Stat 133 Project Part 2 Report

Mengxian Qian

Analysis on removed observations

We first explored the question of whether some questions are more omitted than others. We tried clustering the observations that had omitted the 4 most omitted questions, (Q092, Q083, Q071, Q095) and plotting them out on the US map. (Exhibit7) There appeared to be no distinct geographical boundaries on the 4 most frequently omitted questions.

We then examined the question of whether removed observations were uniformly sampled based on geography. For simplification, we defined a unit of geography as a state. We then calculated the ratio of removed observations to the total number of observations per state. We then performed chi-square test on goodness-of-fit to the null hypothesis, a uniform distribution based on each state. The p value was less than 0.01, so we can strongly reject the null hypothesis.

Cluster Analysis

The data was too large to perform clustering, prcomp, and stability measuring in a reasonable time frame, so we randomly sampled the data. We then performed dimension reduction via PCA before clustering.

1.a **K means Clustering** with 10000 random samples, 3 clusters, 3 PCs (Exhibit 8) : There seems to be geographical boundaries between the three clusters, in the northeast, southeast, and northwest. A plot of PC1 against PC2 also reveals clustering boundaries, which suggest the clusters divide along questions that give more weight in PC1 and PC2, i.e. "Q056.2" "Q073.1" "Q080.1" "Q105.1" for PC1 and "Q050.9" "Q071.5" "Q076.4" "Q103.4" for PC2.

b 10000 random samples, 4 clusters, 3 PCs (Exhibit 9) : With k=4, there still seems to be 3 clusters, and and geographical boundaries are roughly the same as in k=3. Boundaries reveal themselves in PC1 vs PC3 and PC2 vs PC3.

c 3000 random samples, 3 clusters, 3 PCs (Exhibit 10) : Decreasing the number of random samples to 3000 didn't result in much loss of clustering boundaries.

d 10000 random samples, 3 clusters, 4 PCs (Exhibit 11) : Increasing PCs for kmeans didn't make much of a difference.

2.a **Hierarchical Clustering** with 10000 random samples, 3 clusters, 3 PCs (Exhibit 12) : For hclust, geographical boundaries were not as distinct as kmeans, and PC1 vs PC3 revealed the the most number of clustering boundaries.

b 10000 random samples, 4 clusters, 3 PCs (Exhibit 13) : Increasing k to 6 : boundaries get more blurry.

c 3000 random samples, 3 clusters, 3 PCs (Exhibit 14) : As with kmeans, decreasing the number of random samples to 3000 didn't result in much loss of clustering boundaries.

d 10000 random samples, 3 clusters, 4 PCs (Exhibit 15) : Worse boundaries with 4 PCs compared to 3 PCs.

3.a **Clustering Stability** with 10000 random samples, 3 PCs, clusters from 2 to 6 (Exhibit 16)

Next we measured stability of the hclust and kmeans clusters using clValid package. For both kmeans and hclust, stability increased with increasing number of clusters with average proportion of non-overlap (APN) and average distance between means (ADM), but decreased with the average distance (AD) and figure of merit (FOM) measures.

b 10000 random samples, 4 PCs, clusters from 2 to 6 (Exhibit 17)

We changed tol=0.7 in prcomp to get 4 PCs. Stability trends were similar to 3 PCs, with the following optimal scores for 4 PCs (right), which are quite similar to those of 3 PCs (left).

3PCs	Score	Method	Clusters	**4PCs**	Score	Method	Clusters
APN	0.0038	hierarchical	2	APN	0.0037	hierarchical	2
AD	1.9755	kmeans	6	AD	2.2258	kmeans	6
ADM	0.0488	hierarchical	2	ADM	0.0401	hierarchical	2
FOM	1.1128	hierarchical	6	FOM	1.0386	hierarchical	6

PCA Analysis and Mlogit analysis

We firstly conducted principal component analysis, and set the tol value to be 0.8, so that we get three components whose standard deviations are larger than or equal to 0.8 times the standard deviation of the first component. And we found that Q073, Q080 give more weight in the first principal component; Q65,Q76 in the second one; Q97, Q98 in the third one (Exhibit 1). In this case, we can say that the answer to these questions is relatively more informative, and we use Q73 and Q80 as representatives to explore the relationship between different questions, and also the relationship between the answers to the questions and geography.

