MATH 381 Project 2 Autumn 2016

Multidimensional Scaling of the species of National Parks

Group 2

Carrie Ann Beede, Daehyun Kim, Samantha Leigh Larson, Chengjun Zhang

University of Washington

December 1, 2016

Introduction

We will be modeling National Parks in the United States based upon the biological species found in those parks. The United States National Park Service was formed in 1916 by President Woodrow Wilson to protect areas of natural or historical significance. Today, these parks are locations of many natural superlatives such as the worlds largest carnivore (Alaskan brown bear), the lowest point in the Western Hemisphere (Badwater Basin, California, 282 ft below sea level), and America′s deepest lake (Crater Lake, Oregon). These national parks are also valuable preservation sites for hundreds of threatened or endangered plant and animal species.

In this project we will look at the species located in each park. There are 58 national parks, and we are able to collect data for the species in 56 of these. A list of those 56 parks can be found in Appendix. For any two parks utilized in our model, we will calculate a distance between them based upon the number of species they have in common, and create a distance matrix with these distances. Then, using multidimensional scaling, we will create coordinate points and a visual representation of these 56 parks. Our model will be an analysis of this result based upon how we can rotate this visual representation and label our axes. Additionally, we can investigate the goodness of fit and try other dimensional models for this data, exploring the possibilities for multidimensional scaling in R.

Background

The inspiration for this project came from a lifelong interest in wildlife and natural preservation that has led one of our group members to visit and study many different national parks and to discover the astounding variety between them. Olympic National Park, which is on the peninsula of our very own Washington State, is a wealth of natural beauty containing ocean beaches, temperate rainforest, and mountain glaciers. By stark contrast, Death Valley National Park in southern California is a desert of salt flats, borax mines, and extreme heat. How many animal species do these two parks have in common? How many animals that thrive in an arid climate of under 2 inches of rainfall per year in Death Valley can also be found in Olympic National Park, which in the Hoh and Quinault rainforests receives 150 inches of rainfall every year? How "far apart" are these two parks? Is rainfall a possible axes for a two dimensional model of the national parks? We hope that by analysing this model of natural locations all over the United States we can learn more about our natural world as well as the mathematical concepts we will be studying and utilizing in this project.

The primary mathematical tool we will be using for this project is multidimensional scaling, a method for information visualization, orientation, and analysis based on similarities between data points. MDS takes in a "matrix of dissimilarities" and gives us a visual representation of these dissimilarities as distances. Usually this must be done in low dimensions; visualization becomes very difficult after the third dimension. Versions of MDS can be seen in mathematical work as far back as the 17th century, used by cartographers and scientists solving problems with scaling. The idea that similarity can be represented as distance is not a new concept, after all. Psychologically, we see things that are "far apart" as dissimilar and things that are "close together" as similar, evidenced by individual human relationships, cultures, biological diversity, and many other practical examples. How to calculate distance to represent similarity and how to best model this in low-dimensions have been problems continually addressed and improved upon in the study and utilization of MDS [4].

Multidimensional scaling has been used to understand similarities between ecosystems in recent history. One example of this can be found in a 2010 annual report by the U.S. Department of the Interior, looking in particular at Arches National Park. They divided the park into ecologically differing areas with three different categorizations (deep blackbrush, pinyon-juniper blackbrush, and grassland) and, on page 17 of their report, they have the visual two dimensional graph showing how grasslands tend to be high on their y-axis while blackbrush is found closer to the origin [5]. The Department of the Interior has similar papers using similar techniques for several different natural areas and ecosystems. Academic studies of ecology have also utilized multidimensional scaling on ecological problems [6]. These studies generally explored the problem on a smaller scale that we have attempted here, both in terms of geography and number of species.

Model

We were able to get data for the species in 56 national parks from the National Park Service website [8]. For each park we downloaded a database which contained, among a lot of other information, taxon codes and category (mammal, bird, amphibian, etc) for every living species found in that park. The data sheet of one park, Acadia National Park, is attached below in Figure 1. Each unique species has a unique taxon code, so by comparing these codes for every pair of parks we obtained the number of species in common for those two parks. For our two-dimensional model, we want an inverse relationship between number of species in common and ecological distance between any two parks.

