Report for the final Project

Yuchen Dong, Zhenyang Zhang

Methodology and Results

Clustering：

In order to better understand the data, we need clustering to see if certain linguistic trend is linked to geographic locations. First we sample the data, and use kmeans in R to see how points are located around United States with respect to different numbers of clusters. From Figure 1 to 5 we can see that when k equals 3, there are three different groups: northeast New England, Southern and the other. As k increases, the groups become hard to differentiate and the plot is not very informative. To quantify the behaviors of various ks, we calculate silhouette of a datum to see how closely points are around their centers. In the figures, the higher ASW is, the better the clustering is. From Figure 6, in which sample size varies from 2000 to 5000, k=2 is better in one case but mostly k=3 is a reasonable solution. To further verify our argument that k=3 is a good clustering, we do a few more sampling of sample size 2000: k=3 performs the best among all cases.

Principal Component Analysis:

In this part we want to see if the survey can actually be shorten by doing a PCA: to see what answers to what questions matter the most. Unfortunately, we find out that no PCs are dominating: based on table 1, PC1 has 1.9% proportion of variance. However we still manage to pick out PC1 through PC6, which contribute at least 1% of variance, and try to find out what answers are influencing the PC (either positively or negatively). We take into account all variables that have coefficients whose absolute values are greater or equal to 0.1. We count the times that a problem appears among the 6 PCs. The questions we consider as important are 56,63,64,65,66,74,76,80,86,99,103,106. So if we only use answers to these questions and do clustering again, with k=2 and k=3 deduced from previous part, the clustering looks similar to the previous plots. This suggests that our result of major questions that matter work smoothly. But if we repeat the step of silhouette as we did before, we can see that k=2 is better in all cases. This is reasonable since the number of variables is very small. So although our data reduction performs well, it is not suggestive to cut the rest questions since they also play a big role.

Random Forest:

As shown in previous parts, geographical factors are related to linguistic trends. In this part, we are going to use random forest to try to build up a model to estimate or predict the location of the questionee. We do a single random sample based on the data, and use the rest as test group. We calculate the

average difference of longitude and latitude. We replicate the process of sampling and average difference calculation several times and plot out the histogram. From the histogram we can see that difference of latitude is centered around 2.85 degrees, which is around 300 kilometers in reality. But the average difference in longitude is centered at around 91 degrees, which is a very large error. From this we can conclude that the prediction of latitude is acceptable while we cannot rely on the prediction of longitude. There will be discussion about the error in the next part.

Basic Look at Conditional Probabilities Between Two Questions:

Although we cannot easily predict a person's location precisely by his or her answer to the survey, we are interested to see if we can predict the answer of a question based on the answer of another. We calculate the conditional probability and try to find out if there is any pattern. We basically look at problem 65 66 and 76 which we consider as major components in the PCA part. In the appendix, table 2 shows given the answer of 66, how people respond to 65. We can see given the person answers 3 for 66, he is very likely to answer 2 for 65. But in other cases although people tend to answer 1,2,3 in most cases, there is no tendency that we can regard one over another. So we think that we cannot use the data to predict easily.

Conclusion and Discussions

Through clustering and plotting it is easy to find out that language trends are closely linked to geographical factors. With the help of silhouette, we can easily show that the clustering with k=3 is a good choice. If we try to make the survey concise, the clustering will be the northeast-south-the-other pattern. However, the long list of questions cannot be easily cut off or else the results of analysis may change, as pointed out in the silhouette analysis. Moreover, it is hard for us to predict a person's location since points are widely spread out around the country. In order to predict more precisely, we need to omit part of the data. One possible solution is that we simply take east coast data into account and the error in longitude might be reduced. For the last part, although our try is not very successful, it supports our claim before that the questions might not be closely related. We may build possible predictions based on this discrete case with logistic models.