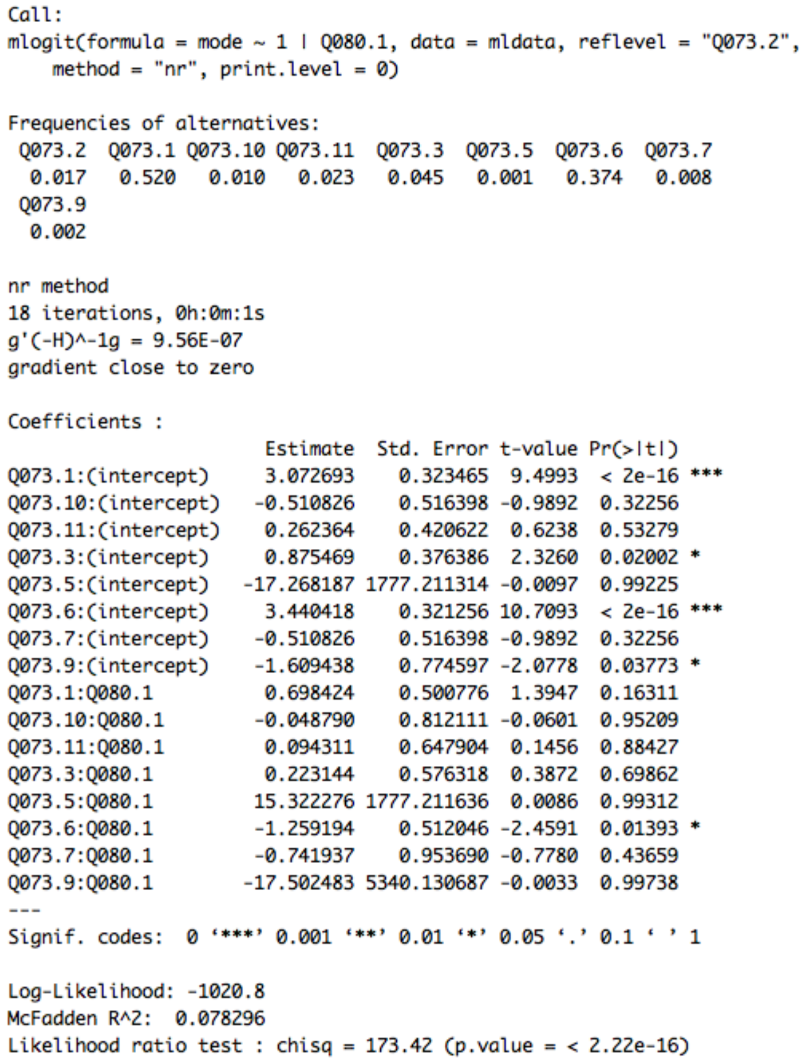
Since the reformatted data is discrete choice type of data, we use the Multinomial logit model in R to conduct multinomial logistic regression. We firstly conducted the analysis on Q73 and Q80.1, and randomly pick Q073.2 as the base alternative. As we can see in Exhibit 2, the coefficient of Q73.6 and Q080.1 is -1.26, which is statistically significant. So we can conclude that if people choose Q80.1, it is less likely that he/she would choose Q73.6, comparing to the base case Q073.2. In this case, we can predict that if a person call the situation when rain falls while the sun is shining as 'sunshower'. It is less likely she/he would call the rubber-soled shoes worn in gym class as 'tennis shoes', compared to the possibility that she/he would call it 'shoes'!

Using the same model above, we implement multinomial logit model on Q73 and the latitude of the respondents, and the result is shown in Exhibit 3. As we can see in the output, the coefficient between Q73.3 is positively related to Q80.1, using Q73.2 as base alternative, and it is statistically significant. So can predict that the higher the latitude, the more likely people would call the rubber-soled shoes 'gymshoes'. As we can see in the map of the continental United States (Exhbit 4) showing the regional distribution of people's answers for this Q73.3, it is the same result as we predicted! Similar to the previous analysis, we conducted the regression analysis on Q73 and longitude. As we can see in Exhibit 5, the coefficient between Q73.1 and longitude is significantly positive and the coefficient between Q73.7 and longitude is significantly negative. So we can predict that more people would call ruber-soled shoes worn in gym class as 'sneaker' in the higher longitude; however, less people would call it 'running schoes'. As we can see in the map of the continental United States (Exhbit 6) showing the regional distribution of people's answers for this Q73.1, it is also the same as what we predicted！

Exhibit 1: PCA Analysis- Questions giving more weight in principal components

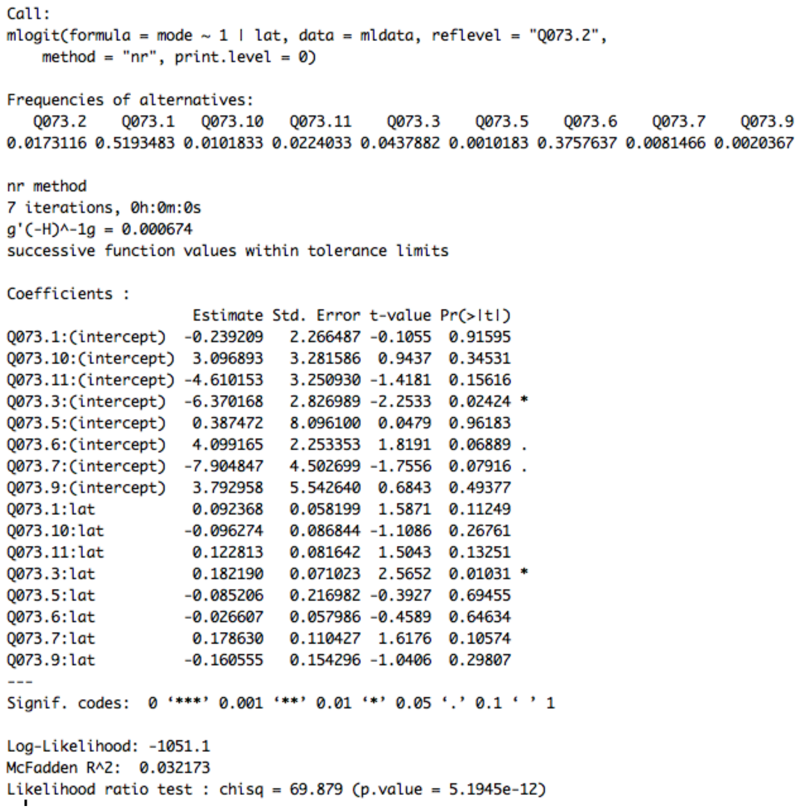
PC1	PC2	PC3
Q073.6	Q065.2	Q097.5
Q080.8	Q076.1	Q098.1
Q105.2	Q103.3	Q107.3
Q106.1	Q118.7	Q120.1