FIGURE 1. Extraction of data for Acadia National Park

	Park Code	Park Name	Category	Category Sort	Order	Family	Taxon Code	TSN	Taxon Record Status	Scientific Name	Common Names	Synonyms	Occurrence
2	ACAD	Acadia National Park	Mammal	1	Artiodacty	Cervidae	95052	180703	Active	Alces alces	Moose		Present
3	ACAD	Acadia National Park	Mammal	1	Artiodacty	Cervidae	95046	180699	Active	Odocoileus virginiar	Northern White		Present
4	ACAD	Acadia National Park	Mammal	1	Carnivora	Canidae	93340	180599	Active	Canis latrans	Coyote, Eastern		Present
5	ACAD	Acadia National Park	Mammal	1	Carnivora	Canidae	93347	180604	Active	Vulpes vulpes	Black Fox, Cross		Present
6	ACAD	Acadia National Park	Mammal	1	Carnivora	Mustelidae	93311	180572	Active	Lutra canadensis	Otter, River Otte		Present
7	ACAD	Acadia National Park	Mammal	1	Carnivora	Mustelidae	93298	180560	Active	Martes pennanti	Blackcat, Fisher,		Present
8	ACAD	Acadia National Park	Mammal	1	Carnivora	Mustelidae	93292	180555	Active	Mustela erminea	Bonaparte Weas		Present
9	ACAD	Acadia National Park	Mammal	1	Carnivora	Mustelidae	93293	180556	Active	Mustela frenata	Long-tailed Wea		Present
10	ACAD	Acadia National Park	Mammal	1	Carnivora	Mustelidae	93290	180553	Active	Mustela vison	American Mink,		Present
11	ACAD	Acadia National Park	Mammal	1	Carnivora	Procyonid	93314	180575	Active	Procyon lotor	Common Racco		Present
12	ACAD	Acadia National Park	Mammal	1	Carnivora	Ursidae	93280	180544	Active	Ursus americanus	Black Bear		Present
13	ACAD	Acadia National Park	Mammal	1	Chiroptera	Vespertilio	88730	180008	Active	Eptesicus fuscus	Big Brown Bat, (		Present
14	ACAD	Acadia National Park	Mammal	1	Chiroptera	Vespertilio	88737	180014	Active	Lasionycteris noctiv	Silver-haired Ba		Present
15	ACAD	Acadia National Park	Mammal	1	Chiroptera	Vespertilio	88739	180016	Active	Lasiurus borealis	Eastern Red Bat		Present
16	ACAD	Acadia National Park	Mammal	1	Chiroptera	Vespertilio	88740	180017	Active	Lasiurus cinereus	Hoary Bat		Present
17	ACAD	Acadia National Park	Mammal	1	Chiroptera	Vespertilio	88707	179989	Active	Myotis keenii	Keen's Myotis, I		Present
18	ACAD	Acadia National Park	Mammal	1	Chiroptera	Vespertilio	88706	179988	Active	Myotis lucifugus	Little Brown Bat		Present
19	ACAD	Acadia National Park	Mammal	1	Lagomorp	Leporidae	89638	180112	Active	Lepus americanus	Snowshoe Hare		Present
20	ACAD	Acadia National Park	Mammal	1	Rodentia	Castoridae	90673	180212	Active	Castor canadensis	American Beave		Present
21	ACAD	Acadia National Park	Mammal	1	Rodentia	Cricetidae	90763	180294	Active	Clethrionomys gapp	Boreal Redback		Present
22	ACAD	Acadia National Park	Mammal	1	Rodentia	Cricetidae	90745	180278	Active	Peromyscus leucopı	White-footed M		Present
23	ACAD	Acadia National Park	Mammal	1	Rodentia	Cricetidae	90743	180276	Active	Peromyscus manicu	Deer Mouse		Present
24	ACAD	Acadia National Park	Mammal	1	Rodentia	Cricetidae	91716	180324	Active	Synaptomys cooper	Cooper Lemmir		Present

We make the following definitions of variables.

$$P = \text{The set of all national parks } \{1, 2, 3, \ldots, 56\}.$$
$$n_i = \text{The number of species that exist in park } i. \ (i \in P)$$
$$c_{i,j} = \text{The number of common species between park } i \text{ and park } j. \ (i, j \in P)$$
$$d_{i,j} = \text{The ecological distance between park } i \text{ and park } j. \ (i, j \in P)$$

Then, we define the ecological distance between park i and park j as the following.

$$d_{i,j} = \left(1 - \frac{2c_{i,j}}{n_i + n_j}\right)^4 , \forall i, j \in P \tag{1}$$

The ecological distance is a measure of how two national parks are similar to each other based on the number of common species they have. The intuition of this distance formula is that $\dfrac{2c_{i,j}}{n_i + n_j} = \dfrac{c_{i,j}}{\frac{1}{2}(n_i + n_j)}$ which is the number of common species between park i and park j divided by the average total number of species in both parks. This can be a measure of

how two parks are similar to each other. Then $\left(1 - \frac{2c_{i,j}}{n_i + n_j}\right)$ is a measure of how different two parks are. Raising this measure to the power of 4 reduced the scale of distance if the difference measure is very close to 0 (two very similar parks); meanwhile, the distance between two very different parks are not effected by much. The reason why power of 4 is chosen is because it has a better GOF for the MDS output than powers of 2 or 3 as shown in Figure 13 in the Appendix, and the MDS will looks clustered after power of 5 which makes it difficult to interpret.
When $i = j$, the ecological distance between the park i and itself j is 0 because $2c_{i,j} = n_i + n_j$. When $i \neq j$ and $c_{i,j} = 0$, that is when park i and park j has no common species, the ecological distance between them is 1 because $2c_{i,j} = 0$. When $i \neq j$ and $c_{i,j} \neq 0$, that is when park i and park j have some common species, the ecological distance between them is between 0 and 1 because $0 < 2c_{i,j} < n_i + n_j$. Based on the formula of ecological distance, the maximum distance between two parks is 1, and the minimum distance between two parks is 0.

Software Implementation:

(1) MATLAB import `xlsx` file one by one, and the name of the park is stored in `name`, the taxon codes for all species in this park will be stored in `spec`, and all the taxon codes for all parks will be stored in the corresponding line of the matrix `total_spec`. The number of total species in each park is stored in `num_spec`.

(2) Then `dist` is initialized to be a 56 by 56 matrix with all 0 entries. We will iterate thought all entries for every two rows of the matrix `total_spec`, and add 1 to the corresponding entry in `dist` once a common taxon code is found. Therefore, the entry of `dist` of row j and column k is now the number of common species between park j and park k.

(3) Now we calculate the ecological distance between each two parks using the formula of ecological distance. The entry in `dist` will be replaced by the calculated distance. Therefore, the entry of `dist` of row j and column k is now the ecological distance between park j and park k.

(4) Finally, we save the `dist` matrix to a `dat` file for MDS in R.

The following is the MATLAB Code for data processing.

```
clear all
close all

%% Import Data
n = 56;
```

```
tic
for j = 1:n
    filename = sprintf('park%d.xlsx',j);
    A = importdata(filename);
    name(1,j) = A(2,1);
    spec = A(2:end, 7)';
    total_spec(j,1:length(spec)) = str2double(spec);
    dimensions = size(spec);
    num_spec(j) = dimensions(2);
end

%% Distance Matrix
dist = zeros(56);
for j = 1:n
    for k = 1:n
        for l = 1:length(total_spec(k,:))
            if total_spec(k,l) == 0
                l = length(total_spec(k,:));
            elseif any(total_spec(j,:) == total_spec(k,l))
                dist(j,k) = dist(j,k) + 1;
            end
        end
    end
end

%% Calaulation of distances
for j = 1:n
    for k = 1:n
        dist(j,k) = (1-(2*dist(j,k))/(num_spec(j)+num_spec(k)))^4;
    end
end
toc

%% Saving files
save('dist_mat.dat', 'dist','-ASCII');
dlmwrite('dist_mattxt.txt',dist,'delimiter','\t','precision',10);
name = name';
dlmwrite('names.txt',name,'delimiter','');
```

Note: This program took 189.890708 seconds to run on the computer we used. The information of the computer can be found in the Appendix.