Appendix

Figure 1

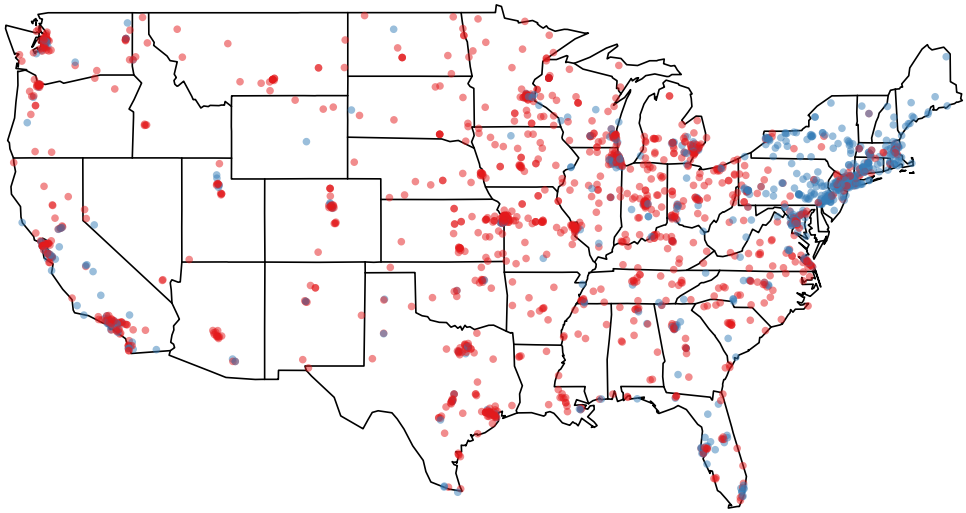

Figure 2

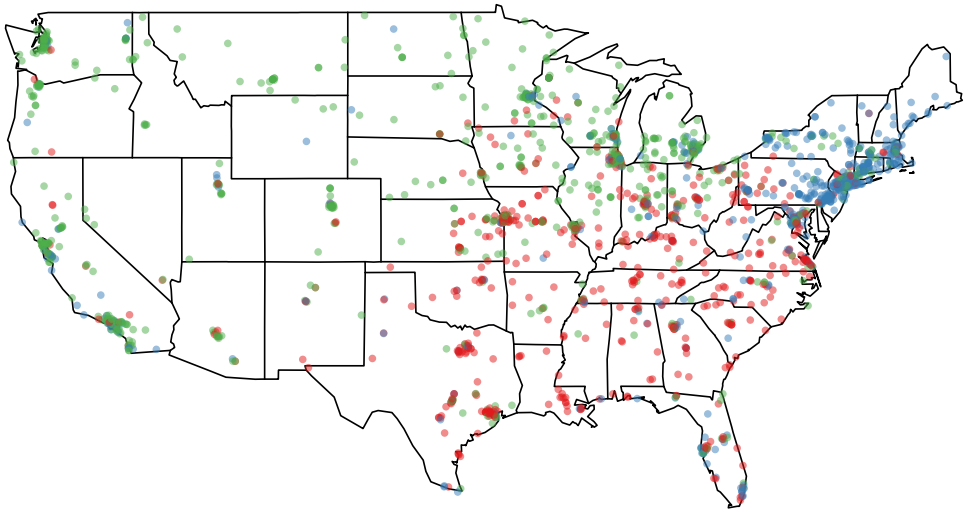

Figure 3

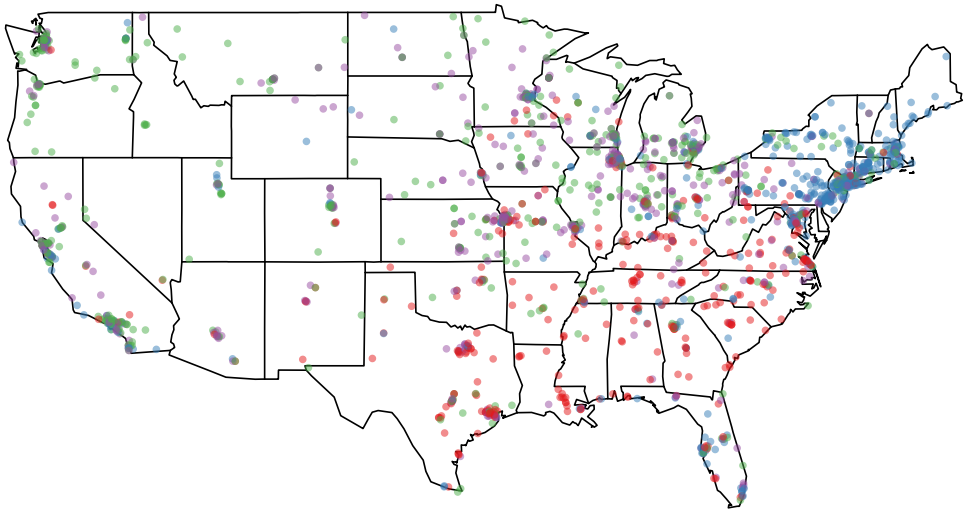

