Lab 2 - Linguistic Survey
Stat 215A

Jonathan Fischer

October 7, 2014

1 Introduction

In this report, we revisit the redwood data of Lab 1 to examine the effects of adding various smoothers to the temperature density and plots of temperature and humidity for a fixed time of day. For this exercise, we wish to observe the differences in performance among different kernel and parameter choices.

The second element of the lab utilizes the lingual data collected by Bert Vaux in his 2003 Dialect Survey. In our treatment, we clean the data before examining a few questions geographically. To allow for better analysis, we convert the data from categorical responses to a sequence of 0's and 1's indicating the choices selected by respondents. This permits the use of PCA and subsequent K-means clustering based on projections to the principal axes. Optimal clustering seems to occur for K=3, and we observe groupings based in the northeast, south, and midwest/west. Perturbation by subsampling and different initial conditions for K-means produced identical clusters, engendering confidence in our results.

2 Redwood Data

2.1 Temperature Density Estimates via Kernel Smoothing

Our first task is to generate kernel density estimates of temperature based on the data obtained on the macroscope network. We used the cleaned data set from Lab 1 and aggregated over time and node id to get the entire temperature profile. Four kernels were considered for estimation–biweight, Epanechnikov, Gaussian, and rectangular. We examined the density estimates for five choices of the bandwidth, as shown in Figure 2.1. Smaller bandwidths produced excessively spiky densities, though an adjustment factor of 1 or greater mitigated this concern. The high bandwidths resulted in oversmoothing, more or less yielding gaussian distributions. This reflects the bias-variance tradeoff in which bandwidth positively correlates with bias but negatively correlates with variance (see Homework 2.) The smaller bandwidth estimators suggest a trimodal density with peaks at around 10, 15, and 20 degrees Celsius. All kernels performed similarly with the exception of the rectangular kernel, which gave unstable estimates highly dependent on the bandwidth.

2.2 Loess Smoothed Humidity vs Temperature

To investigate Loess smoothing, we fixed the time of day at 10:05 AM and plotted humidity against temperature. We observe the negative correlation as before and fit Loess smoothers with polynomial degrees 0, 1, and 2 for four different spans. Here, the span is not an absolute bandwidth, rather it represents the proportion of points used in the smoother. For example, a span of .25 means we take the x closest points where x/n = .25 and n is the number of observations. Thus changing this value is akin to varying the bandwidth. Figure 2.2 shows the results. The degree 0 smoother does rather poorly compared to the higher degree versions as it is too flat for all spans. The degree 1 and 2 smoothers behave similarly, though the second-degree fit seems subtly more consistent across spans. As in the KDE plots, higher spans give a smoother trend while lower ones are more sensitive to the local behavior.