Figure 2 is an example of the distance matrix generated by MATLAB. The figure only shows the first 25 rows and columns of the matrix. The actual matrix is 56 by 56.

FIGURE 2. Example of the Distance Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0	0.5432	0.3805	0.5464	0.6203	0.5471	0.5216	0.5627	0.5022	0.4942	0.5839	0.3407	0.4704	0.1204	0.3869	0.6038	0.4915	0.5081	0.4885	0.3048	0.4195	0.5894	0.5564	0.4275	0.2127
2	0.5432	0	0.2256	0.3688	0.7491	0.1883	0.0730	0.0059	0.0209	0.2657	0.5104	0.6254	0.4039	0.5298	0.6614	0.3137	0.6260	0.6698	0.6969	0.4389	0.6249	0.2084	0.2539	0.1898	0.6315
3	0.3805	0.2256	0	0.3566	0.6267	0.2405	0.2302	0.2277	0.2235	0.2381	0.4843	0.4769	0.3945	0.3425	0.5434	0.3947	0.4476	0.5180	0.6116	0.2846	0.5302	0.3050	0.3332	0.1250	0.4574
4	0.5464	0.3688	0.3566	0	0.5735	0.4564	0.4241	0.3409	0.3433	0.0363	0.4637	0.5051	0.5684	0.5020	0.7301	0.3221	0.4873	0.4597	0.7880	0.5410	0.6771	0.5259	0.2277	0.3727	0.5505
5	0.6203	0.7491	0.6267	0.5735	0	0.7876	0.7364	0.7142	0.6992	0.5923	0.6491	0.5325	0.7754	0.6226	0.8122	0.6602	0.0482	0.0393	0.8351	0.7416	0.7522	0.8134	0.7242	0.6994	0.6462
6	0.5471	0.1883	0.2405	0.4564	0.7876	0	0.1424	0.1890	0.1707	0.3385	0.5778	0.6950	0.3475	0.5733	0.6007	0.3607	0.6893	0.7188	0.6912	0.3265	0.5844	0.1793	0.2846	0.1383	0.6692
7	0.5216	0.0730	0.2302	0.4241	0.7364	0.1424	0	0.0573	0.0329	0.3269	0.5309	0.6962	0.3276	0.5568	0.5584	0.3635	0.6173	0.6677	0.6371	0.2955	0.5302	0.1258	0.2628	0.1366	0.6629
8	0.5627	0.0059	0.2277	0.3409	0.7142	0.1890	0.0573	0	0.0153	0.2540	0.5107	0.6486	0.4008	0.5445	0.6443	0.2939	0.6195	0.6500	0.7114	0.4159	0.6097	0.2164	0.2101	0.1917	0.6348
9	0.5022	0.0209	0.2235	0.3433	0.6992	0.1707	0.0329	0.0153	0	0.2407	0.4763	0.6244	0.3857	0.4777	0.5853	0.2765	0.5947	0.6259	0.6781	0.3254	0.5584	0.1645	0.1864	0.1679	0.5781
10	0.4942	0.2657	0.2381	0.0363	0.5923	0.3385	0.3269	0.2540	0.2407	0	0.4314	0.4982	0.4956	0.4606	0.6981	0.2994	0.4429	0.4818	0.7635	0.4769	0.6375	0.4084	0.1727	0.2496	0.5219
11	0.5839	0.5104	0.4843	0.4637	0.6491	0.5778	0.5309	0.5107	0.4763	0.4314	0	0.6477	0.5550	0.5840	0.6876	0.3734	0.5424	0.5657	0.7384	0.5344	0.4283	0.5916	0.4586	0.4772	0.6530
12	0.3407	0.6254	0.4769	0.5051	0.5325	0.6950	0.6962	0.6486	0.6244	0.4982	0.6477	0	0.6916	0.2011	0.7977	0.6613	0.5015	0.3519	0.8001	0.6473	0.7204	0.7173	0.6592	0.6128	0.1290
13	0.4704	0.4039	0.3945	0.5684	0.7754	0.3475	0.3276	0.4008	0.3857	0.4956	0.5550	0.6916	0	0.5674	0.4280	0.5120	0.6779	0.7058	0.5757	0.2080	0.3710	0.2903	0.4720	0.2693	0.6403
14	0.1204	0.5298	0.3425	0.5020	0.6226	0.5733	0.5568	0.5445	0.4777	0.4606	0.5840	0.2011	0.5674	0	0.5836	0.5798	0.5261	0.4931	0.6354	0.4065	0.5788	0.6332	0.5176	0.4676	0.0773
15	0.3869	0.6614	0.5434	0.7301	0.8122	0.6007	0.5584	0.6443	0.5853	0.6981	0.6876	0.7977	0.4280	0.5836	0	0.6633	0.7176	0.7261	0.0198	0.1681	0.0887	0.5331	0.6216	0.3887	0.7057
16	0.6038	0.3137	0.3947	0.3221	0.6602	0.3607	0.3635	0.2939	0.2765	0.2994	0.3734	0.6613	0.5120	0.5798	0.6633	0	0.5378	0.5576	0.7376	0.4894	0.5923	0.3355	0.1671	0.3572	0.6699
17	0.4915	0.6260	0.4476	0.4873	0.0482	0.6893	0.6173	0.6195	0.5947	0.4429	0.5424	0.5015	0.6779	0.5261	0.7176	0.5378	0	0.1394	0.7211	0.6496	0.6432	0.7354	0.6679	0.5375	0.6131
18	0.5081	0.6698	0.5180	0.4597	0.0393	0.7188	0.6677	0.6500	0.6259	0.4818	0.5657	0.3519	0.7058	0.4931	0.7261	0.5576	0.1394	0	0.7569	0.6504	0.6700	0.7615	0.6355	0.5973	0.5052
19	0.4885	0.6969	0.6116	0.7880	0.8351	0.6912	0.6371	0.7114	0.6781	0.7635	0.7384	0.8001	0.5757	0.6354	0.0198	0.7376	0.7211	0.7569	0	0.2943	0.2077	0.6369	0.7377	0.5237	0.7674
20	0.3048	0.4389	0.2846	0.5410	0.7416	0.3265	0.2955	0.4159	0.3254	0.4769	0.5344	0.6473	0.2080	0.4065	0.1681	0.4894	0.6496	0.6504	0.2943	0	0.2115	0.2663	0.3521	0.1930	0.5117
21	0.4195	0.6249	0.5302	0.6771	0.7522	0.5844	0.5302	0.6097	0.5584	0.6375	0.4283	0.7204	0.3710	0.5788	0.0887	0.5923	0.6432	0.6700	0.2077	0.2115	0	0.5681	0.5880	0.4146	0.6654
22	0.5894	0.2084	0.3050	0.5259	0.8134	0.1793	0.1258	0.2164	0.1645	0.4084	0.5916	0.7173	0.2903	0.6332	0.5331	0.3355	0.7354	0.7615	0.6369	0.2663	0.5681	0	0.2405	0.1655	0.6921
23	0.5564	0.2539	0.3332	0.2277	0.7242	0.2846	0.2628	0.2101	0.1864	0.1727	0.4586	0.6592	0.4720	0.5176	0.6216	0.1671	0.6679	0.6355	0.7377	0.3521	0.5880	0.2405	0	0.2415	0.5886
24	0.4275	0.1898	0.1250	0.3727	0.6994	0.1383	0.1366	0.1917	0.1679	0.2496	0.4772	0.6128	0.2693	0.4676	0.3887	0.3572	0.5375	0.5973	0.5237	0.1930	0.4146	0.1655	0.2415	0	0.5911
25	0.2127	0.6315	0.4574	0.5505	0.6462	0.6692	0.6629	0.6348	0.5781	0.5219	0.6530	0.1290	0.6403	0.0773	0.7057	0.6699	0.6131	0.5052	0.7674	0.5117	0.6654	0.6921	0.5886	0.5911	0