Exhibit 2: Multinomial logit model (Q73~Q80)



```
Call:
mlogit(formula = mode ~ 1 | Q080.1, data = mldata, reflevel = "Q073.2",
    method = "nr", print.level = 0)

Frequencies of alternatives:
 Q073.2  Q073.1 Q073.10 Q073.11  Q073.3  Q073.5  Q073.6  Q073.7
  0.017   0.520   0.010   0.023   0.045   0.001   0.374   0.008
 Q073.9
  0.002

nr method
18 iterations, 0h:0m:1s
g'(-H)^-1g = 9.56E-07
gradient close to zero

Coefficients :
                       Estimate  Std. Error t-value Pr(>|t|)
Q073.1:(intercept)     3.072693    0.323465  9.4993  < 2e-16 ***
Q073.10:(intercept)   -0.510826    0.516398 -0.9892  0.32256
Q073.11:(intercept)    0.262364    0.420622  0.6238  0.53279
Q073.3:(intercept)     0.875469    0.376386  2.3260  0.02002 *
Q073.5:(intercept)   -17.268187 1777.211314 -0.0097  0.99225
Q073.6:(intercept)     3.440418    0.321256 10.7093  < 2e-16 ***
Q073.7:(intercept)    -0.510826    0.516398 -0.9892  0.32256
Q073.9:(intercept)    -1.609438    0.774597 -2.0778  0.03773 *
Q073.1:Q080.1          0.698424    0.500776  1.3947  0.16311
Q073.10:Q080.1        -0.048790    0.812111 -0.0601  0.95209
Q073.11:Q080.1         0.094311    0.647904  0.1456  0.88427
Q073.3:Q080.1          0.223144    0.576318  0.3872  0.69862
Q073.5:Q080.1         15.322276 1777.211636  0.0086  0.99312
Q073.6:Q080.1         -1.259194    0.512046 -2.4591  0.01393 *
Q073.7:Q080.1         -0.741937    0.953690 -0.7780  0.43659
Q073.9:Q080.1        -17.502483 5340.130687 -0.0033  0.99738
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -1020.8
McFadden R^2:  0.078296
Likelihood ratio test : chisq = 173.42 (p.value = < 2.22e-16)
```

Exhibit 3: Multinomial logit model (Q73~lan)



```
Call:
mlogit(formula = mode ~ 1 | lat, data = mldata, reflevel = "Q073.2",
    method = "nr", print.level = 0)

Frequencies of alternatives:
   Q073.2    Q073.1   Q073.10   Q073.11    Q073.3    Q073.5    Q073.6    Q073.7    Q073.9
0.0173116 0.5193483 0.0101833 0.0224033 0.0437882 0.0010183 0.3757637 0.0081466 0.0020367

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 0.000674
successive function values within tolerance limits

Coefficients :
                     Estimate Std. Error t-value Pr(>|t|)
Q073.1:(intercept)  -0.239209   2.266487 -0.1055  0.91595
Q073.10:(intercept)  3.096893   3.281586  0.9437  0.34531
Q073.11:(intercept) -4.610153   3.250930 -1.4181  0.15616
Q073.3:(intercept)  -6.370168   2.826989 -2.2533  0.02424 *
Q073.5:(intercept)   0.387472   8.096100  0.0479  0.96183
Q073.6:(intercept)   4.099165   2.253353  1.8191  0.06889 .
Q073.7:(intercept)  -7.904847   4.502699 -1.7556  0.07916 .
Q073.9:(intercept)   3.792958   5.542640  0.6843  0.49377
Q073.1:lat           0.092368   0.058199  1.5871  0.11249
Q073.10:lat         -0.096274   0.086844 -1.1086  0.26761
Q073.11:lat          0.122813   0.081642  1.5043  0.13251
Q073.3:lat           0.182190   0.071023  2.5652  0.01031 *
Q073.5:lat          -0.085206   0.216982 -0.3927  0.69455
Q073.6:lat          -0.026607   0.057986 -0.4589  0.64634
Q073.7:lat           0.178630   0.110427  1.6176  0.10574
Q073.9:lat          -0.160555   0.154296 -1.0406  0.29807
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -1051.1
McFadden R^2:  0.032173
Likelihood ratio test : chisq = 69.879 (p.value = 5.1945e-12)
```

Exhibit 4:

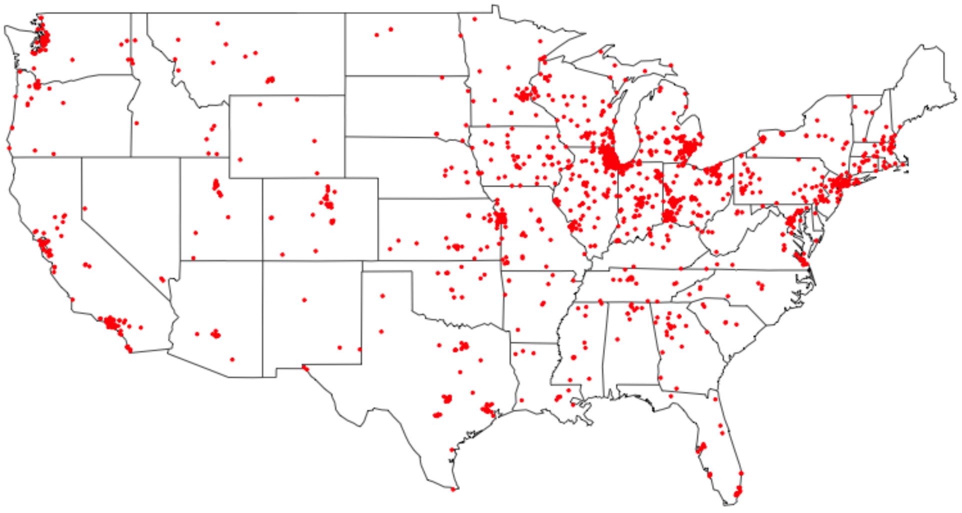
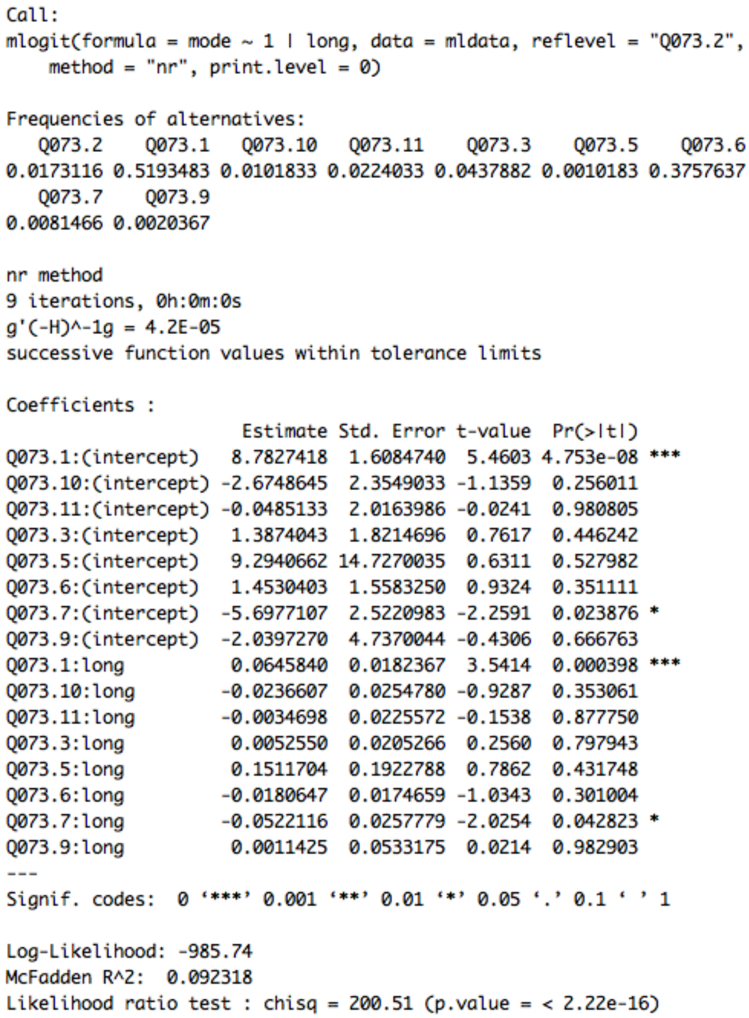