Figure 4

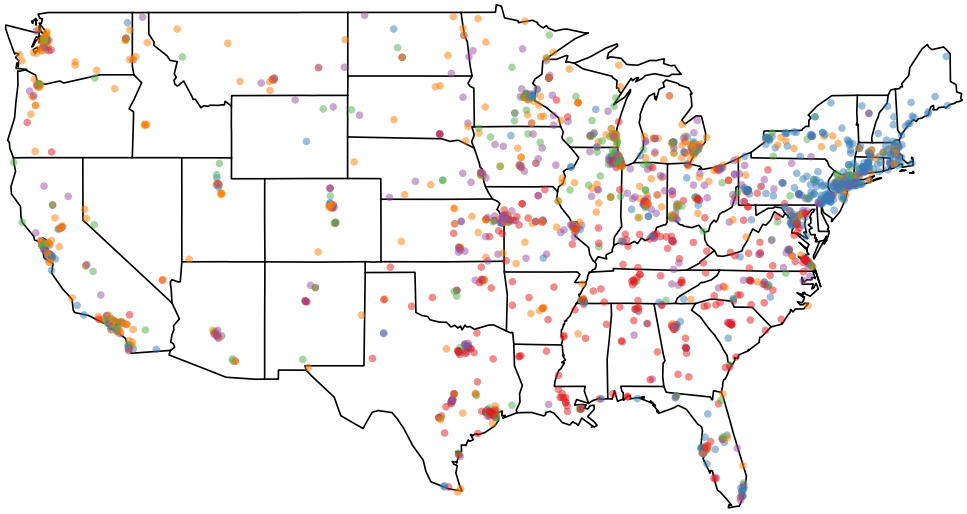

Figure 5

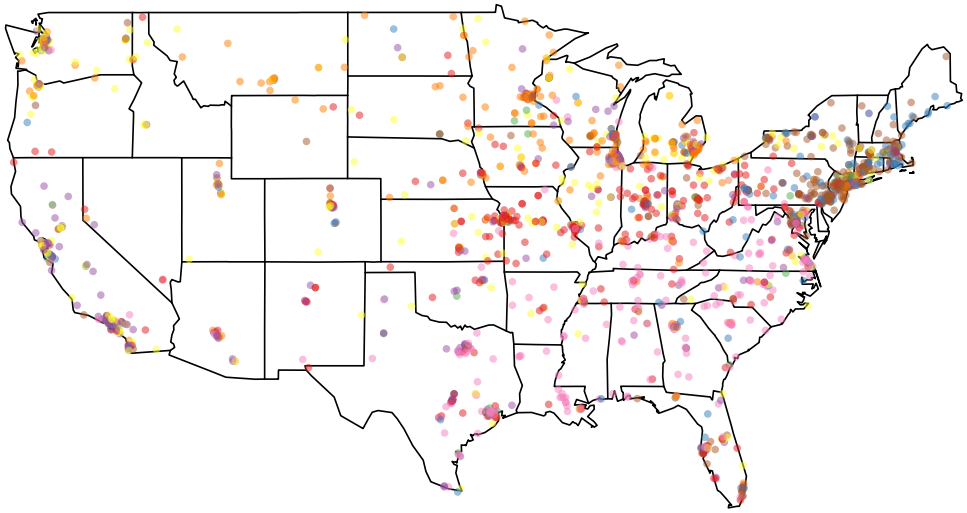

Figure 6