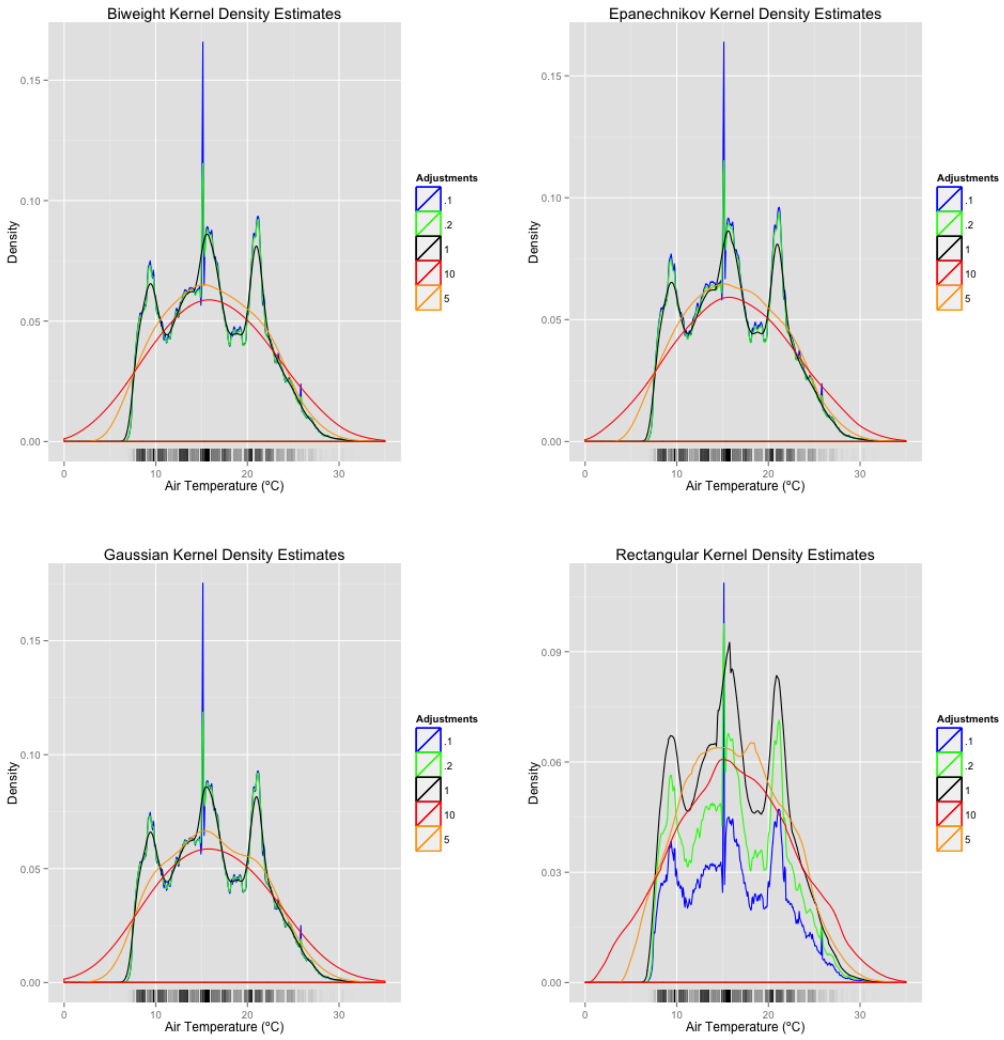

Figure 2.1: Kernel smoothed estimates of the temperature density for four different kernels. The rectangular kernel performs the worst while the others produce similar densities. Bandwidth = adjustment * .392 where .392 is the optimal bandwidth as calculated in R by bw.nrd0.

