Homework #7

Jonathan Beaubien
Math 381 - Discrete Modeling
University of Washington

March 5th, 2021

Multidimensional Scaling

In this write-up I will analyze two sets of cities by using multidimensional scaling.
For our first set of cities, we will choose cities within Washington State so that the distances between each city is relatively small and euclidean in distance.
I have chosen the cities Seattle, Renton, Kent, Issaquah, Federal Way, and Tacoma.
I have retrieved the distances between each city by using Google Maps and picking the shortest route between each pair.
I made this distance matrix (where the distances are in kilometers):

$$A = \begin{bmatrix} 0.00000 & 19.63400 & 30.09473 & 27.68072 & 35.56650 & 54.07396 \\ 19.63400 & 0.00000 & 13.84040 & 19.31210 & 23.97923 & 41.36014 \\ 30.09473 & 13.84040 & 0.00000 & 36.04931 & 11.58730 & 29.77286 \\ 27.68072 & 19.31210 & 36.04931 & 0.00000 & 49.08499 & 60.67227 \\ 35.56650 & 23.97923 & 11.58730 & 49.08499 & 0.00000 & 19.95587 \\ 54.07396 & 41.36014 & 29.77286 & 60.67227 & 19.95587 & 0.00000 \end{bmatrix}$$

$$B = \{\text{Seattle, Renton, Kent, Issaquah, Federal Way, Tacoma}\}$$

The ij^{th} entry of matrix A is the distance from city B_i to city B_j. Notice that $A_{ij} = A_{ji}$ because the distance doesn't depend on the starting/ending city. Also, the entries $A_{ii} = 0$ because the distance from a city to itself is 0 km.
I used the programming language R to do multidimensional scaling of this distance matrix using 1-3 dimensions. 2 dimensions should resemble the true euclidean distance and thus look similar to a map of the cities.
Here is a graph of the 1 and 2 dimensional models using the command cmdscale():