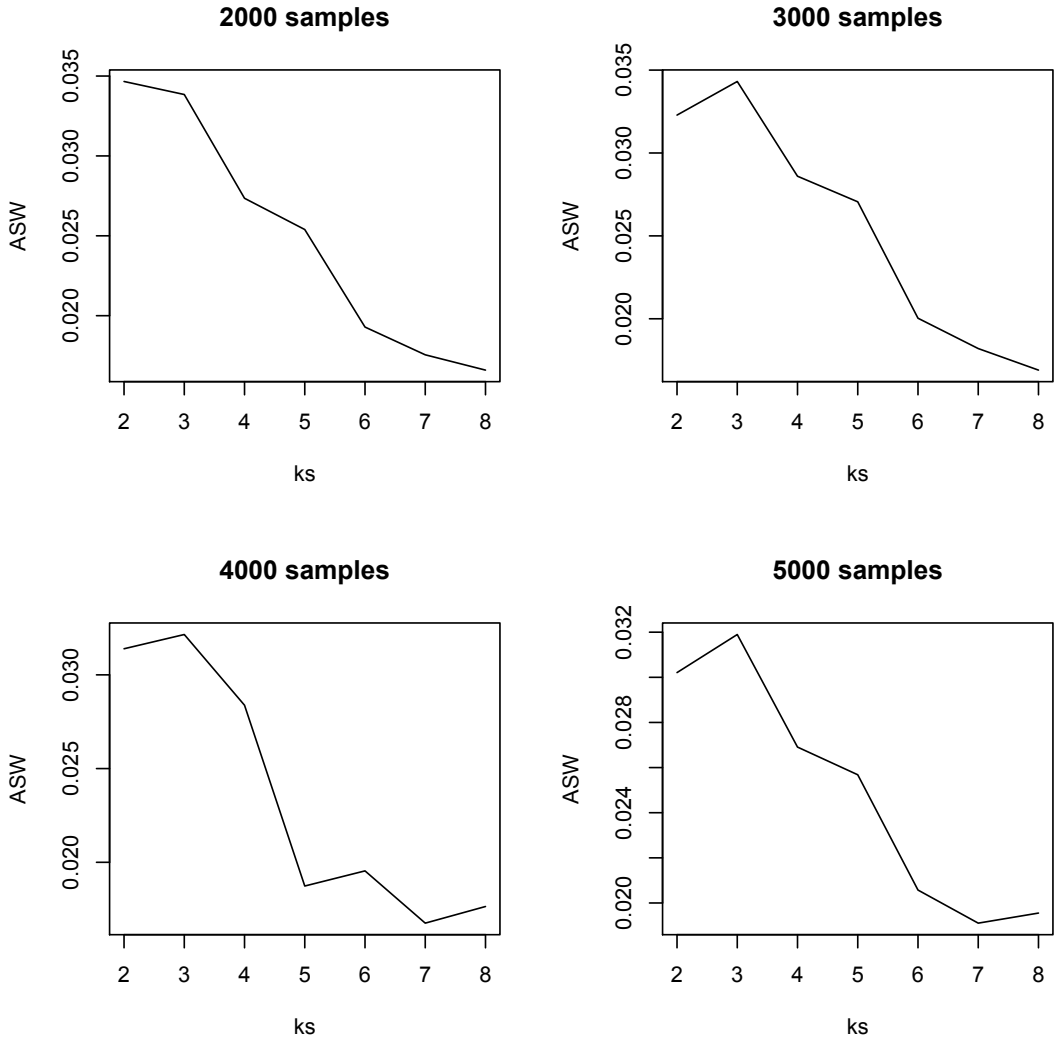
2000 samples
3000 samples
4000 samples
5000 samples
ASW
ks
0.020
0.025
0.030
0.035
0.024
0.028
0.032
2
3
4
5
6
7
8