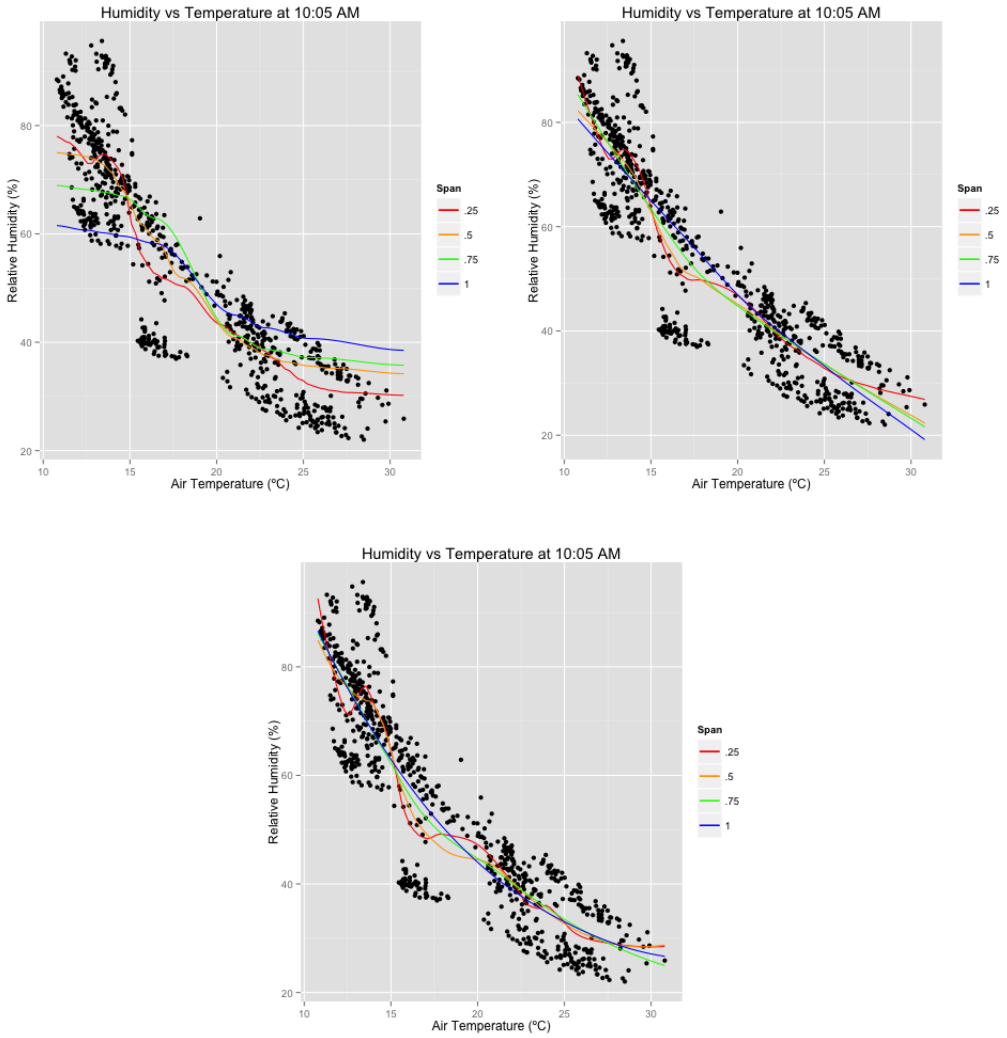

Figure 2.2: Upper Left: Degree 0 Loess smoother. Upper right: Degree 1 Loess smoother. Bottom: Degree 2 Loess smoother. The second degree smoother produces more consisten results, but the degree one gives similar trends.

3 Linguistic Data

3.1 Data quality and cleaning

Each row of the data contains a participant's answers and each column corresponds to a different question. The entries are categorical variables that give the responses in the form of numbers where each number is associated with a specific answer. There were initially 47471 participants and 67 questions considered. Since we want to eventually generate maps using latitude and longitude, we began by removing all rows for which a latitude and longitude were not present. This was a total of 1020 respondents. Several systematic reasons for the missing coordinate data were considered, including zip codes starting with 0 and outdated zip code databases. Investigation of these possiblities yielded no improvement. Additionally, some participants answered few of the questions and were removed. We set a cutoff of 20 missing answers. This left us with 45107 observations. While they were included in the data for clustering, any points corresponding to locations in Alaska and Hawaii were removed for plotting. The latitude and longitude columns were separated from the response columns before encoding to binary variables since that procedure would not work if they remained. Furthermore, they were not readded until clustering was complete since PCA and the clusters would be influenced by the presence of location data and we would likely get geographic clusters by default.

Having the data in categorical form was sufficient for the EDA as it allowed for plotting and tracking how often responses were recorded together. However, further analysis required a transformation into binary variables as described in the lab's prompt. This was accomplished by defining a function based on model.matrix and applying it to each row of the data. The intercept columns were then removed to give a set of 468 binary response variables where a 0 indicates that answer was not chosen and a 1 says the opposite.

3.2 Exploratory Data Analysis

Our EDA began with maps of the US overlaid with responses by latitude and longitude. Each point is colored by the responder's answer. For each question only the popular choices were included to prevent overplotting. The included plots correspond to the most important questions as measured by the loadings on the first 3 principal components. These are the plural second person, name for athletic shoes, and name for a water fountain questions. In Figure 3.1, the south is shown to strongly prefer using "y'all" while the rest of the nation favors "you", "you all", or "you guys" with no clear geographical trend among them. Figure 3.2 illustrates the divide in usage of "sneakers" and "tennis shoes." The northeast refers to the shoes in question as sneakers while the remainder of the country uses tennis shoes. In both examples, we see south Florida's behavior mimicking that of the northeast, perhaps as a consequence of the large number of transplants in the region. The final of our EDA plots, Figure 3.3, shows the geographic separation between the terms "water fountain" and "drinking fountain," with "drinking fountain" only being common in parts of the midwest and on the west coast. Considering these plots together, we note that each one defines a group. In fact, these correspond nicely to the clusters output by PCA and K-means.