This distance matrix is converted into a `dat` file, and then fed to R to produce MDS. R code for MDS is on the following page.

We now run the MDS in R using the data file generated from MATLAB.

(1) Set the directory of the `dat` file containing the distance matrix generated from MATLAB.
(2) Assign `distance` to be the distance matrix.
(3) Assign `names` to be a previously generated file with the abbreviated and full names of each national park.
(4) Use the `cmdscale` command to create a list of 2-dimensional coordinates (`k=2` determines the dimension).
(5) The angle of rotation `theta` is set in radius counter-clockwise, then we multiply the coordinates matrix with the rotation matrix `rotation`.
(6) Plot the MDS graph with the rotated coordinates.
(7) Plot the MDS graph with the names of the national parks. (`ll[,1],ll[,2]` refer to x and y coordinates of the model)
(8) Calculate the goodness of fit (GOF) of the model. The GOF is a value between 0 and 1 and the model is a better fit if its GOF is closer to 1.
(9) Generate a random matrix of the same size `n` by `n`, apply MDS, and calculate its GOF.

The following is the R Code for multidimensional scaling.

```
setwd('C:/Users/FILE_DIRECTORY')
distance <- read.table("dist_mat.dat")
theta = 0 # rotation angle in radius, counter-clockwise
rotation <- matrix(c(cos(theta), -sin(theta), sin(theta),
                  cos(theta)), nrow = 2, ncol = 2, byrow = TRUE)
                  # the rotation matrix
ll <- cmdscale(distances, k=2)
ll <- ll %*% rotation # multiply with the rotation matrix
library(wordcloud)
plot(ll)
textplot(ll[,1],ll[,2],names[,2],xlim=c(-0.5,0.5),cex=1)
cmdscale(distances,k=2,eig=TRUE)$GOF
n = 56
cmdscale(dist(replicate(n,runif(n))),k=2,eig=TRUE)$GOF
```

Note: This program took 0.06 seconds to run on the computer we used. The information of the computer can be found in the Appendix.

Results

The goal of our project was to create a model for the species in National Parks and to be able to say something about that model and what it can tell us about National Parks or the species therein. A model is more than just a figure or plot. We could perform MDS on any random data set and get coordinate points that best fit that data set, but it could be meaningless. Fortunately, we were able to actually label the axes on our figure and create a true model, though certainly a limited one. This analysis of our figure proved to be the part of the model which required the most outside knowledge. Being able to create a model using MDS requires a good deal of knowledge (or research) into what is being modeled, not just pure mathematical knowledge.

The following Figure 3 is the MDS output of R code. Figure 4 shows the state in which each park is located.

