Lab 2 - Linguistic Survey
Stat 215A, Fall 2013

Your name

1 Introduction

In the study of dialects, there is a long-standing connundrum about the relation of regional dialects and spatial distribution of particular linguistic choices. While it is evidently possible to qualitatively identify a person's regional dialect in practice, the analysis of any particular linguistic choice does not generate these boundaries. Here, we attempt to take the "blind" approach: to use extensive data across multiple linguistic features to generate those boundaries that are spatially pertinent, without explicit reference to the boundaries themselves.

2 The Data

This data is linguistic survey data collected from the internet pertaining to a person's common english language word choice.

2.1 Data quality and cleaning

For simplicity, I deleted all rows with instances of NA from my dataset. This removed about 2% of the data, which I deemed acceptable. Many question responses were missing, labelled with a zero. These responses were left in the data set and handled further downstream.

To create a spatial grid of data that did not depend on the local density of responses, I used a nearest-neighbor search. For each point on the grid, the k=100 nearest neighbors were found. Then, these 100 neighbors' "binarized" responses were added together. At this point, zeros were omitted from responses; thus, each question was not necessarily normalized to 100 responses.

It was assumed that zero-responses were from either respondent uncertainty or simple malfunction. Assuming the malfunctions to be randomly distributed and the respondent uncertainty to be meaningful, the resulting bins of fraction of non-zero responses of the 100 nearest neighbors were assumed to be meaningful.

2.2 Exploratory Data Analysis

To explore the data, first I took all of the responses from the data, and I generated a distance function between two elements. This distance function is that used by Seguy: the distance is the number of elements which are not shared in common. This allows us to immediately begin clustering data with a non-metric method. Thus, I began by using a heirarchical clustering.

It was clear from this initial clustering that this did not create a geographical distribution of responses. Furthermore, the maximum variation generated by individiuals was larger than that generated by spatial variation. This indicated that spatial aggregation would be necessary to reduce variance from individuals within a region, to be able to see the broader trend.

I also wanted to explore the number of redundant area codes that showed up in the raw data.

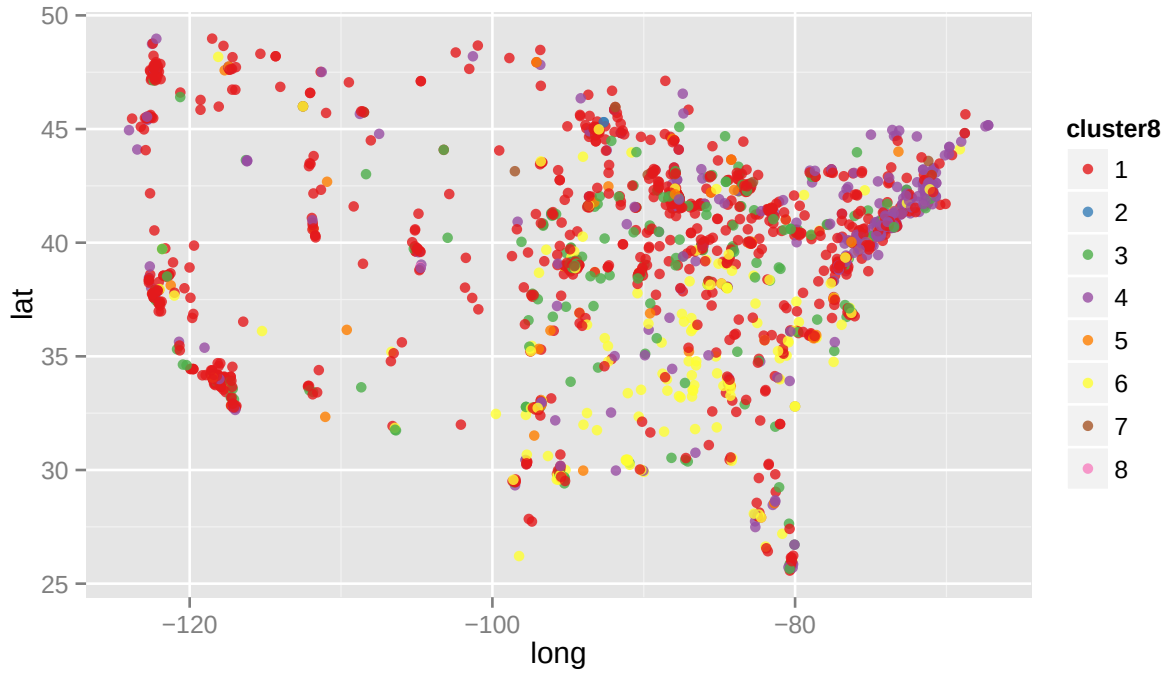

Figure 2.1: Heirarchical clustering on individuals shows excessive sensitivity to particular outliers. Data may give a sense of the natural variation of responses.