To look at the relationship between a pair of questions, we look at the second person plural and shoes questions.

4 Dimension reduction methods

This is where you discuss and show plots about the results of whatever dimension reduction techniques you tried - PCA, hierarchical clustering, K-means, random projections, etc.

`http://youtu.be/-qgCEqmbsO4`
`http://youtu.be/FuRNZofyJgs`

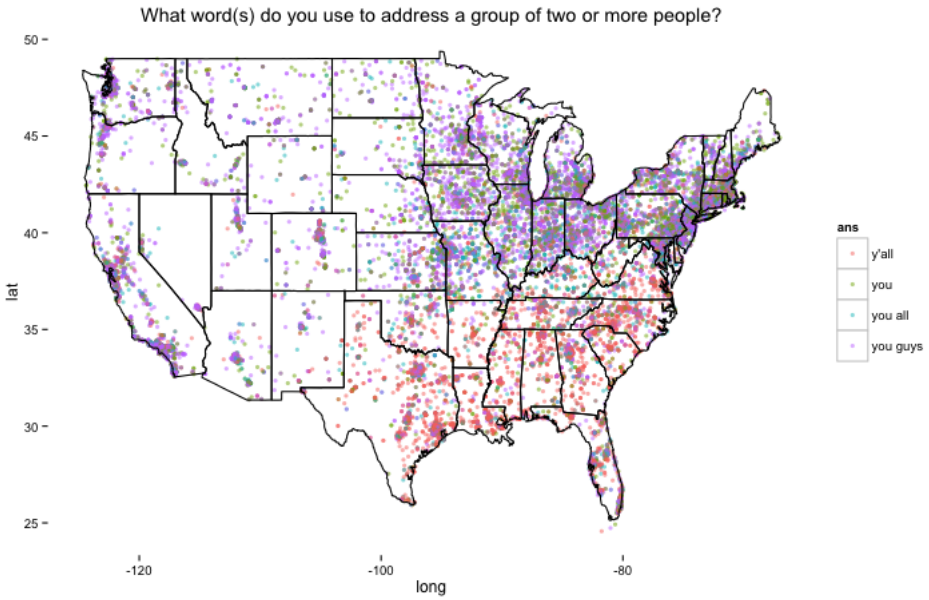

Figure 3.1: Most popular ways to address a group of people mapped by latitude and longitude. Y'all is popular in the south, but the other choices are evenly spread across the rest of the country.

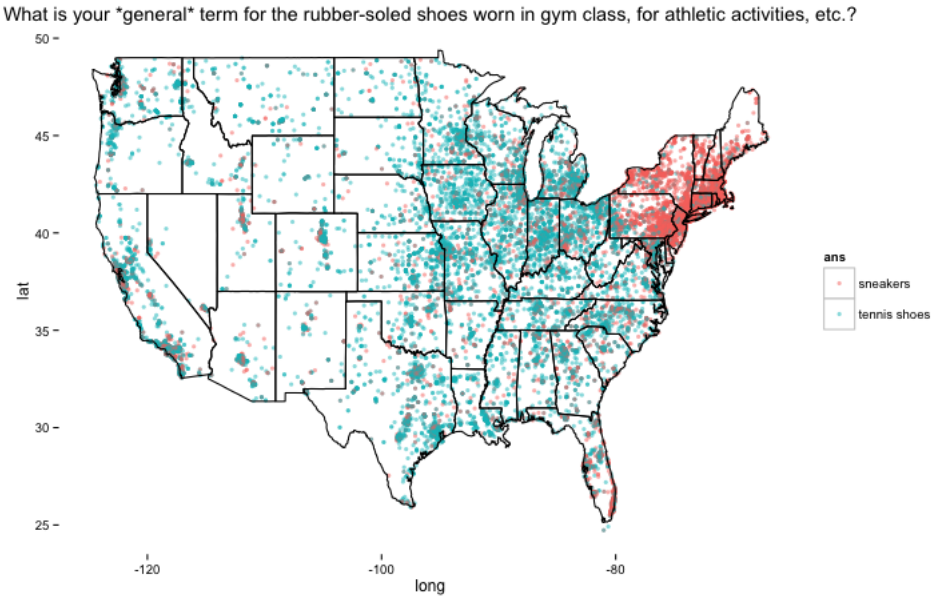

Figure 3.2: Mapped responses for tennis shoes/sneakers. The northeast is the only region to prefer sneakers.