FIGURE 3. MDS Output with Abbreviations

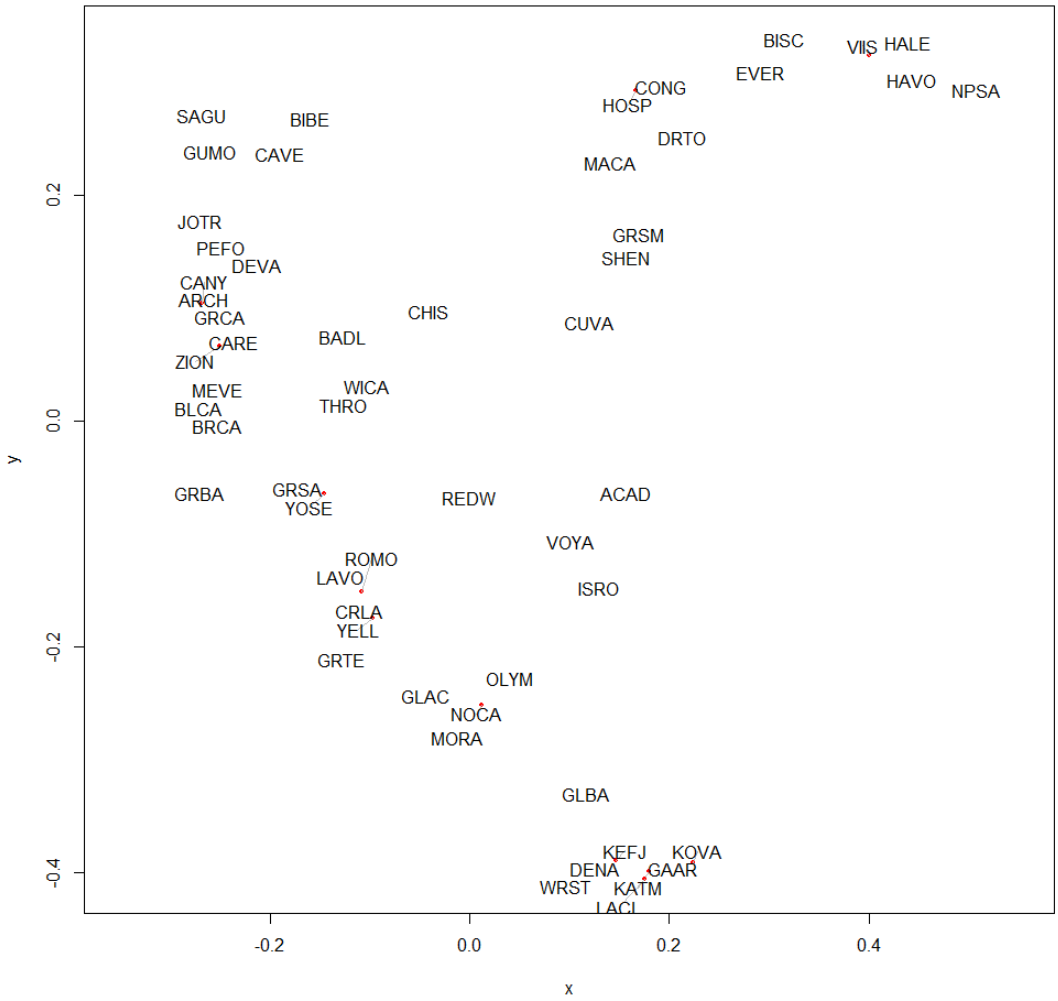
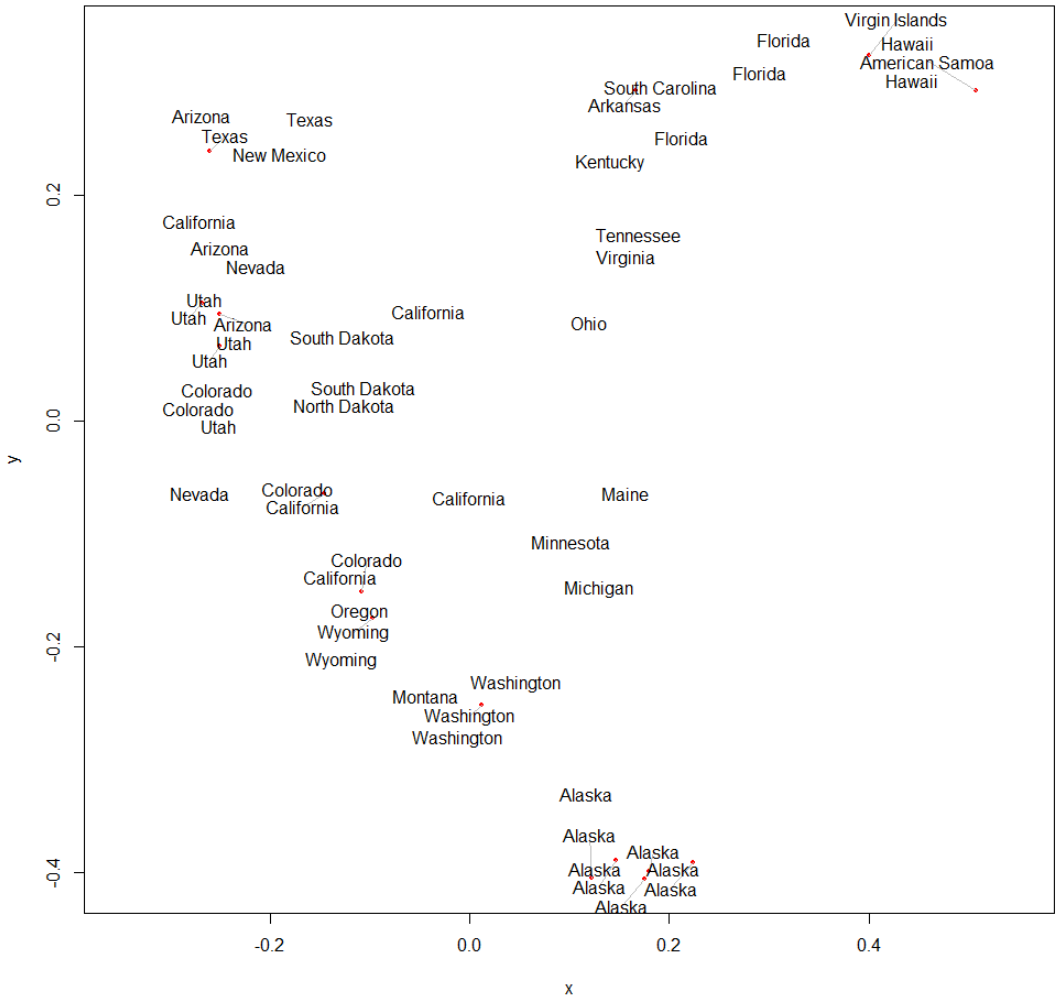
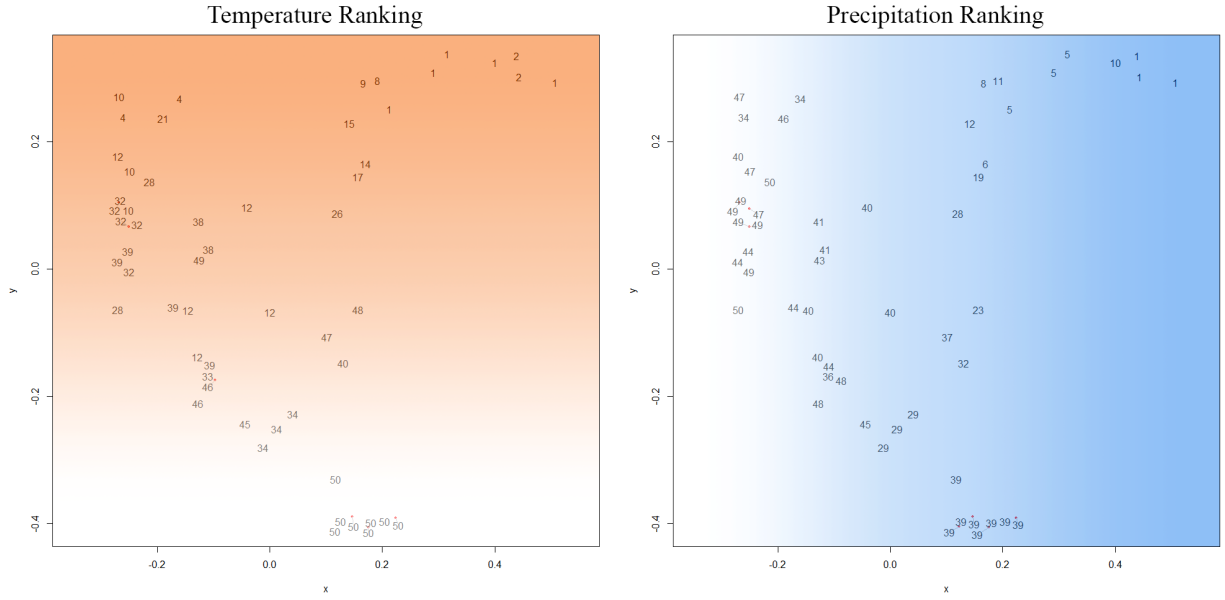