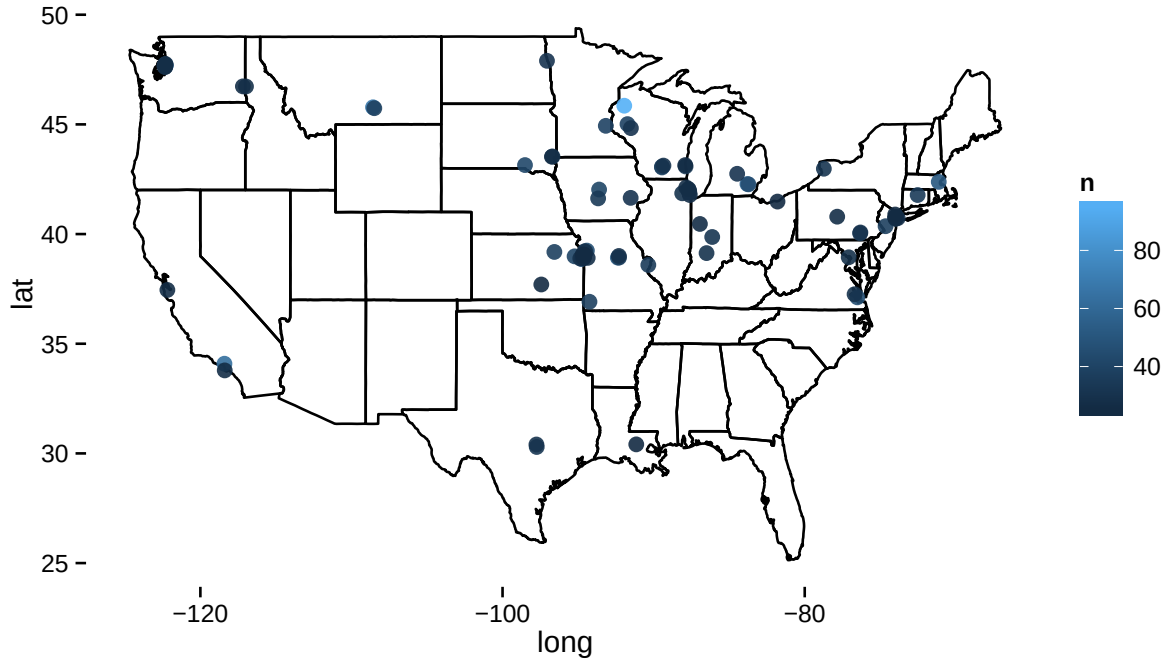

Figure 2.2: Redunancy in area code shows a strong signal near Minneapolis. This feature emerges in the heirarchical clustering of spatially coarse grained data.