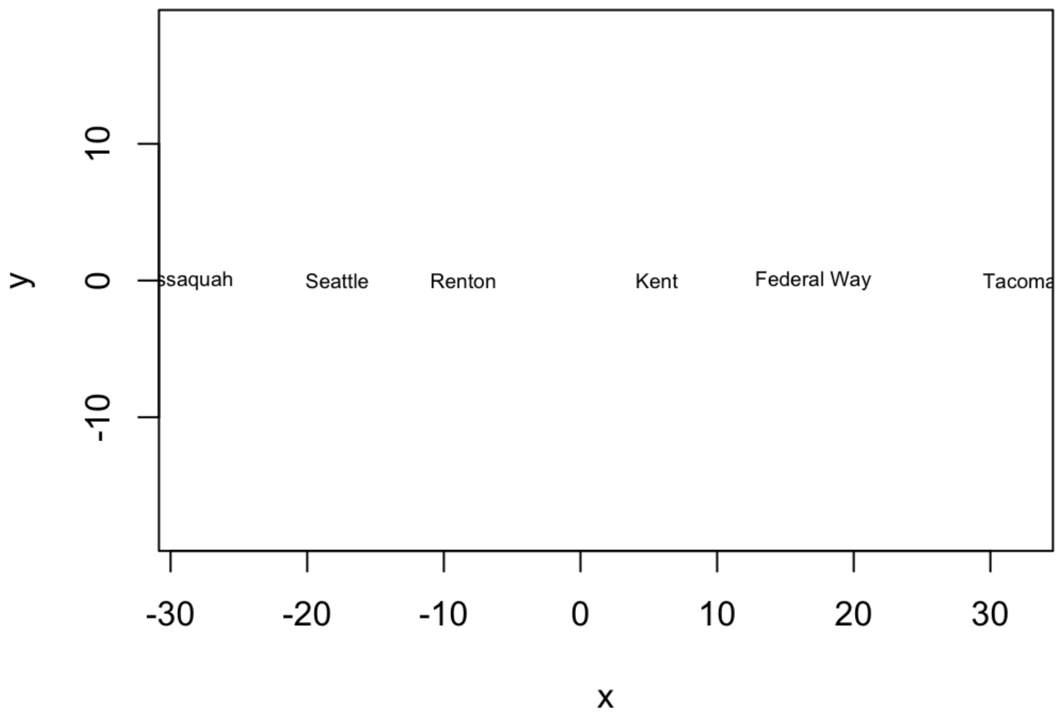

Figure 1: 1 Dimensional Model

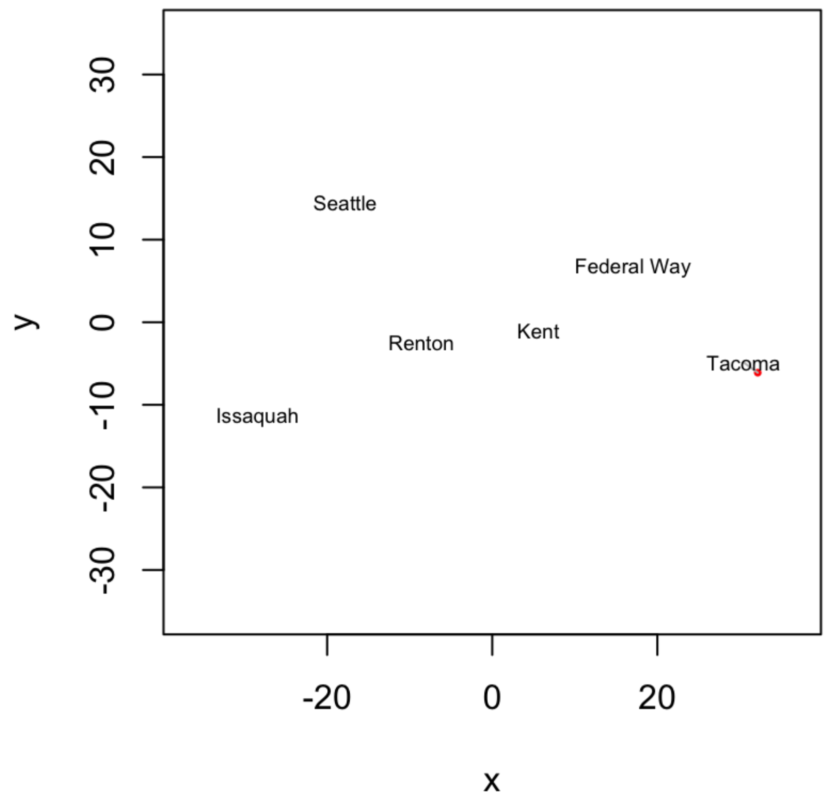

Figure 2: 2 Dimensional Model

The model using 3 dimensions is difficult to graph and is thus left out.
We can compare this to a real map of the cities. In comparison, it looks like the 2-d model is definitely not perfect. At first I though the x-axis may be north-south (low to high) and y-axis was east-west (high to low) yet this seems to not fit with Issaquah's placement on the model coordinates.
This is included in figure 3.

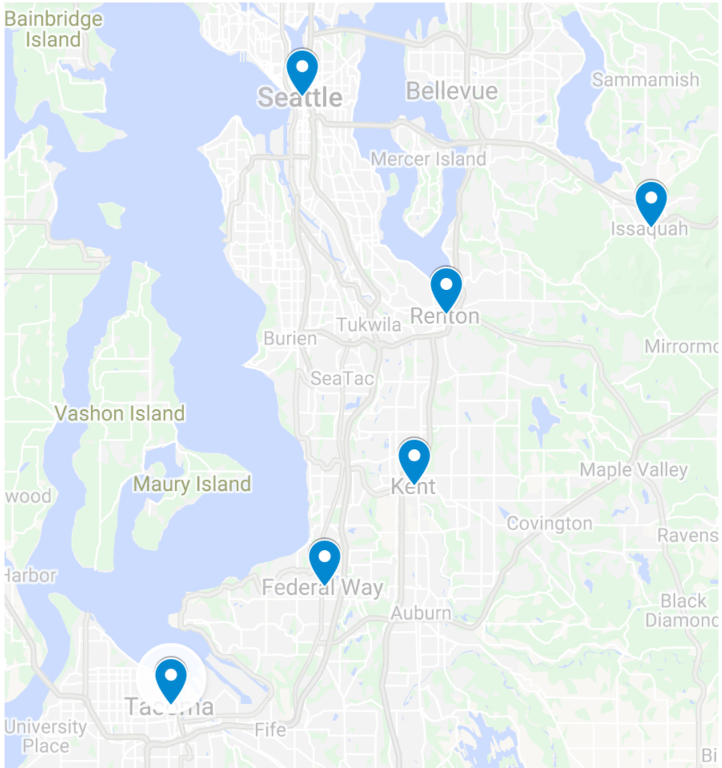

Figure 3: True Locations of Cities

We can now compare the goodness of fit for each model. Here is a table comparing the goodness of fits:

Model Dimension	GOF Value
1	0.7956923
2	0.9301851
3	0.9648809

It looks like there is diminishing returns after a 2 dimensional model (which makes sense because our real life model is 2 dimensional).
We can see this also in the graph of the eigenvalues of the matrix A.
This is included in figure 4.