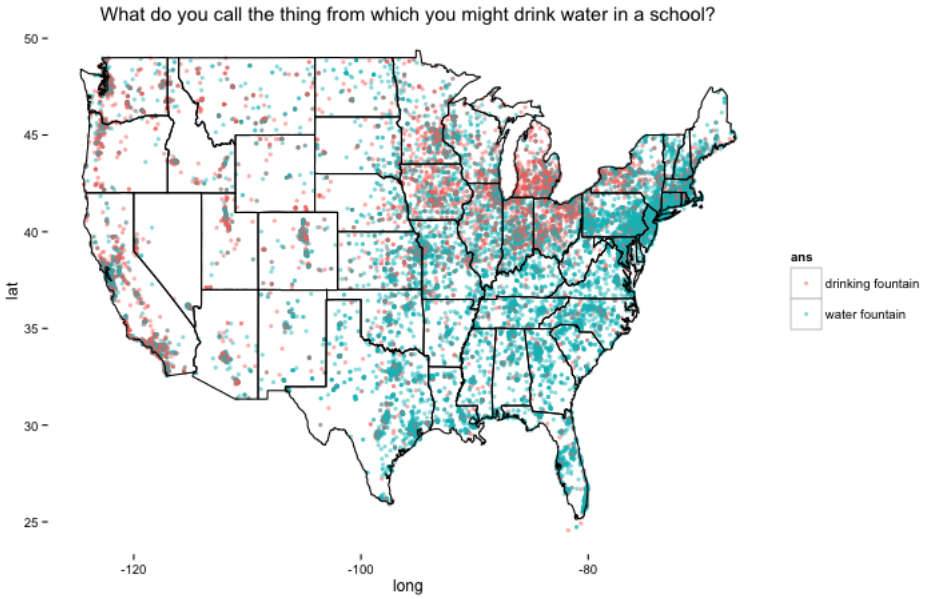

Figure 3.3: Map of water/drinking fountain usage. The upper midwest and west coast show a preference for drinking fountain while everyone else says water fountain.

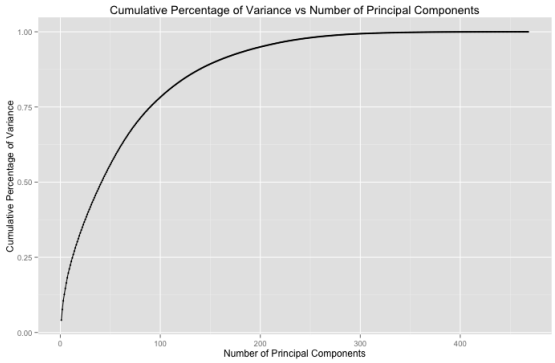

Figure 4.1: Cumulative variance as a function of the number of principal components.

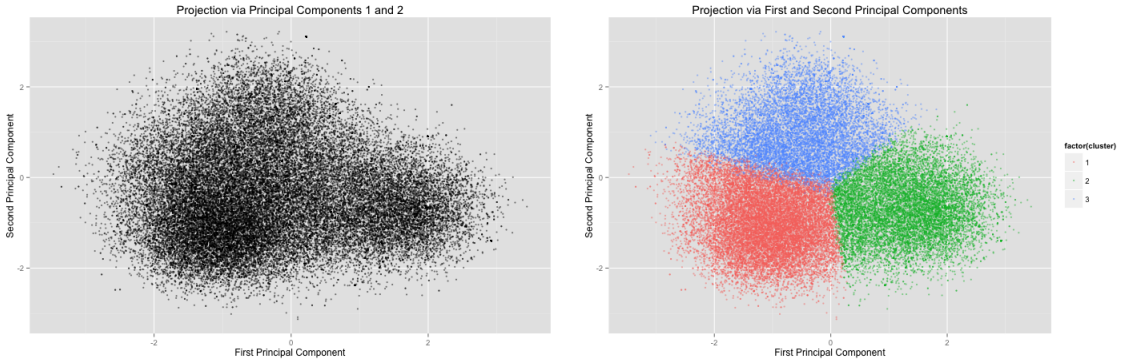

Figure 4.2: Scatter plot of projections onto the first and second principal component axes with and without annotation of K-means defined clusters.