Exhibit 5: Multinomial logit model (Q73~long)



```
Call:
mlogit(formula = mode ~ 1 | long, data = mldata, reflevel = "Q073.2",
    method = "nr", print.level = 0)

Frequencies of alternatives:
   Q073.2    Q073.1   Q073.10   Q073.11    Q073.3    Q073.5    Q073.6
0.0173116 0.5193483 0.0101833 0.0224033 0.0437882 0.0010183 0.3757637
   Q073.7    Q073.9
0.0081466 0.0020367

nr method
9 iterations, 0h:0m:0s
g'(-H)^-1g = 4.2E-05
successive function values within tolerance limits

Coefficients :
                      Estimate Std. Error t-value  Pr(>|t|)
Q073.1:(intercept)   8.7827418  1.6084740  5.4603 4.753e-08 ***
Q073.10:(intercept) -2.6748645  2.3549033 -1.1359  0.256011
Q073.11:(intercept) -0.0485133  2.0163986 -0.0241  0.980805
Q073.3:(intercept)   1.3874043  1.8214696  0.7617  0.446242
Q073.5:(intercept)   9.2940662 14.7270035  0.6311  0.527982
Q073.6:(intercept)   1.4530403  1.5583250  0.9324  0.351111
Q073.7:(intercept)  -5.6977107  2.5220983 -2.2591  0.023876 *
Q073.9:(intercept)  -2.0397270  4.7370044 -0.4306  0.666763
Q073.1:long          0.0645840  0.0182367  3.5414  0.000398 ***
Q073.10:long        -0.0236607  0.0254780 -0.9287  0.353061
Q073.11:long        -0.0034698  0.0225572 -0.1538  0.877750
Q073.3:long          0.0052550  0.0205266  0.2560  0.797943
Q073.5:long          0.1511704  0.1922788  0.7862  0.431748
Q073.6:long         -0.0180647  0.0174659 -1.0343  0.301004
Q073.7:long         -0.0522116  0.0257779 -2.0254  0.042823 *
Q073.9:long          0.0011425  0.0533175  0.0214  0.982903
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -985.74
McFadden R^2:  0.092318
Likelihood ratio test : chisq = 200.51 (p.value = < 2.22e-16)
```

Exhibit 6:

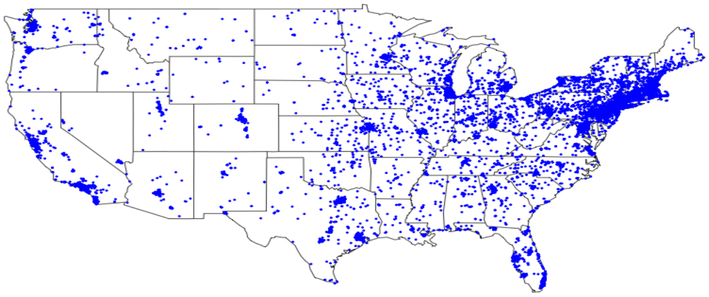


Exhibit 7.

