitable for the creation of summary variables that directly indicate geographical distinctions. Nevertheless, transforming the dataset into the PC space is essential for analyzing the distribution of responses across these principal components, which will be discussed in further detail in Section 4.

4 Clustering

In this section, we examined a variety of clustering techniques to identify patterns and groupings within our dataset. However, due to computational limitations—specifically, the processing capabilities of our hardware (2 GHz Quad-Core Intel Core i5 processor and 16 GB of 3733 MHz LPDDR4X RAM)—we were unable to utilize hierarchical models and several other sophisticated algorithms due to their excessive resource requirements and time constraints.

Consequently, two clustering methods were successfully applied and reported on in consideration of the aforementioned computational constraints and the characteristics of our dataset. The two clustering methods that we successfully applied and reported on were K-means and Gaussian mixture models (GMM). Both methods are well-suited for the handling of large datasets and can provide valuable insights into the structure of the data without requiring excessive computational resources.

4.1 K-Means

Prior to executing the clustering process, we employed the dataset that had been dimension-reduced and transformed into principal component (PC) space as detailed in Section 3. This transformation was essential for managing the high dimensionality of our data and for enhancing the clustering performance by focusing on the most informative dimensions.

To determine the optimal number of clusters, we leveraged the silhouette score, a metric that measures the quality of clustering by assessing how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, with higher values indicating better-defined clusters. We performed a 5-fold cross-validation to objectively evaluate the silhouette score for various values of k, the number of clusters.

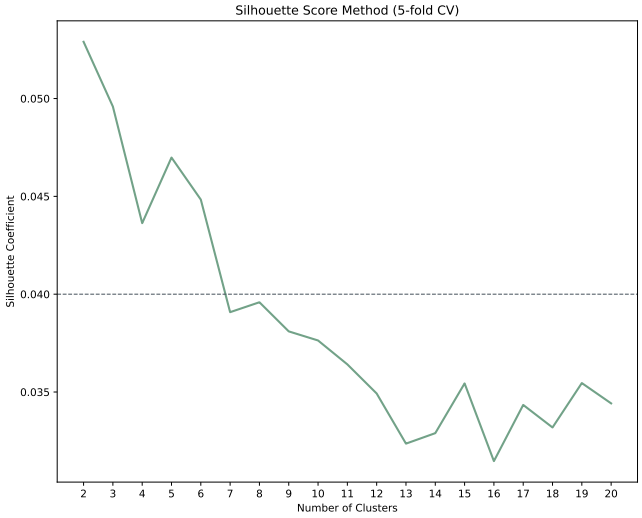

Figure 4: 5-Fold Cross-Validation of Silhouette Scores for Determining the Optimal k

Upon examining the silhouette scores from the 5-fold cross-validation (Figure 4), it was observed that a smaller number of clusters initially yielded higher silhouette scores, suggesting better intra-cluster similarity and inter-cluster dissimilarity. However, we also considered the geographical and linguistic diversity across the United States, recognizing that a smaller number of clusters might oversimplify the complex linguistic landscape. It was observed that while a value of k equal to 3 resulted in relatively high silhouette scores, this value may not be sufficiently robust in capturing the nuanced linguistic variations across different regions. Conversely, a k value of 6 provided a more granular partitioning of the data, potentially allowing for the identification of more localized linguistic patterns without overly fragmenting the data, so we analyzed six clusterings as well.

Examination of Figure 5 reveals that data points assigned to the same cluster tend to be geographically concentrated. Specifically, the data points categorized under cluster 2 are predominantly found in the southeastern United States. Cluster 1 encompasses a significant portion of the northern U.S. and the regions bordering Canada. Meanwhile, cluster 3 appears to be associated with coastal areas, with notable concentrations of data points along the U.S. coastlines. The geographic clustering of dialects is particularly evident in states like California and New York, where the prevalence of distinct regional dialects significantly

influences the clustering patterns. These observations suggest that the linguistic variations captured by our dataset are closely tied to geographical regions, aligning with the known distribution of American dialects.