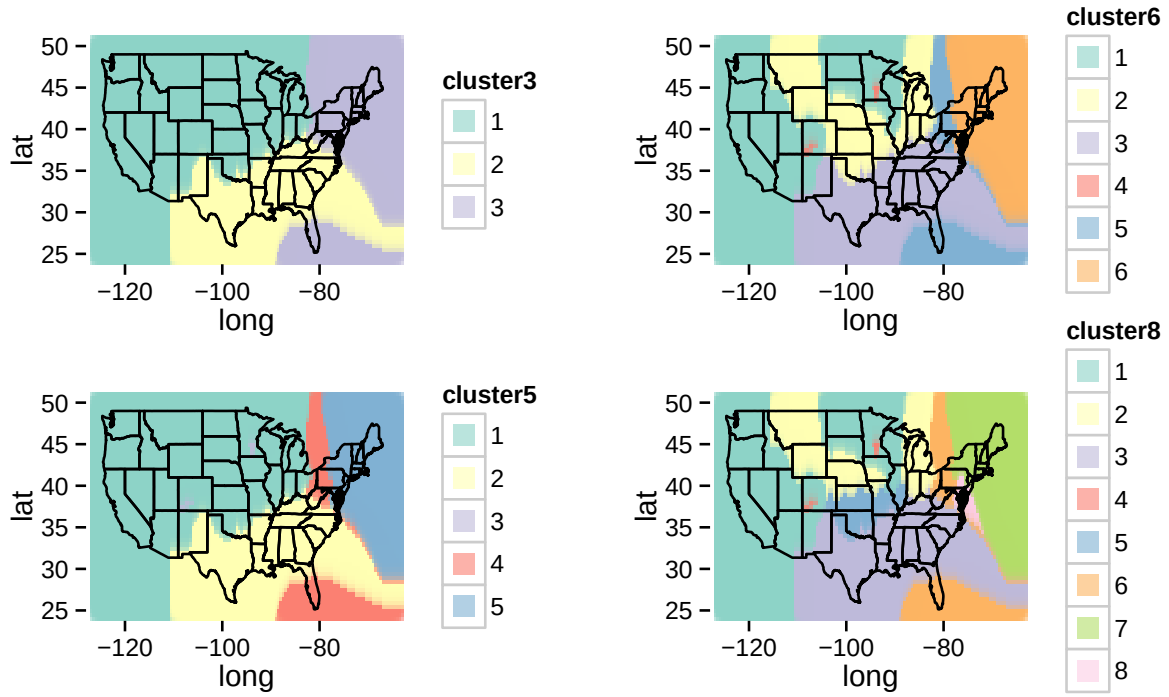

Figure 3.1: Clusters formed from spatial data shows much more regularity and the iterative emergece of cultural regions is apparent – save for the outliers in Minneapolis and Denver.

3 Dimension reduction methods

3.1 Hierarchical Clustering of spatialized data

To produce more stable clusters, the sptailization method described in the data cleaning was used. After performing this clustering, a standard minkowski distance was used for the clustering. This is a reasonable metric to use because the data now spans the range of 0-100, and a minkowski distance will create more distance between quite similar pairs and less distance between quite dissimilar pairs. This will reduce the effect of extreme outliers.

When we analyze the clusters generated, we find that the heirarchical clustering captures layers of relatedness that first separates Northeast, South, and West, then it goes on to separate Florida from the south, Pennsylvania from the Northeast, and to create a band between Wisconsin/Minnesota/Wisconsin and the South or East. To my cultural understanding, all of these distinctions are relevant and interesting.

One of the problems with this clustering, though, is the association of spatially distinct clusters with one another. Further cutting of the tree, as shown in the Cluster-8 diagram, does not serve to resolve these associations; it instead creates groups near the borders of previous regions that likely have intermediate linguistic properties.

3.2 PCA binning

After analyzing this data, it is interesting to ask whether there are questions with maximal and minimal variance that vary from region to region. The natural analysis for this is a principal component analysis – to see whether there are principal components within subsets of our macroscopic dataset.

In this analysis, it seems that the lowest eigenvectors live in some sort of textured subspace of the data, while the largest eigenvectors are pretty well distributed among all of the responses. There are a few questions that stick out above the signal for the largest eigenvectors, but their loadings are pretty far from 1.

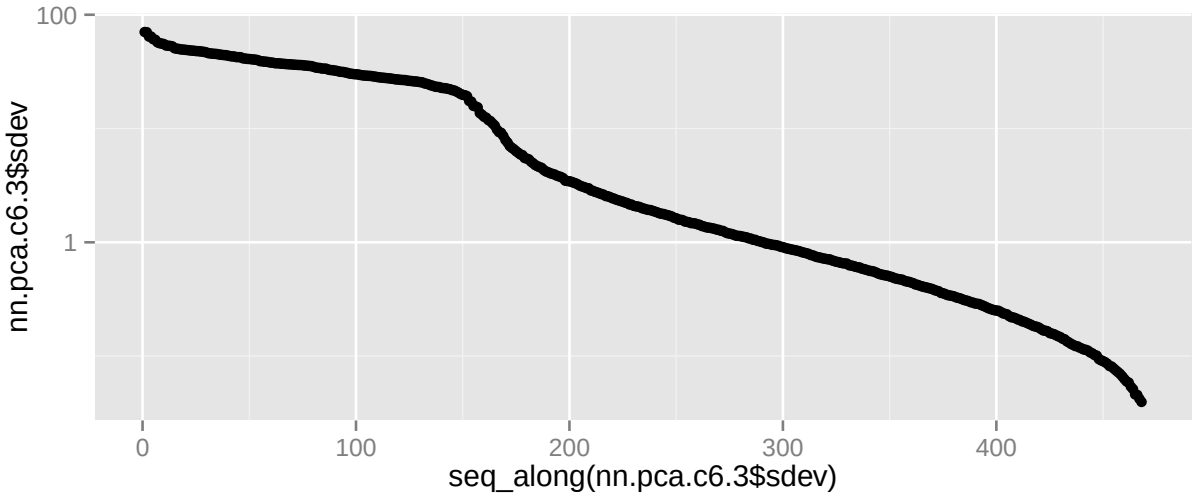

Figure 3.2: PCA analysis for the south, i.e. cluster 3 of 6 from the cluster6 analysis