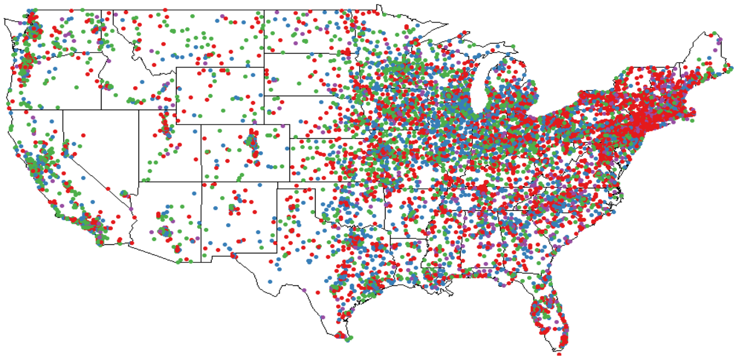
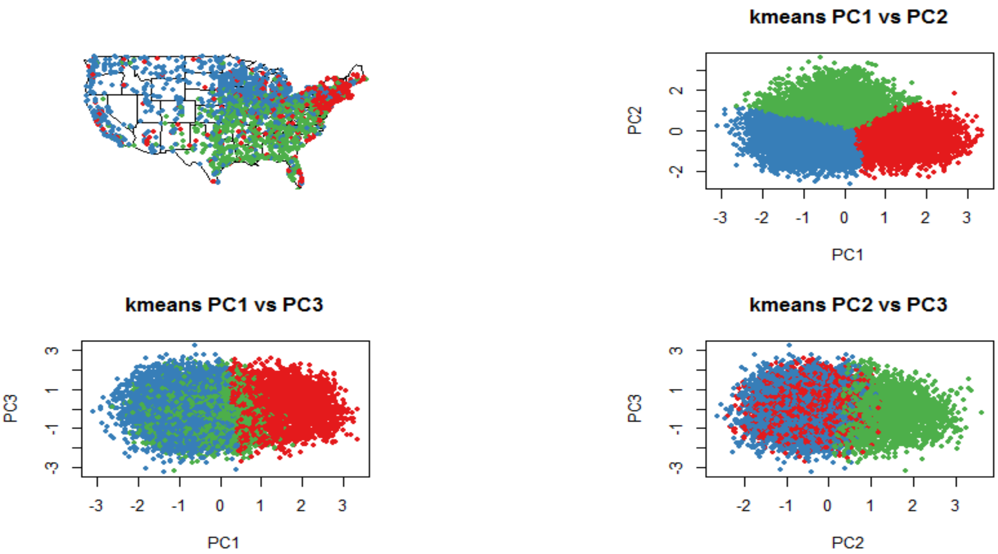


Exhibit 8.



kmeans PC1 vs PC2

kmeans PC1 vs PC3

kmeans PC2 vs PC3

Exhibit 9.

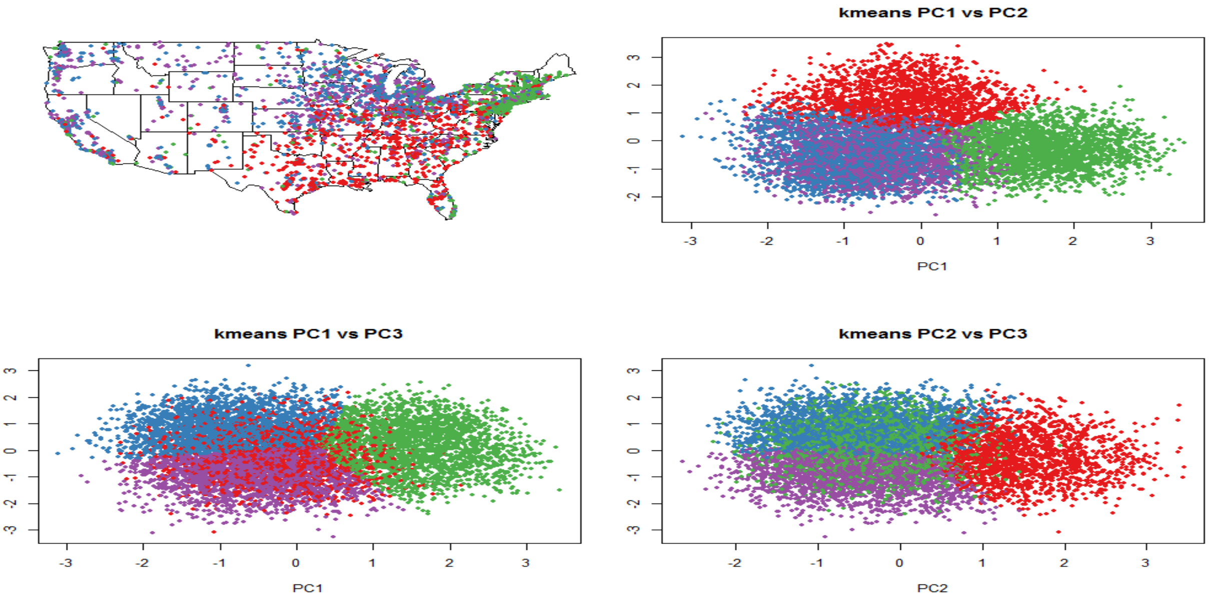


Exhibit 10.

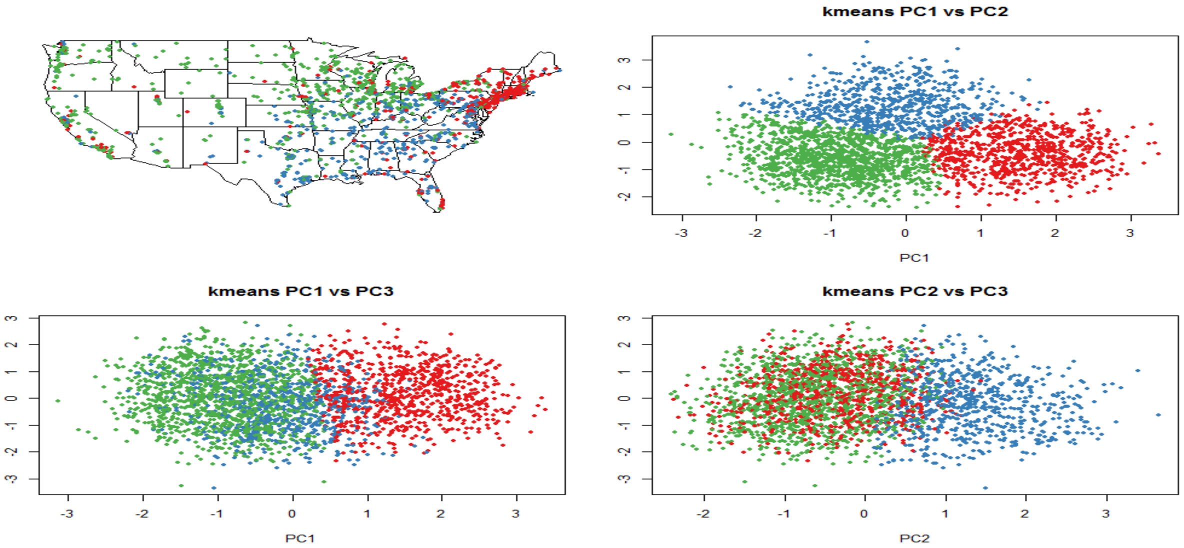


Exhibit 11.