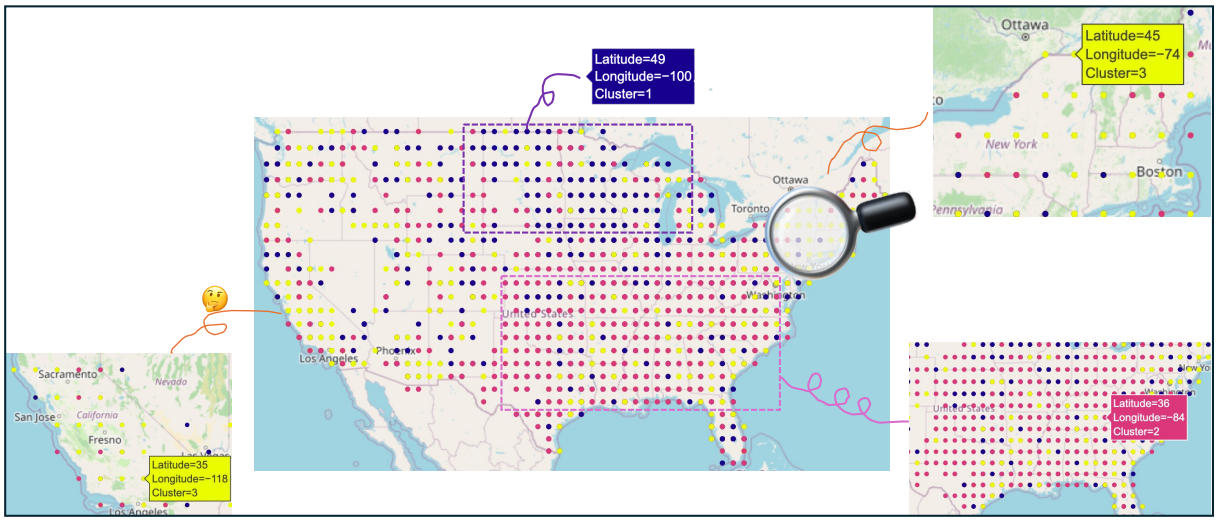

Figure 5: K-means Clustering with 3 Clusters

Figure 6 presents the results of our analysis with k-means clustering set at 6 clusters. Consistent with our previous findings for k=3, a distinct and strong clustering (cluster 6) is observable in the southeastern United States, particularly concentrated in Mississippi, Texas, Kentucky, and North Carolina. This geographical concentration suggests the presence of shared linguistic features or historical connections that have influenced the dialects in these regions. Additionally, we identified another prominent cluster (cluster 5) in and around New York and its neighboring areas. The distinct nature of this cluster could be attributed to the impact of early settlements, cultural hubs, or economic centers that have played a significant role in shaping the dialects of this region. Clusters 1 and 2 are noted to be sporadically distributed along the coastal regions and at state borders. This pattern may be a reflection of historical migration patterns, trade routes, or cultural exchanges that have contributed to the diversity of dialects in these areas.

However, it is evident that other clusters are intermingled across various states. This blending of clusters could be the result of several factors: 1- **Historical Migration**[2]: The movement of people from one region to another may carry their dialects with them, leading to a mixing of dialect features. 2- **Cultural Exchange**[3]: The exchange of ideas, media, and communication across state borders could foster a blending of dialects. 3- **Linguistic Variation**[1]: The natural evolution and variation within a language can result in diverse dialects even within the same region. These factors collectively contribute to the complex linguistic landscape of the United States, as revealed by our clustering analysis.

4.2 Gaussian Mixture Model

Consistent with the approach we took for K-means clustering, we utilized the dataset that had been transformed into principal component (PC) space as described in Section 3. We employed the silhouette score to determine the optimal number of clusters for our analysis. Upon reviewing the silhouette scores from Figure 8, we observed that all evaluated configurations resulted in scores below 0, indicating that the clustering was suboptimal. The K-means algorithm is best suited for datasets with distinct cluster separation and a roughly spherical shape. If a dataset aligns with these characteristics, K-means tends to perform well. Conversely, the Gaussian Mixture Model (GMM) assumes that the data is composed of a mixture of multiple Gaussian distributions. If the actual data distribution significantly deviates from the Gaussian distribution, GMM may not perform effectively. This assumption discrepancy is likely the reason why the silhouette scores for K-means were considerably higher, albeit still below the desired threshold, compared to GMM.