Figure 7

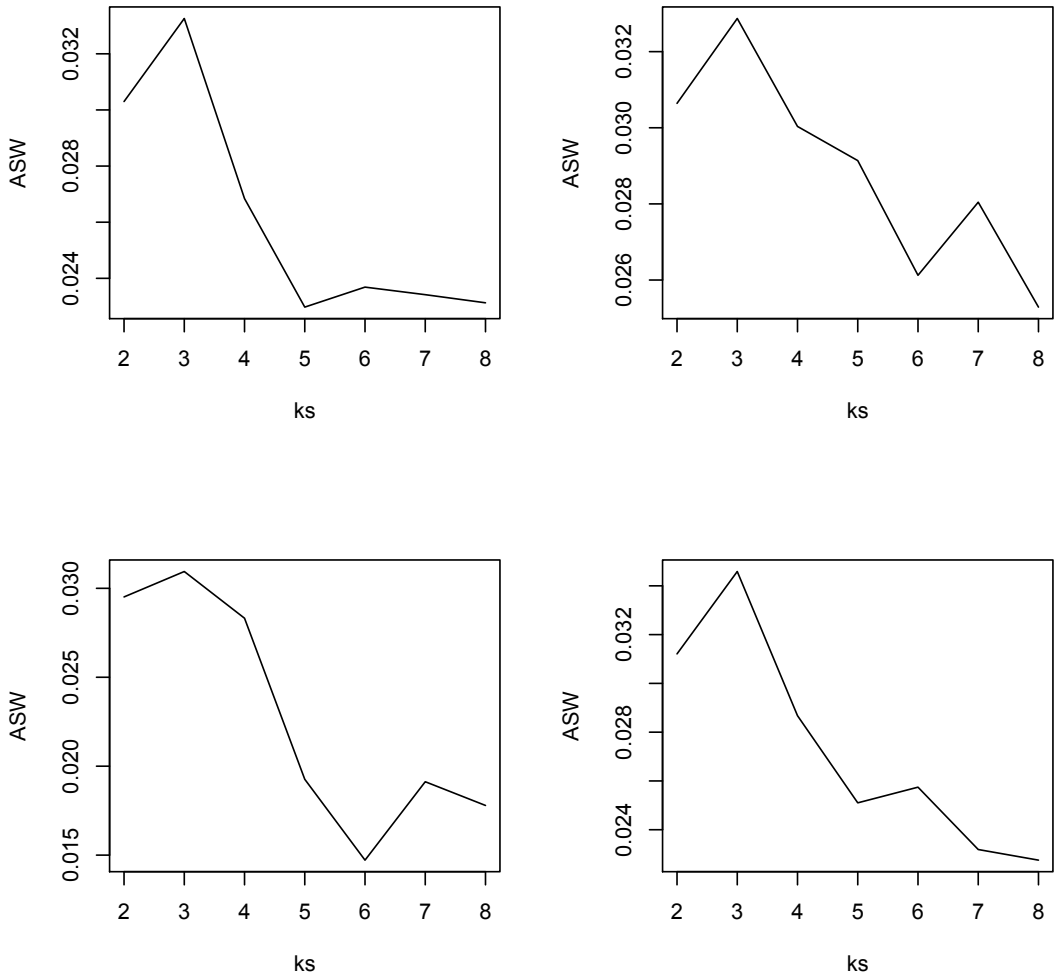

Table 1

	PC1	PC2	PC3	PC4	PC5	PC6
SD	3.02545	2.75745	2.57682	2.38151	2.2235	2.16336
Proportion of Var	0.01956	0.01625	0.01419	0.01212	0.01056	0.01
Cumulative Proportion	0.01956	0.03581	0.04999	0.06211	0.07268	0.08268