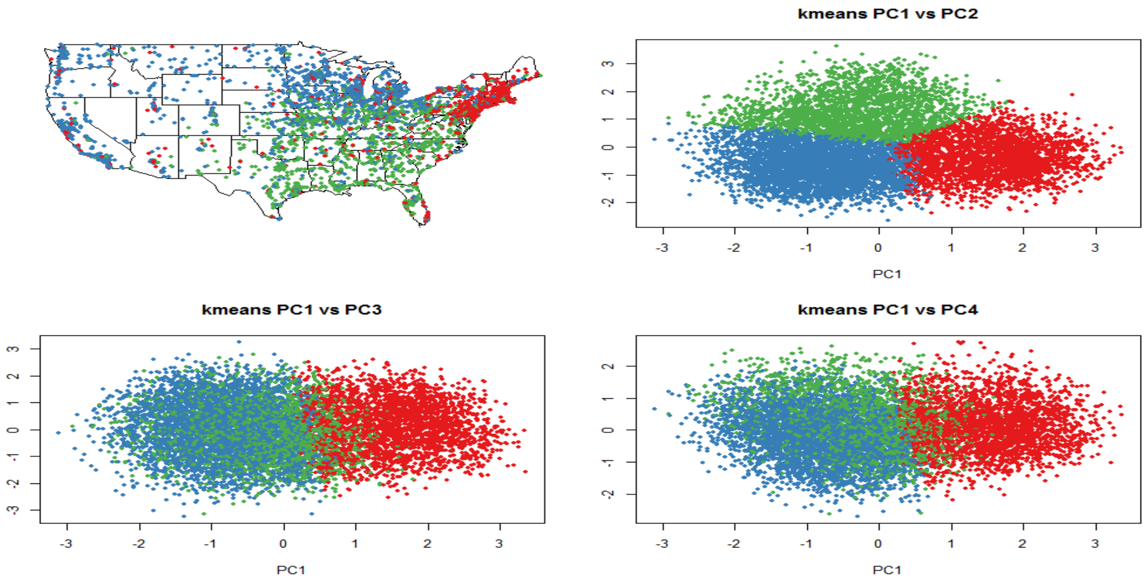


Exhibit 12.

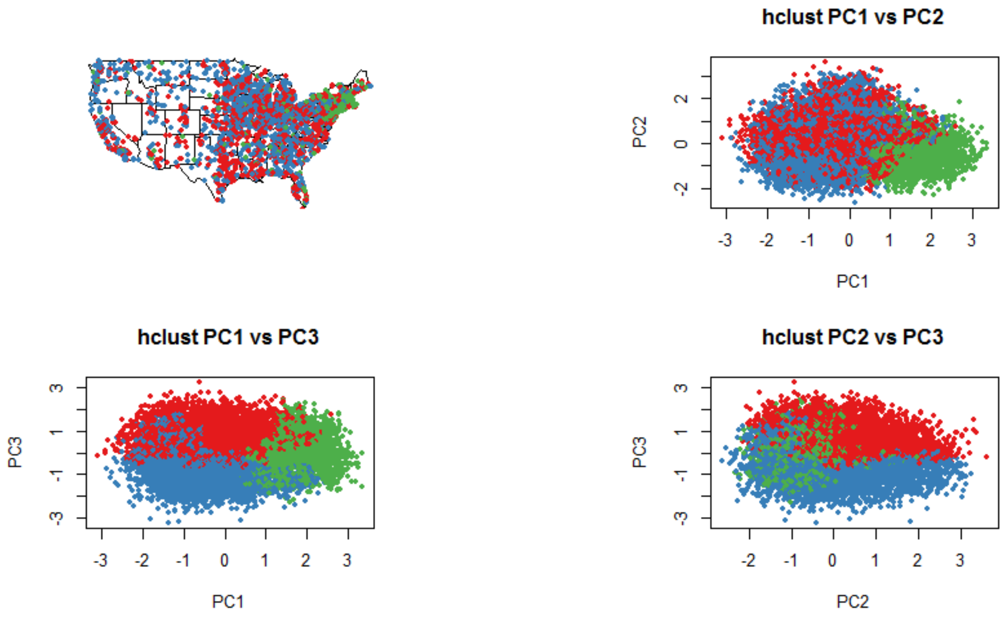


Exhibit 13.