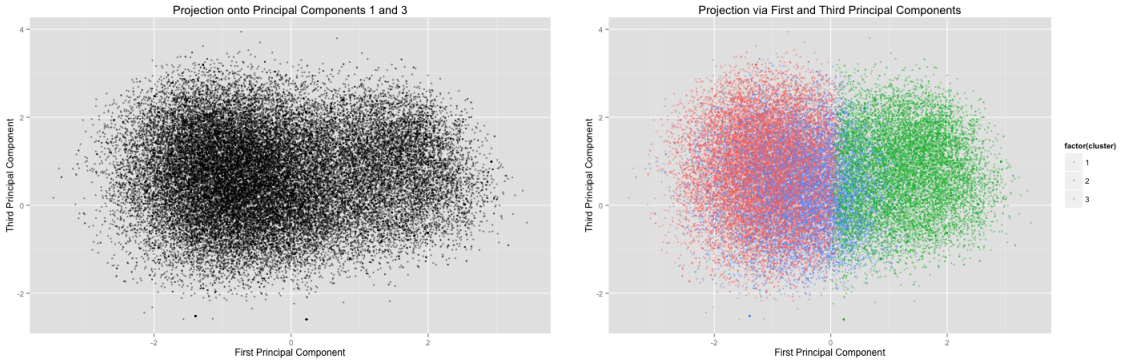

Figure 4.3: Scatter plot of projections onto the first and third principal component axes with and without annotation of K-means defined clusters.

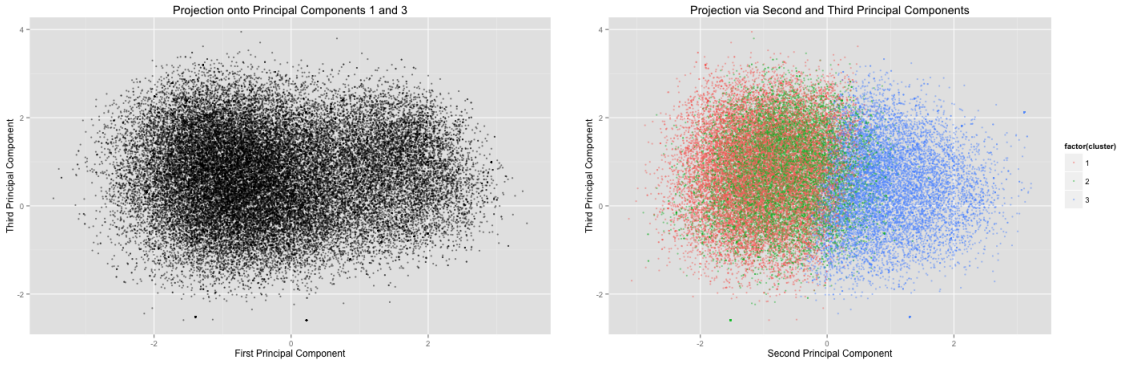

Figure 4.4: Scatter plot of projections onto the second and third principal component axes with and without annotation of K-means defined clusters.

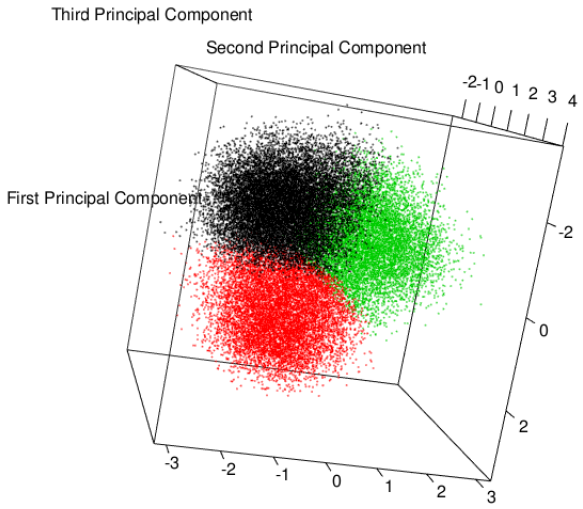

Figure 4.5: Screenshot of projection to first three principal axes with cluster coloring.