Despite these challenges, we selected the option that maximized the silhouette score, which was k=6. In Figure 7, we present the first two principal components labeled with their respective clusters in a 2D space,

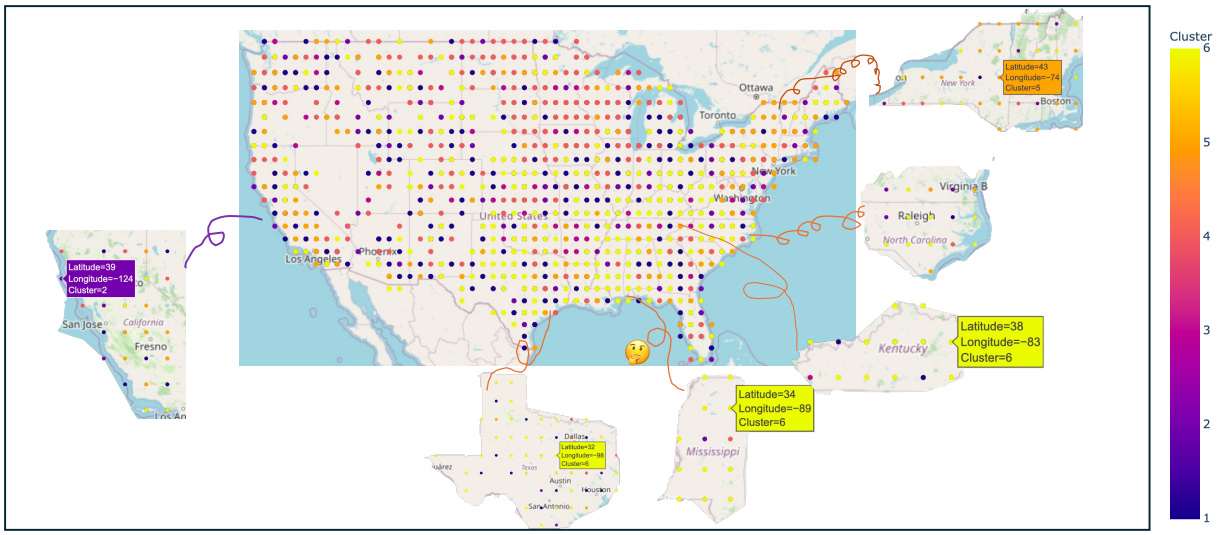

Figure 6: K-means Clustering with 6 Clusters

and we extend this visualization to a 3D space using the first three principal components. The visualizations reveal a dominant cluster, with other clusters intermingling. Upon closer inspection, particularly in Figure 8, it becomes evident that the GMM clustering does not delineate clear regional groupings as distinctly as K-Means does.

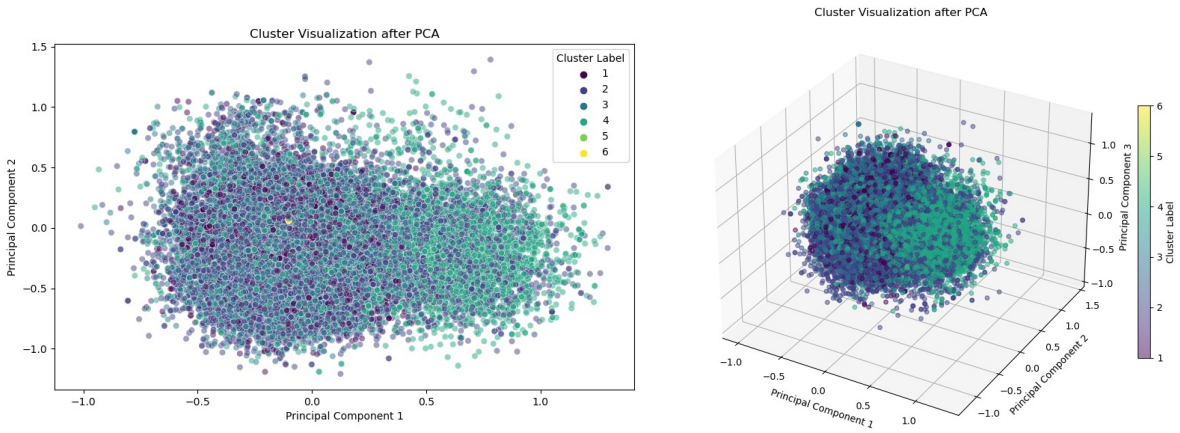

Figure 7: PCA of Linguistics data under GMM clustering. Left: 2D Space. Right: 3D Space

5 Stability of findings to perturbation

To assess the robustness of our clustering results, we introduced perturbations to the dataset and re-evaluated the clustering outcomes. The perturbation process was designed to simulate real-world variations and potential noise in data collection. This section details the perturbation method, its impact on clustering, and the overall stability of our findings.