FIGURE 4. MDS Output with States of Parks' Locations



One possibility for the axes appeared to be temperature and precipitation.

FIGURE 5. Temperature and Precipitation Ranking by Located States



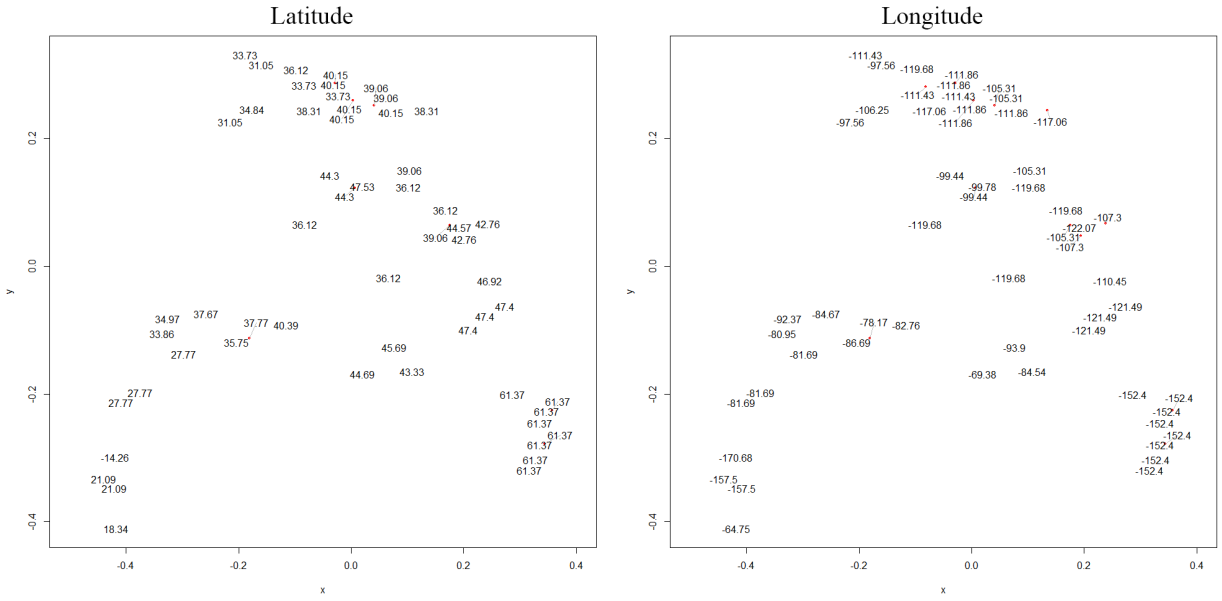
We replaced the park or state names in the Figures with rankings of average annual precipitation level [1] by state, and ranking of average temperature [3] by state. A ranking of 1 indicates the highest level of precipitation or temperature, and 50 indicates the lowest.

Assumption: Each park has the precipitation and temperature that are the same as the average of the state it is located in.