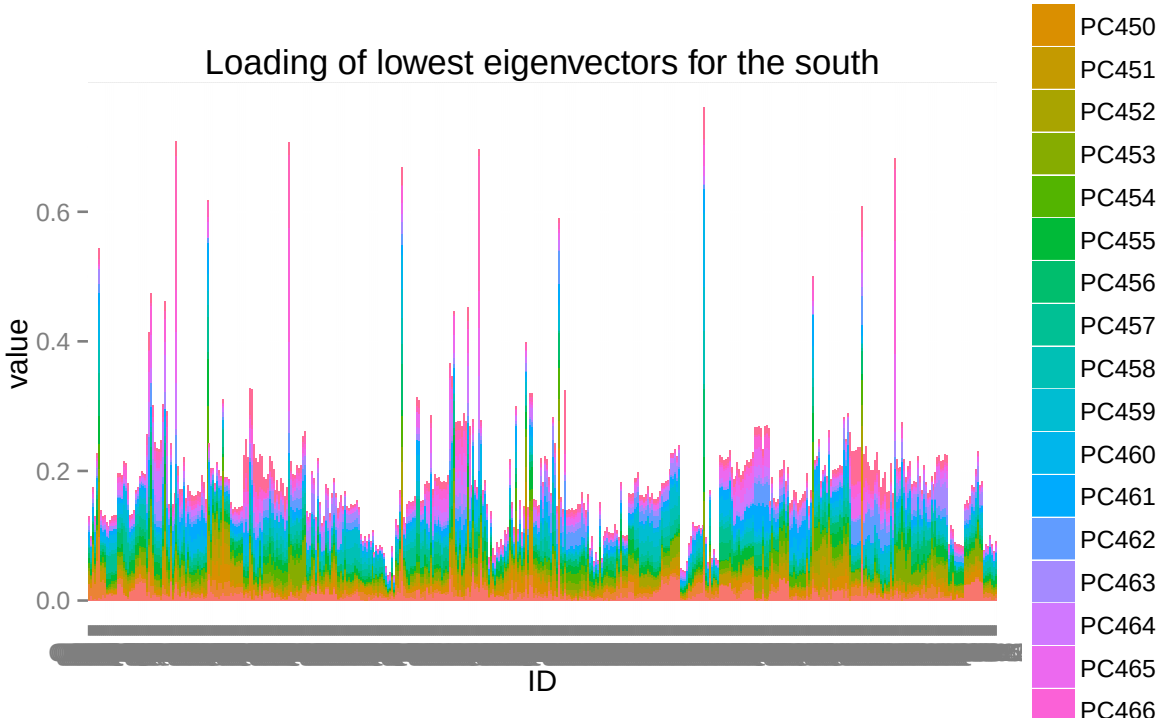

Figure 3.3: PCA analysis for the south, i.e. cluster 3 of 6 from the cluster6 analysis