In our perturbation approach, we initially create a duplicate of the original dataset to ensure that we do not alter it directly. This copy is then subjected to two key perturbation steps: random noise addition and bootstrap sampling. For the random noise addition, we calculate the mean of the dataset and introduce noise scaled to 20% of its maximum value, ensuring the noise level is relative to the data's scale and avoiding

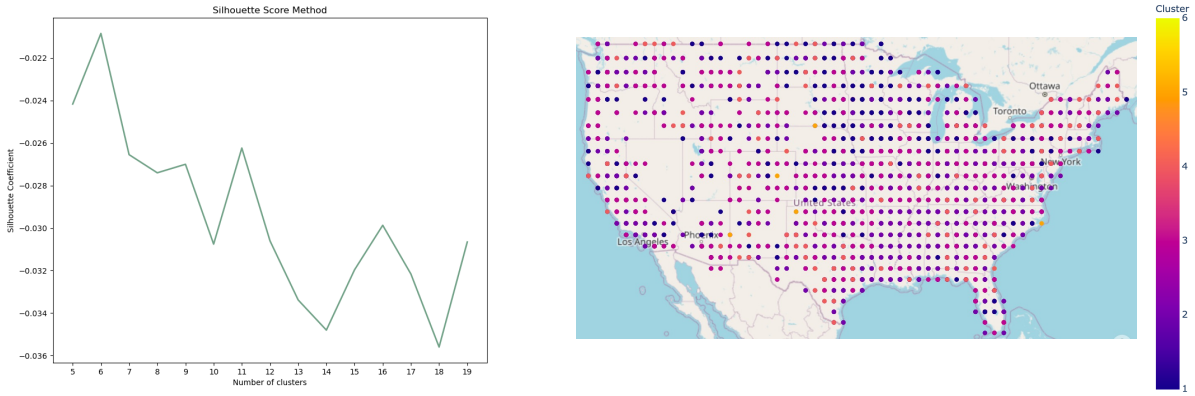

Figure 8: Silhouette Scores and Corresponding Clustering Maps. Left: Determining the Optimal k Using Silhouette Scores. Right: Clustering Map.

skewed impacts on smaller value. This noise is generated from a normal distribution with a mean of 0 and a standard deviation based on the calculated scale value. Then we proceed to perform a bootstrap sample of the perturbed dataset, enabling random resampling with replacement.

After perturbing the data, we applied SVD K-means clustering with six predefined clusters. Despite the introduction of noise and resampling, the clustering results demonstrated remarkable stability in Figure 9:

- Southeastern United States: Although some minor shifts were observed in the clustering of the southeastern part of the United States, the overall structure remained consistent.
- New York State: Notably, the clustering for New York State showed no significant change, indicating that the perturbation did not influence its cluster affiliation.