Figure 8

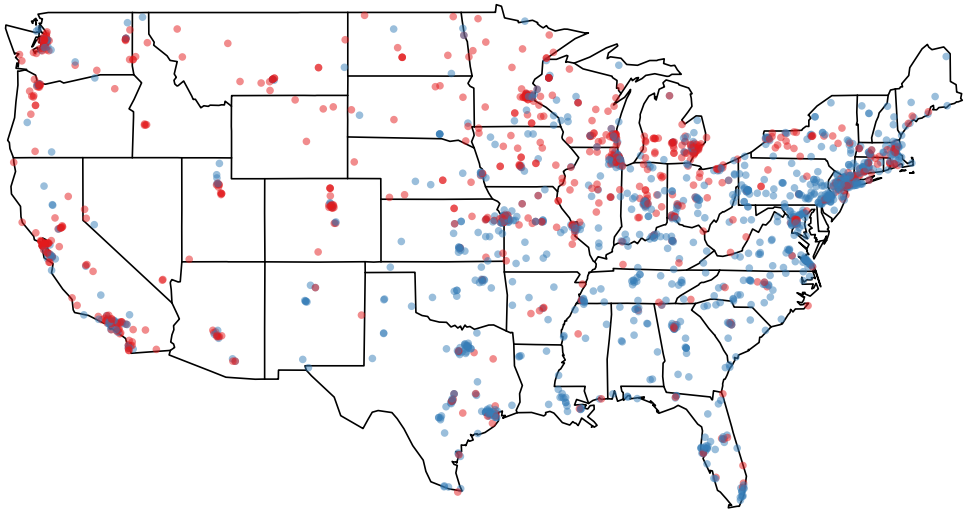

Figure 9

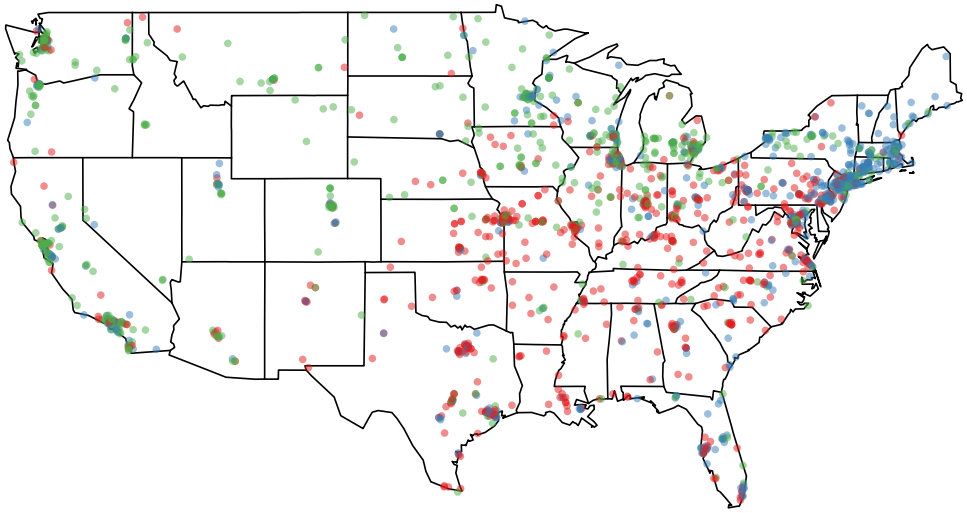

Figure 10

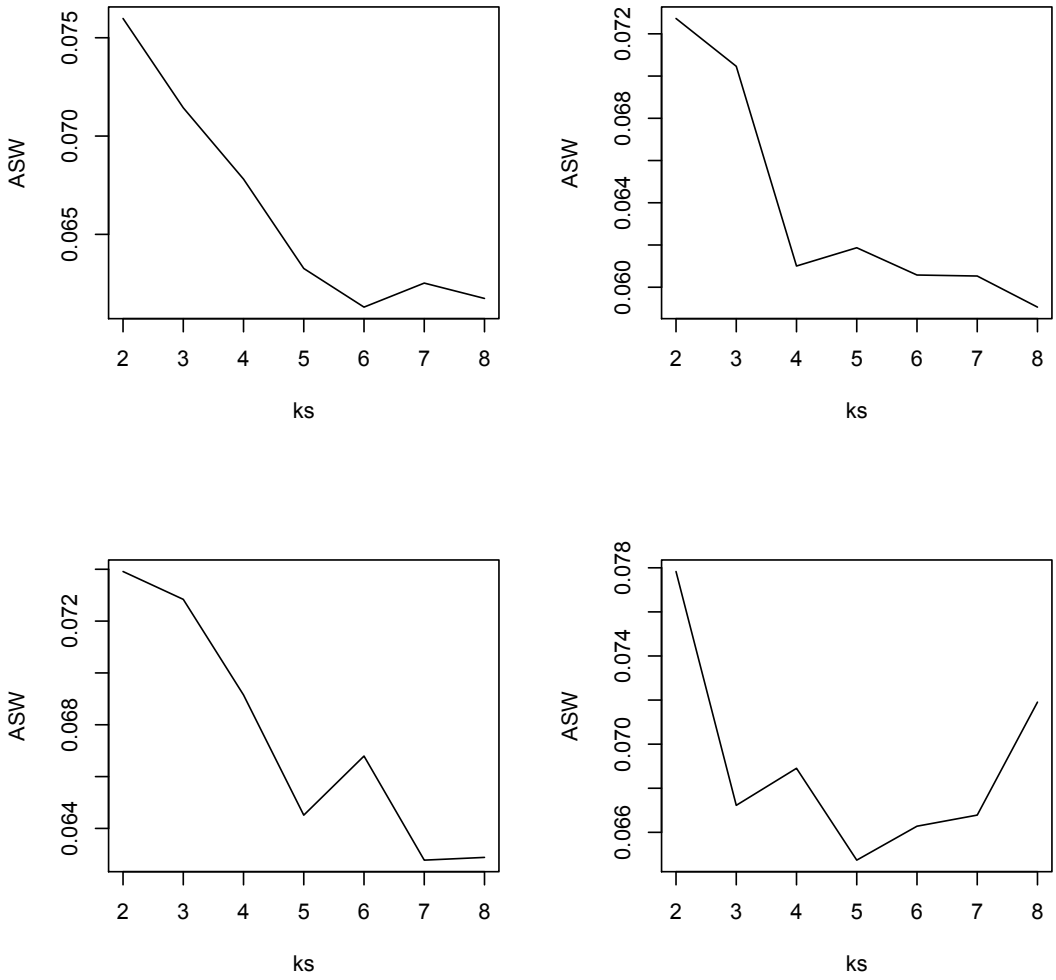
ASW
0.065
0.070
0.075
2
3
4
5
6
7
8
ks
ASW
0.060
0.064
0.068
0.072
2
3
4
5
6
7
8
ks
ASW
0.064
0.068
0.072
2
3
4
5
6
7
8
ks
ASW
0.066
0.070
0.074
0.078
2
3
4
5
6
7
8
ks