From this rotated version of the output for MDS, we can justify temperature as the y-axis dimension and precipitation as the x-axis dimension. The collection of outliers low on the y-axis (all parks in Alaska) do not quite fit with this interpretation for precipitation, but this clustering of Alaskan parks could be due to factors unique to geographical locations that are perpetually frozen; perhaps our axis for precipitation only accurately represents precipitation in its liquid phase, or water available to the species at that location.
If this is accurate, then the biological population in a National Park is dependent on rainfall and temperature. This makes sense, because living organisms depend on both water to thrive, and many organisms have very specific temperatures that they require to flourish, Varying amounts of water and different temperatures would be expected to foster different types of living organisms. We would intuitively expect different animals to be found in warmer, dryer climates than would be present in cold, wet environments. Rainfall was, in fact, one of our guesses for what we might see on our axes before we even created our model. We have some confidence in our model, therefore, just because it makes rational sense.
Another possibility we considered was that our axes simply corresponded to the latitude and

longitude of the park; this is shown in the un-rotated MDS figure below.

FIGURE 6. Latitude and Longitude by Located States



Although there could be an argument made for labeling the axes using latitude and longitude data [7], especially for the latitudinal data, there are clearer trends for precipitation and temperature. Unfortunately, regardless of how we label our axes, our model is not completely realistic. In fact, all we can conclude is that, as a general model, it does provide very general information. If we look at our Goodness of Fit (GOF) measure for our model, it is around 0.5. For some distance formulas we tried, GOF was even lower. We cannot say that our model is extremely accurate because it does not "fit" perfectly with our actual data. Still, the GOF of the distance matrix from a random matrix of the same size turns out to be 0.122 and our model provides significantly better fit values than random. Part of the result of this project was simply the realization that there are many different ways to model this same concept. Based upon how we define "distance" for these points, which do not have any dimensionality or physical sense of distance of their own, we might come up with a different model and be able to say different things about the similarity of animal species between parks. Some of these things we might conclude would be more accurate than others. Ultimately, to have confidence in our model, we want a high GOF. In our adjustments and extensions of the model, we explore what happens with a modified distance formula, and what happens with a smaller number of parks concentrated in the Northwest.

Adjustments and Extensions

Adjustment 1

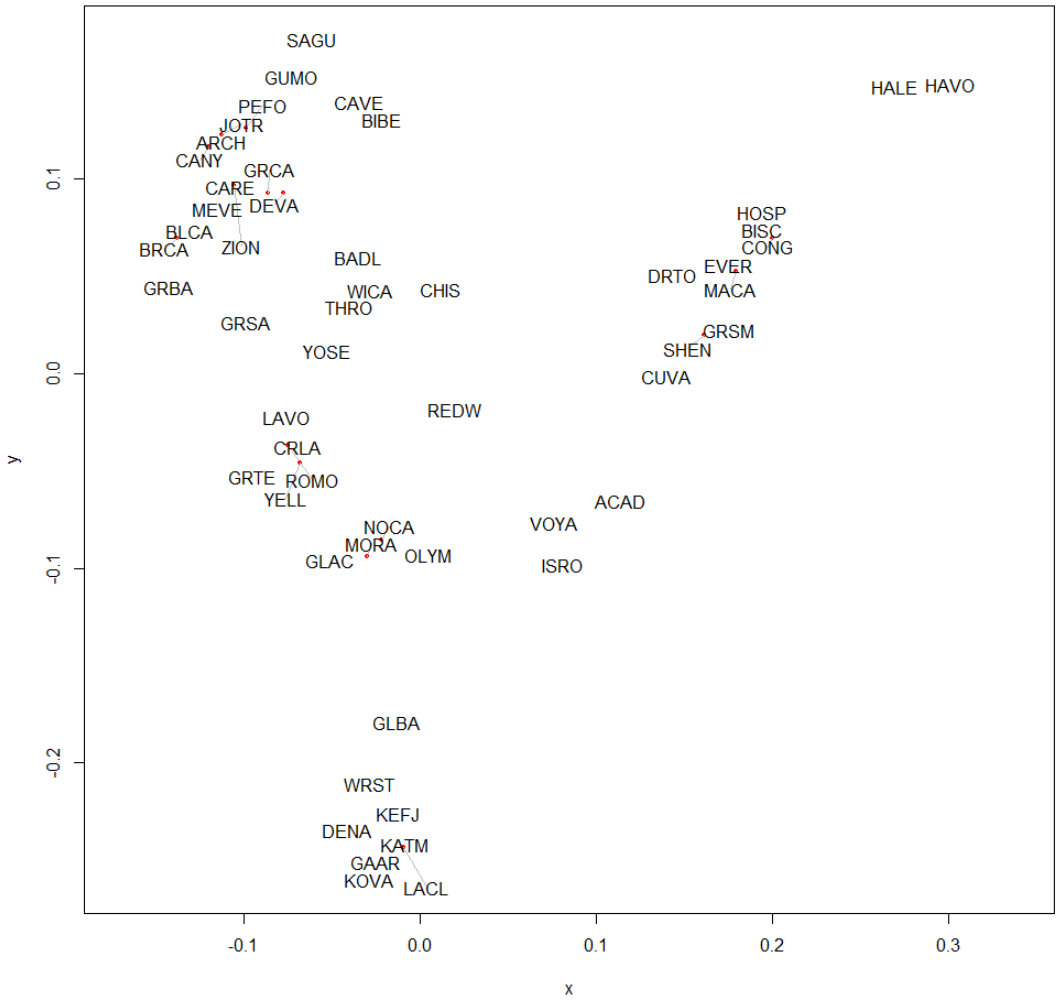
This adjustment will change the formula that defines the ecological distances between parks, formula (1). We make the following new definition of ecological distance.

$$d_{i,j} = \begin{cases} \dfrac{1}{(1+c_{i,j})^{\frac{1}{32}}} & \text{if } i \neq j,\ \forall i,j \in P \\ 0 & \text{if } i = j,\ \forall i,j \in P \end{cases}$$