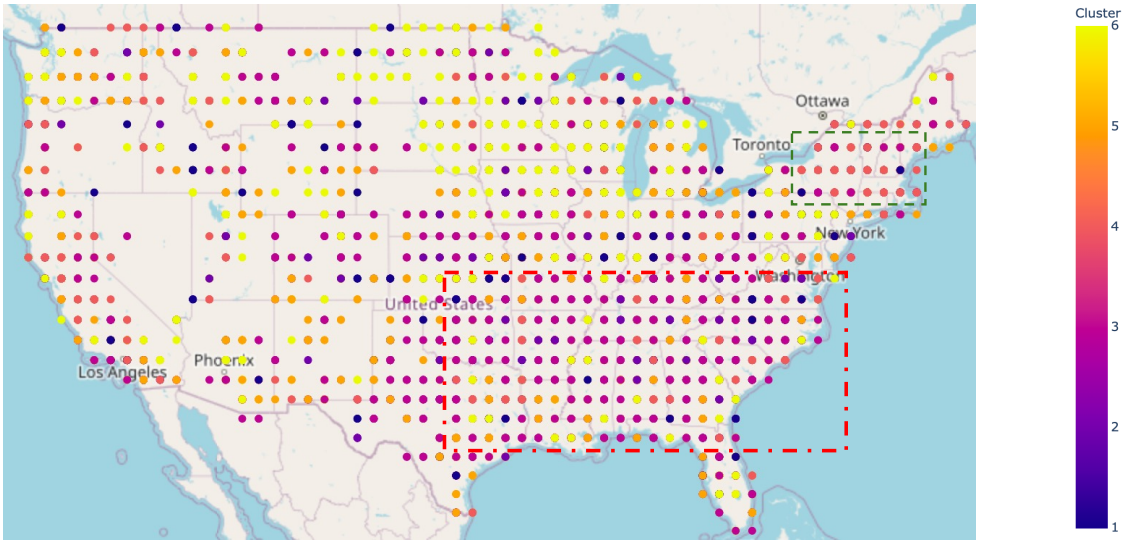

Figure 9: Post-Perturbation K-means Clustering with 6 Clusters

6 Conclusion

In conclusion, the three realms of data science have played crucial roles. The linguistic survey data provided a rich tapestry of dialectological information. The inclusion of geographical coordinates allowed us to analyze linguistic variations in the context of locations. The binary encoding of categorical responses facilitated the

application of mathematical algorithms to this qualitative data. We employed SVD PCA and K-means clustering to uncover patterns and groupings within the dataset. The choice of these algorithms was driven by their effectiveness in handling large datasets and their ability to reveal underlying structures. The perturbation analysis further strengthened our confidence in the robustness of these algorithms. Our analysis revealed distinct clustering patterns that correspond to geographical regions. Notably, the southeastern United States and New York State exhibited stable clustering characteristics, suggesting that these groupings are robust and reflect genuine linguistic differences. The stability of these clusters after perturbation underscores the reliability of our findings.

7 Academic Honesty

7.1 Statement

I hereby certify that the work presented in this document is my original work, completed in fulfillment of the requirements for the course STAT 215A. Unless otherwise indicated by proper citation, the findings, analyses, and conclusions presented herein are the result of my own efforts.

In the course of this work, I have maintained the highest standards of academic honesty and integrity. All data utilized in this study were subjected to ethical analysis, and no unapproved assistance or resources were employed in the completion of this assignment. I am aware of the significance of academic integrity and the severe repercussions associated with plagiarism, fabrication of data and results, or falsification.

7.2 LLM Usage

I used DeepL for the grammar revision.

7.3 Collaborators

I am grateful to Anthony for his assistance in resolving my coding issues and to Yan for his valuable suggestions.

8 Bibliography

References

[1] James Burridge. "Spatial evolution of human dialects". In: *Physical Review X* 7.3 (2017), p. 031008.

[2] Robin Dodsworth. "Migration and dialect contact". In: *Annual Review of Linguistics* 3.1 (2017), pp. 331–346.

[3] Oliver Falck et al. "Dialects, cultural identity, and economic exchange". In: *Journal of urban economics* 72.2-3 (2012), pp. 225–239.

[4] Ian T Jolliffe and Jorge Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202.

[5] Michael B Montgomery, Ellen Johnson, and Charles Reagan Wilson. *The New Encyclopedia of Southern Culture: Volume 5: Language*. Vol. 5. UNC Press Books, 2014.

[6] John Nerbonne and William Kretzschmar. "Introducing computational techniques in dialectometry". In: *Computers and the Humanities* 37 (2003), pp. 245–255.

[7] John Nerbonne and William Kretzschmar. "Progress in dialectometry: toward explanation". In: *Literary and Linguistic Computing* 21.4 (2006), pp. 387–397.

[8] Ben Sienicki. *" Hey, y'guys!": A diachronic usage-based approach to changes in American English address*. The University of New Mexico, 2014.

[9] Kenneth G Wilson. *The Columbia Guide to Standard American English*. Columbia University Press, 1996.