Figure 11

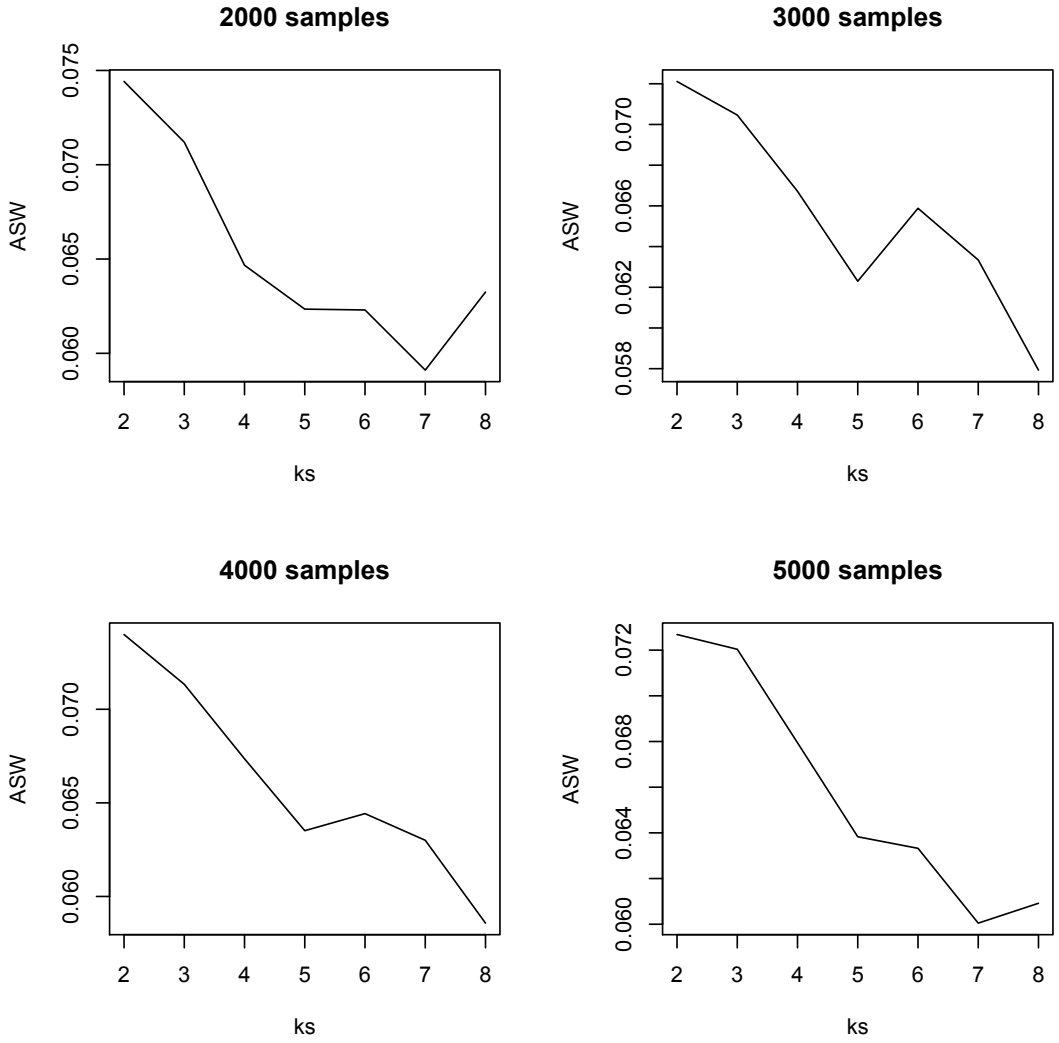
2000 samples
3000 samples
4000 samples
5000 samples
ASW
ks
0.060
0.065
0.070
0.075
0.058
0.062
0.066
0.064
0.068
0.072
2
3
4
5
6
7
8