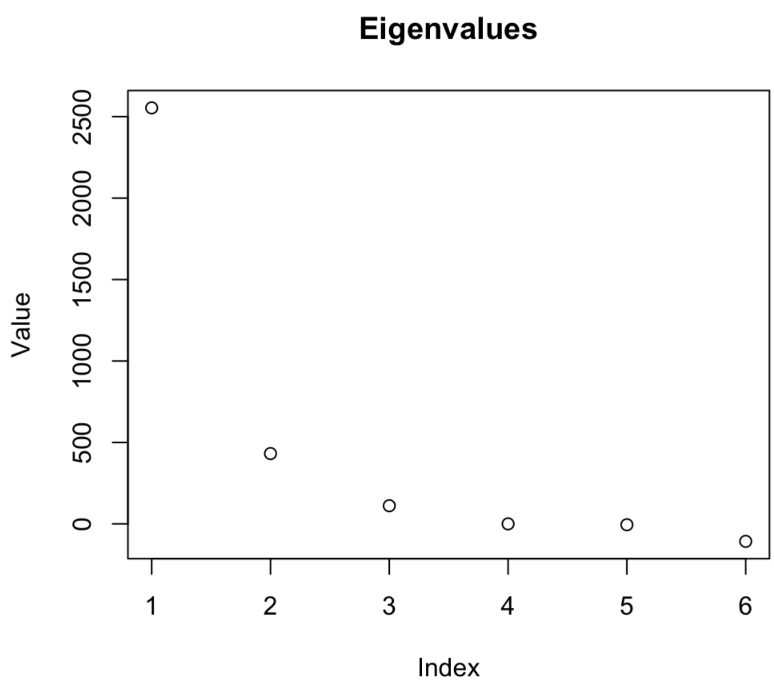

Figure 4: Eigenvalues

Another tool for analysis we have is comparing the distances between cities the 2-d model came up with to the original ones.
By plotting these two things against each other, along with a line of $y = x$ to see how closely the models align, can help us see the accuracy of each model.
This is included in figure 5.

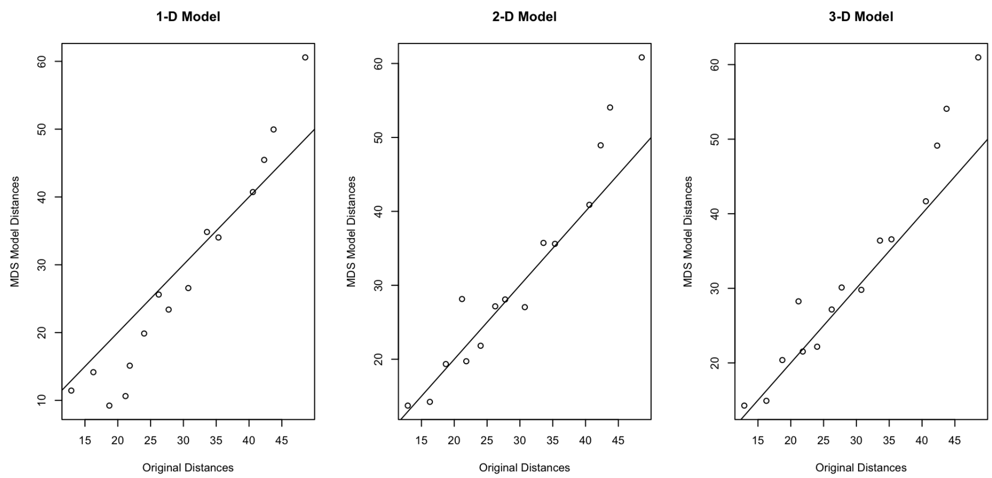

Figure 5: Model Distance Comparisons

The biggest difference in the 2-d model to the orginal distances is the distance from Issaquah to Tacoma.
This distance is more than 12 km less than the true distance.
Here is the matrix of distance differences:

```
> print(C)
                 Seattle      Renton        Kent    Issaquah Federal Way
Renton        -0.6343566
Kent          -0.3359128   2.0481268
Issaquah      -6.9645662   2.1893414  -0.2672696
Federal Way   -2.1430224  -0.9183186  -0.8124168  -6.6302796
Tacoma       -10.3119177  -0.3015089   3.6998042 -12.2649460   2.0990849
> 
```