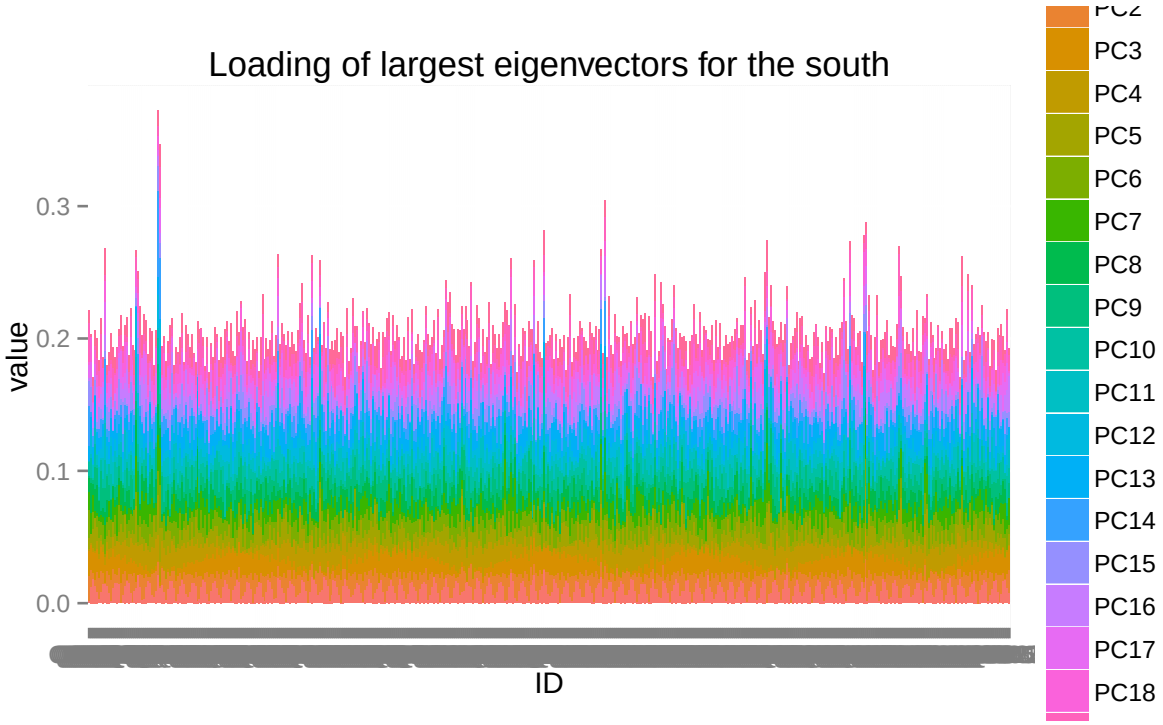

Figure 3.4: PCA analysis for the south, i.e. cluster 3 of 6 from the cluster6 analysis

4 Stability of findings to perturbation

There was not sufficient time for thorough stability analysis, but the results were stable to a coarsening of the grid by an area factor of 4, as demonstrated below. Many of the fine features of the borders disappear, but the important regional trends remain.

5 Conclusion

From this blind analysis, we were able to extract a culturally relevant partition of the united states. This analysis did not leverage much domain knowledge of linguistics; a deeper analysis that included correlations and clustering of questions themselves into categories would be able to more robustly analyze regions without resulting in a "crumbly" edge as the number of clusters increased.

Furthermore, a predictive aspect to the analysis seems possible from this starting point. Given the large individual variability discovered from the first clustering, there would be much to learn from a predictive methodology. Ways of accounting for the variability that is spread across all of the answers from within region to region will be crucial for prediction.

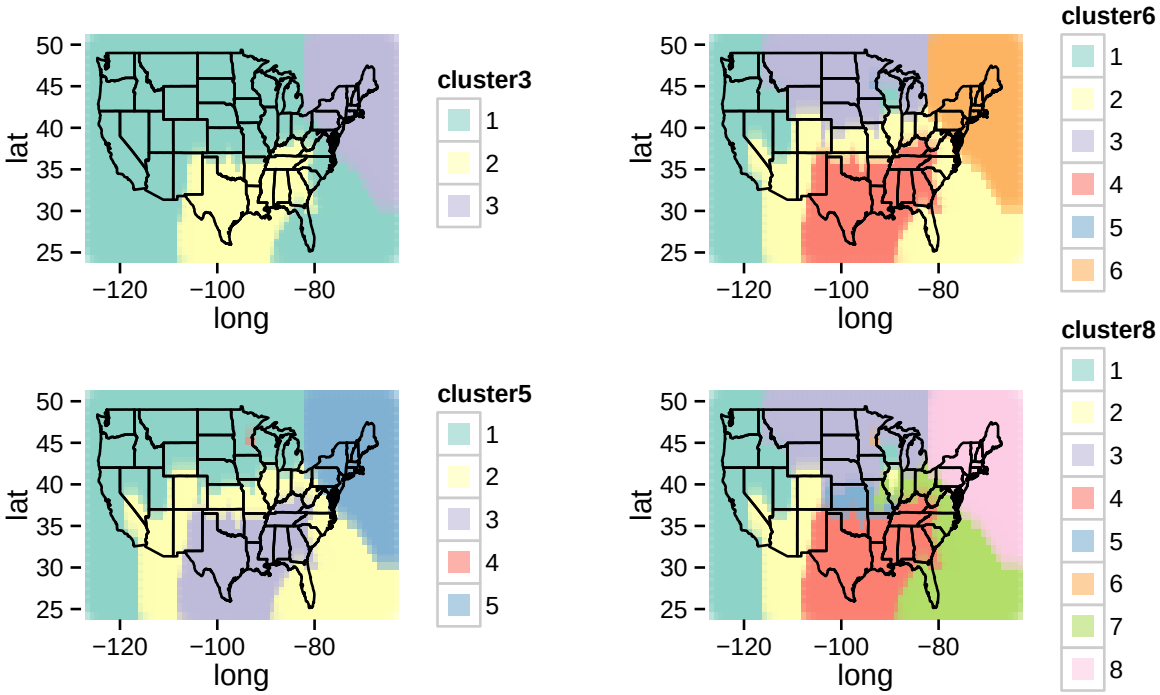

Figure 4.1: A coarser validation of the clustering analysis