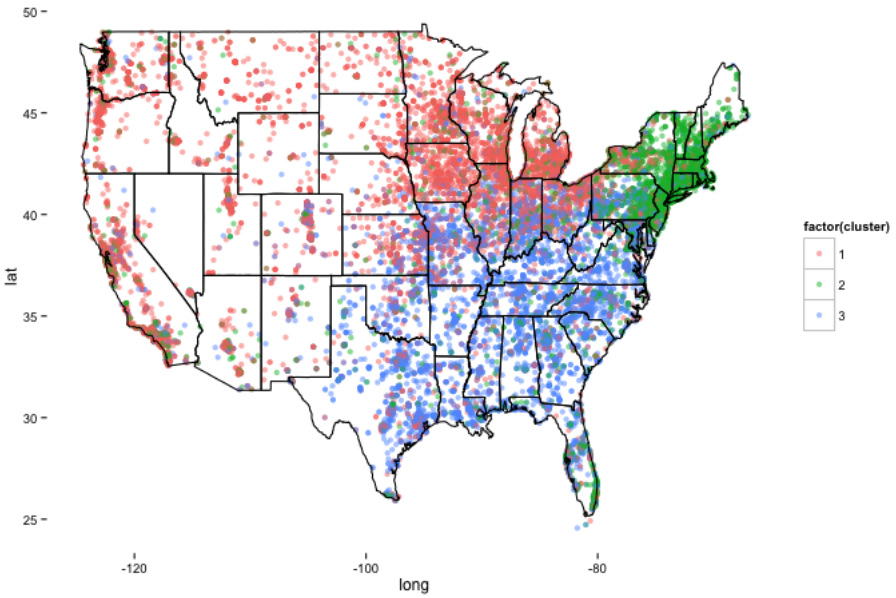

Figure 4.6: Responders mapped onto USA and colored by K-means generated clusters with K=3. We see 3 geographically well-defined clusters–northeast, south, and midwest/west.

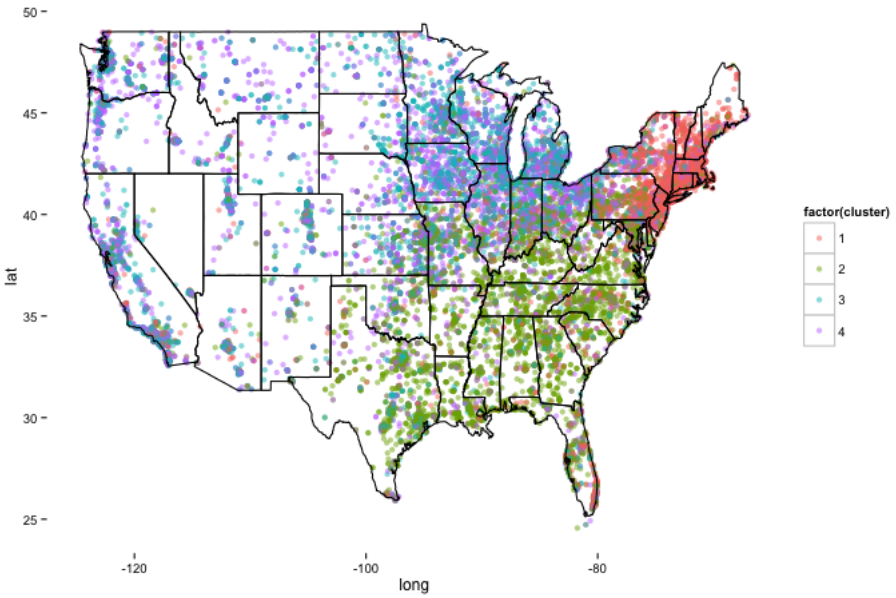

Figure 4.7: Responders mapped onto USA and colored by K-means generated clusters with K=4. The south and northeast clusters remain but two clusters are mixed throughout the west/midwest. This suggests 3 clusters was more appropriate.