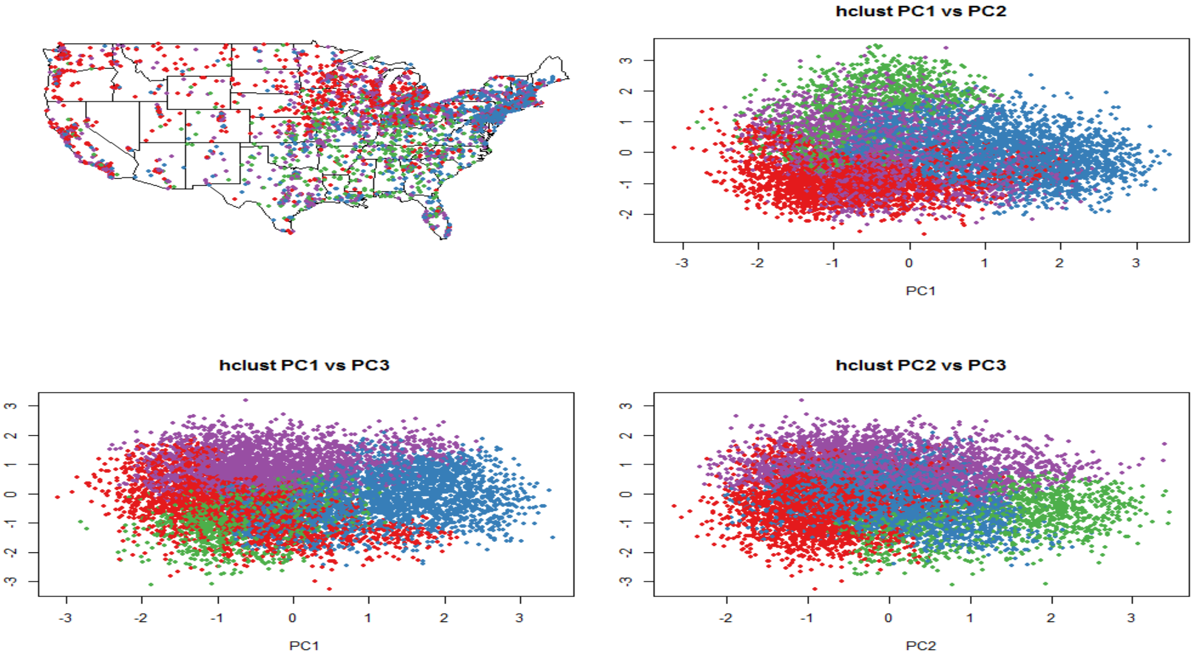


Exhibit 14.

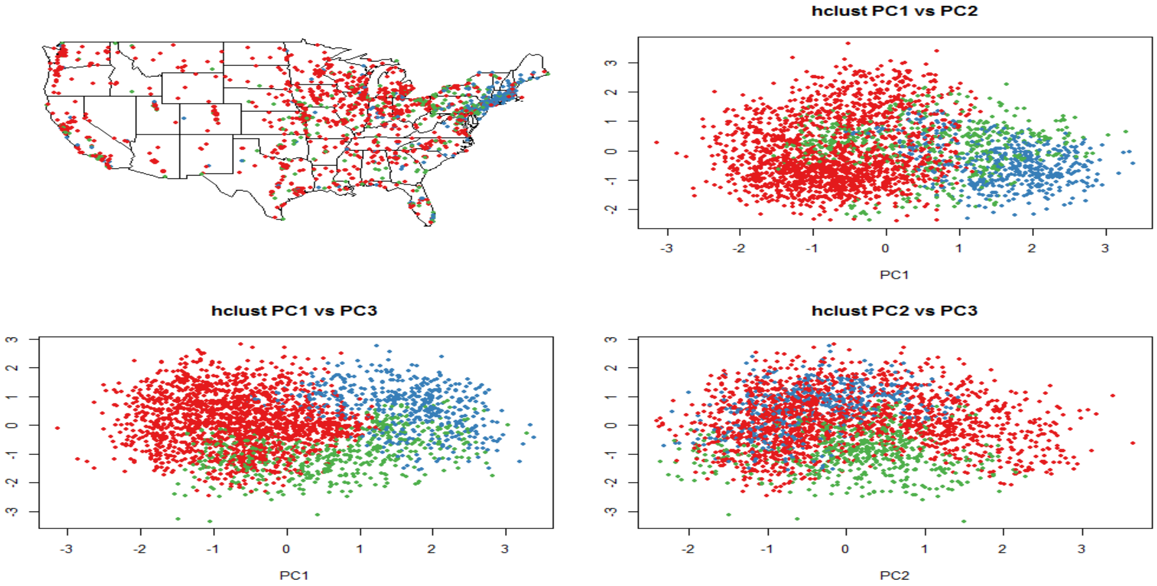


Exhibit 15.

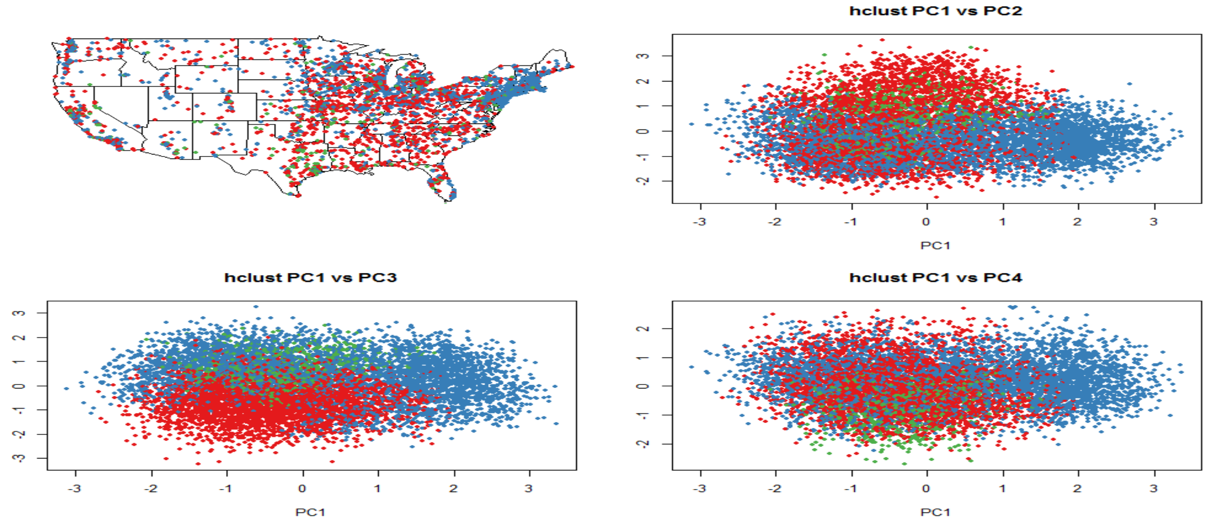


Exhibit 16.