Figure 6: Model Distance Difference Matrix

The mean absolute value difference for all pairs of cities is 3.441392 km.
Here is a histogram of these differences:

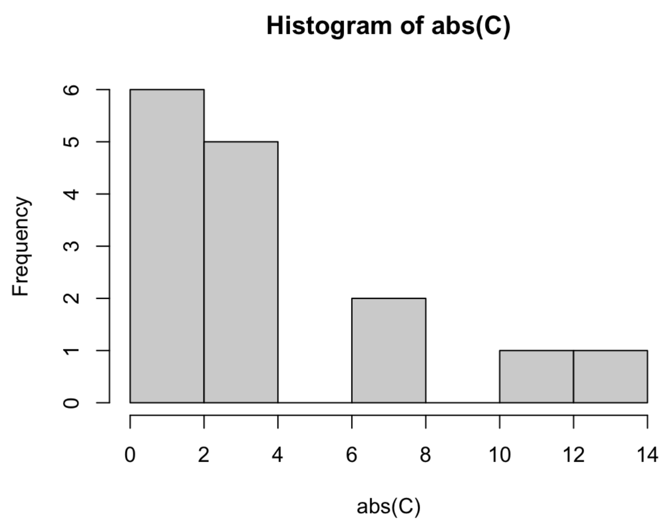

Figure 7: Histogram of Absolute Value of Differences

Most of the differences are less than 4 km with a skew to the right where there are a couple outliers above 10 km.
This is the code used to create the plots and analysis.

```r
library(wordcloud)
rownames = c("Seattle", "Renton", "Kent", "Issaquah", "Federal Way", "Tacoma")
colnames = c("Seattle", "Renton", "Kent", "Issaquah", "Federal Way", "Tacoma")
A <- matrix(
  c(0, 19.634, 30.09473, 27.68072, 35.5665, 54.07396,
    19.634, 0, 13.8404, 19.3121, 23.97923, 41.36014,
    30.09473, 13.8404, 0, 36.04931, 11.5873, 29.77286,
    27.68072, 19.3121, 36.04931, 0, 49.08499, 60.67227,
    35.5665, 23.97923, 11.5873, 49.08499, 0, 19.95587,
    54.07396, 41.36014, 29.77286, 60.67227, 19.95587, 0),
  nrow = 6, byrow = TRUE, dimnames = list(rownames, colnames))
print(A)
model1d <- cmdscale(A, k = 1, eig=TRUE)
model2d <- cmdscale(A, k = 2, eig=TRUE)
model3d <- cmdscale(A, k = 3, eig=TRUE)
print(model1d$GOF)
print(model2d$GOF)
print(model3d$GOF)

plot(model1d$eig, main="Eigenvalues", ylab="Value")

plot(model2d,asp = 1,xlim=c(-35, 35), ylim=c(-35,35),ylab="",xlab="")
textplot(model1d[,1],c(0,0,0,0,0,0),gsub("( [ \\s]+\\s{1})","",rownames(A),perl=TRUE),asp=1,cex=0.6)

par(mfrow=c(1,3))
print(dist(model2d))
print(dist(A)/2)
plot(dist(A)/2, asp=1,dist(model1d), xlab="Original Distances", ylab="MDS Model Distances", main="1-D Model")
abline(0, 1)
plot(dist(A)/2, asp=1,dist(model2d), xlab="Original Distances", ylab="MDS Model Distances", main="2-D Model")
abline(0, 1)
plot(dist(A)/2, asp=1,dist(model3d), xlab="Original Distances", ylab="MDS Model Distances", main="3-D Model")
abline(0, 1)

C <- dist(A)/2 - dist(model2d)
print(C)
print(mean(abs(C)))
hist(abs(C))
```