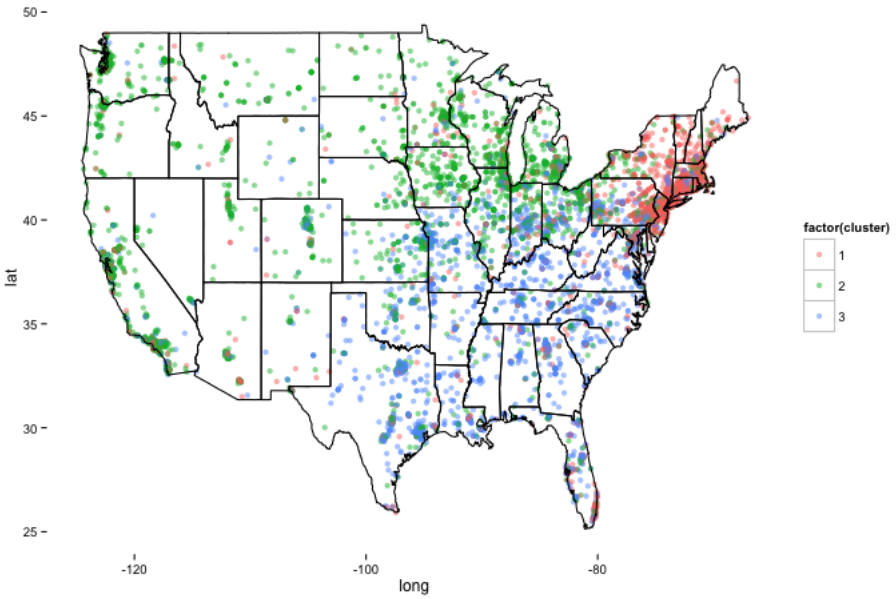

Figure 5.1: The results of PCA and clustering with K=3 with a random subsample of 10000 points. The same geographic clusters arise.

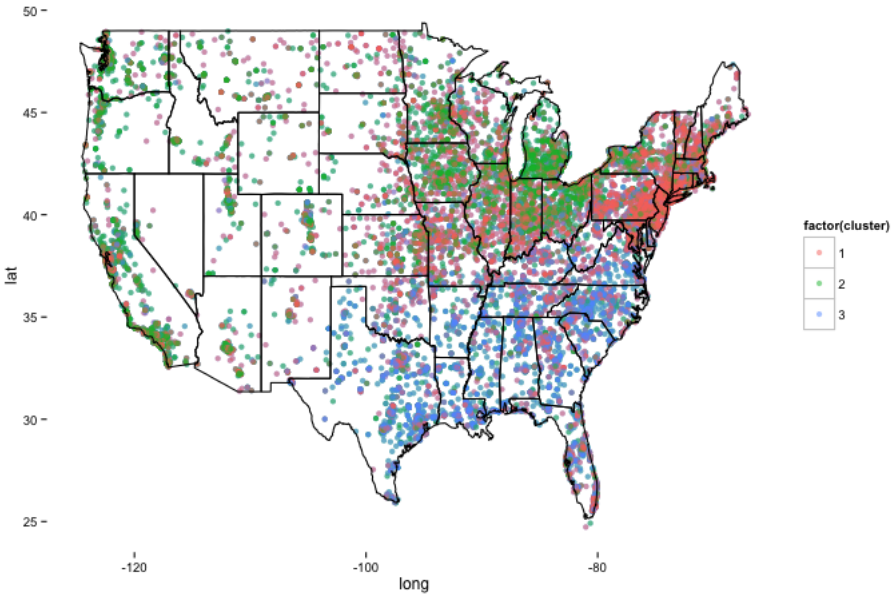

Figure 5.2: The results of PCA and clustering with K=3 where the data have been reduced to responses to questions 50 (not y'all/y'all), 73 (sneakers/tennis shoes), and 103 (water/drinking fountain). There is more overlap in the clusters than when the full data are used, but we still see the distinct geographic pattern.

5 Stability of findings to perturbation

6 Conclusion