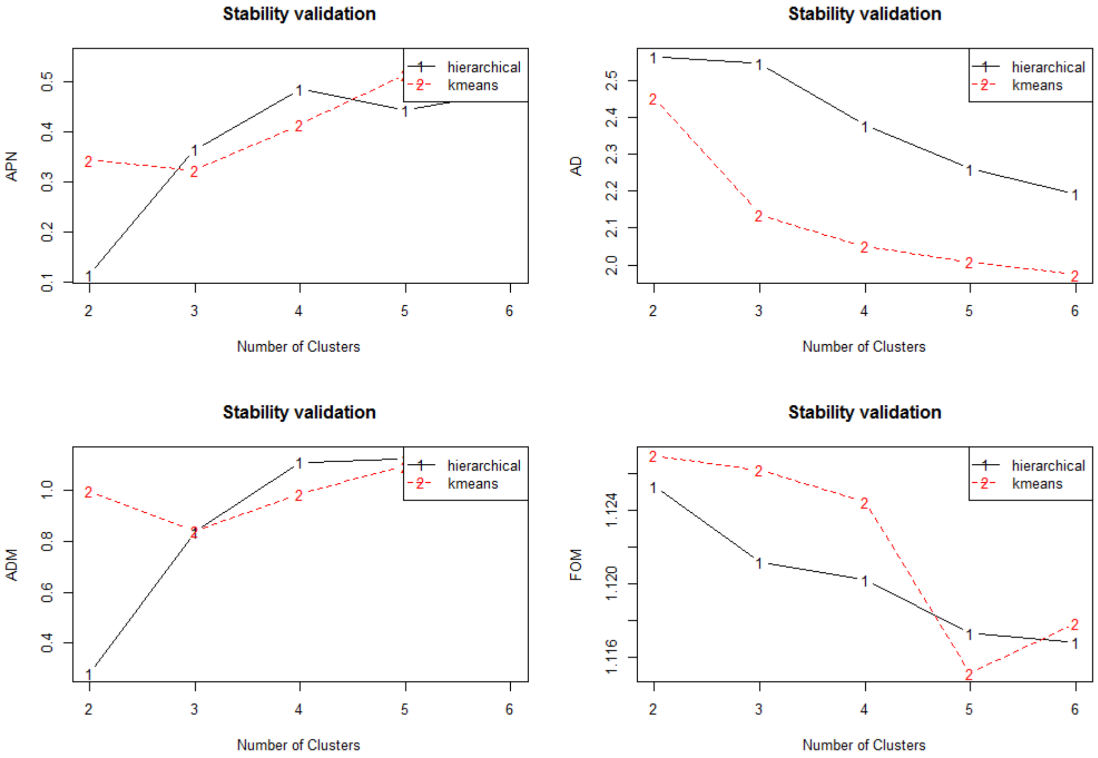
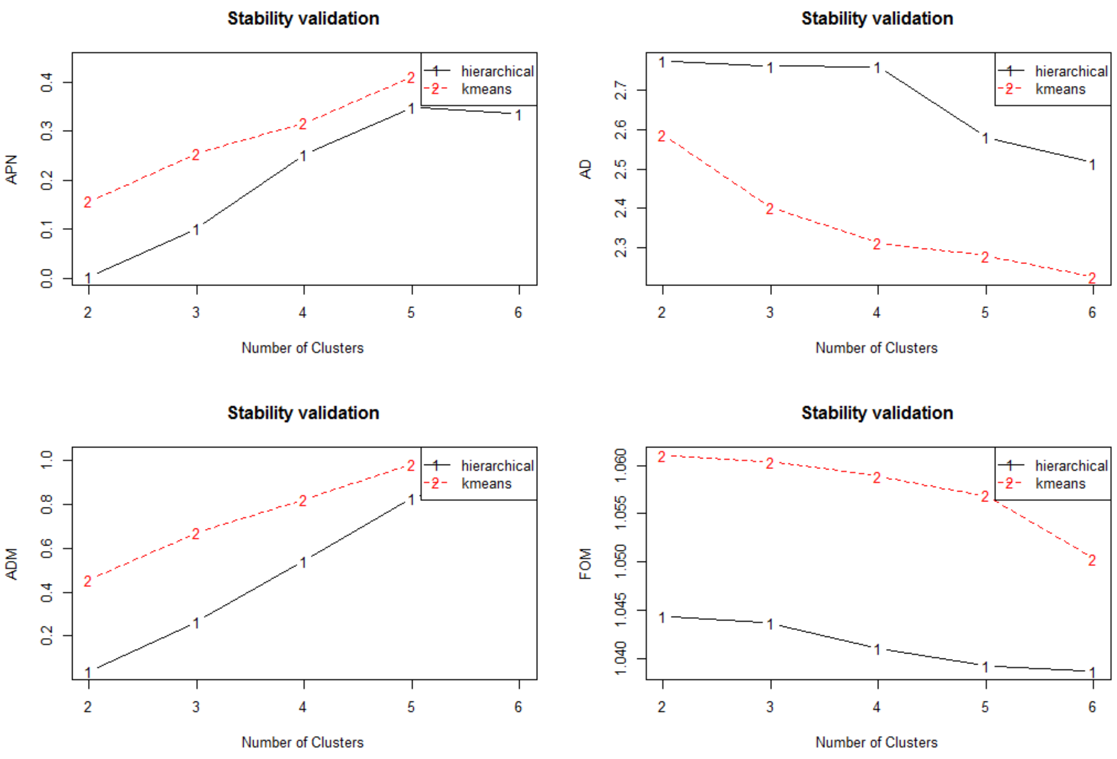


Exhibit 17.


Stability validation
APN
0.0
0.1
0.2
0.3
0.4
hierarchical
kmeans
2
3
4
5
6
Number of Clusters
Stability validation
AD
2.3
2.4
2.5
2.6
2.7
hierarchical
kmeans
2
3
4
5
6
Number of Clusters
Stability validation
ADM
0.2
0.4
0.6
0.8
1.0
hierarchical
kmeans
2
3
4
5
6
Number of Clusters
Stability validation
FOM
1.040
1.045
1.050
1.055
1.060
hierarchical
kmeans
2
3
4
5
6
Number of Clusters