Figure 12

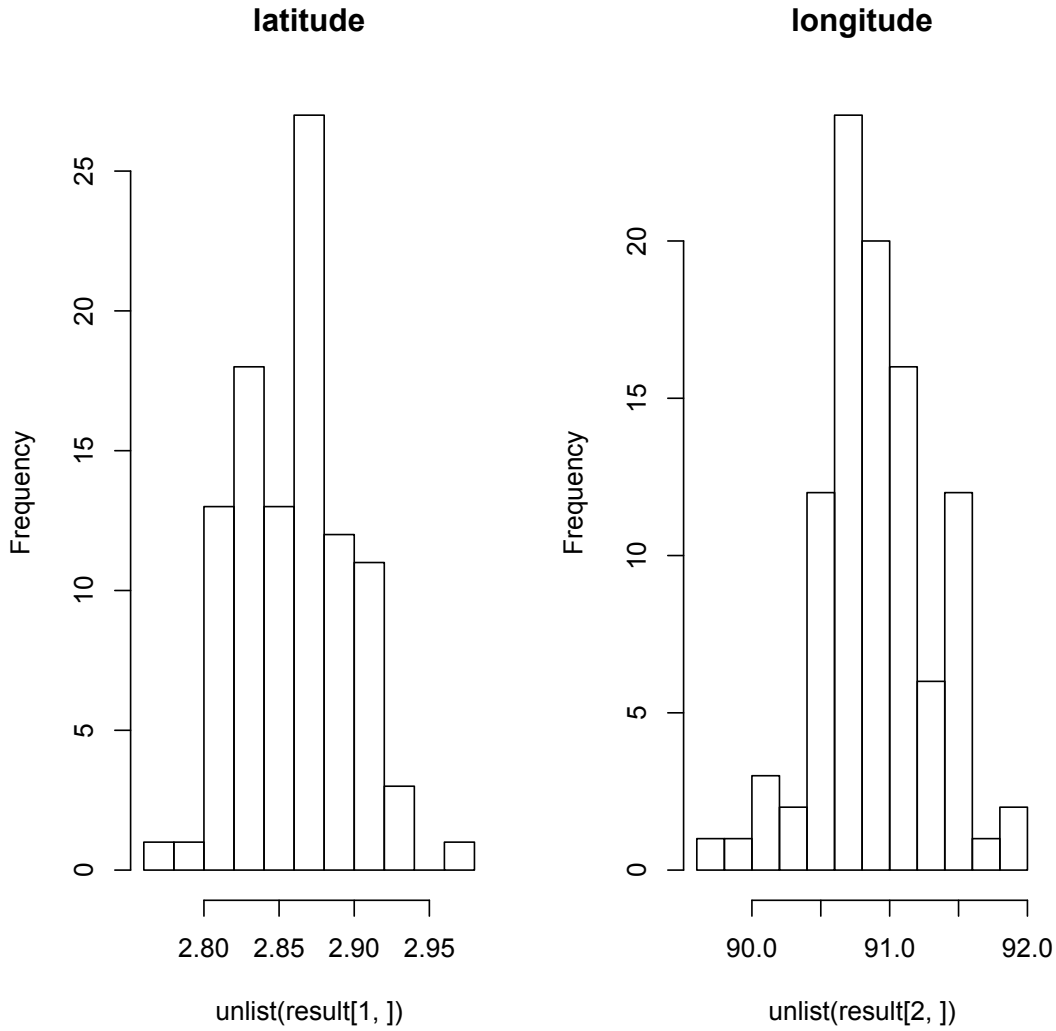

Table 2

Given 66		66						
		1	2	3	4	5	6	7
65	1	0.345895094	0.262434865	0.13157895	0.26923077	0.444940061	0.46938776	0.303086997
	2	0.248044632	0.337754619	0.57894737	0.15384615	0.229346715	0.14285714	0.342843779
	3	0.401575234	0.39428391	0.05263158	0.38461538	0.318093114	0.2244898	0.329279701
	4	0.000273478	0.00102637	0.05263158	0.03846154	0.001300994	0.10204082	0.000935454
	5	0.001804955	0.002763303	0.10526316	0.11538462	0.003995911	0.04081633	0.021047708
	6	0.002406607	0.001736934	0.07894737	0.03846154	0.002323204	0.02040816	0.002806361