The definition of $c_{i,j}, P$ can be found in the Model section.
The new formula for ecological distance is a function only of the number of common species; the function does not contain n_i, or n_j which is the number of species in each park. That means when park a and park b each with over 2000 species have k number of common species, and park c and park d each with only 500 species also have k number of common species, the ecological distances $d_{a,b}$, and $d_{c,d}$ are the same. $(a, b, c, d \in P)$
The formula is in reciprocal form because the ecological distance should have an inverse relationship with the number of common species. The more common species two parks have, the smaller the ecological distance is. We use $(1 + c_{i,j})$ because there might be two parks with 0 common species ($c_{i,j} = 0$). We take the exponent to be $\frac{1}{32}$ to decrease the rate of the ecological distance approaching 0. The number $\frac{1}{32}$ is chosen is because it's the reciprocal of the 5th power of 2, and the reciprocals of 1-4th powers of 2 does not decrease the rate of distances tending to 0 enough. That is, $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$ still produce cluster MDS figures that are difficult to interpret.
The MDS process is run in R, the output figure after some rotation and flipping is shown in Figure 7.

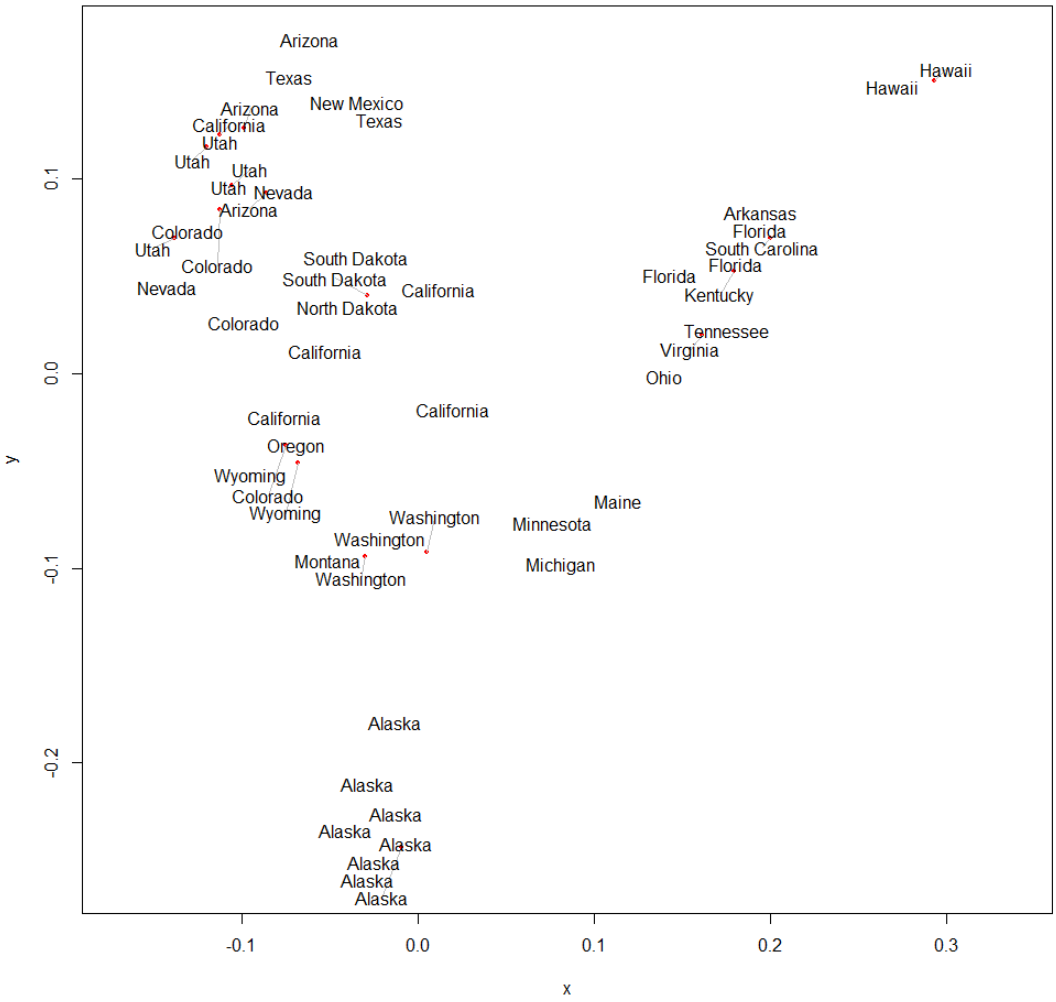
Note: Park NPSA and VIIS are removed from this model adjustment because of lack of climate data.

FIGURE 7. MDS Output



The corresponding state where each park is located in is shown in Figure 8.

FIGURE 8. States of Parks' Location



As we can see when x value increases, the state changes from Utah, Colorado, Arizona, Alaska, Michigan, Florida, and finally to Hawaii. Therefore, we make the interpretation of x-axis to be increasing in the amount of rainfall, or precipitation.

From the increasing direction of y-axis, the state changes from Alaska, Washington, Oregon, California, Virginia, Colorado, Utah, Texas, Hawaii, and finally Arizona. This generally follows the direction from the north pole to the equator. However, average temperature does not fit the graph very well since Florida is the hottest state instead of Arizona, and California is almost as hot as Arizona. The violation of the tendency drove us to seek other interpretations. One interpretation of the vertical axis is the average sunshine hours. As latitude becomes lower, closer to the equator, the sunshine hours generally go up. However, sunshine hours takes in to the account of cloudy, rainy climates, instead of purely based on latitude.

We have generated two graphs with ranking of average annual precipitation level [1] by state, and ranking of average annual sunshine hours [2] by state. A ranking of 1 indicates the highest level of precipitation or sunshine hours, and 50 indicates the lowest.

Assumption: Each park has the precipitation and sunshine hours that are the same as the average of the state it is located in.

Ranking graphs are shown in Figure 9 and Figure 10. A ranking of 1 indicates the highest level of sunshine hours or precipitation, and 50 indicates the lowest.

FIGURE 9. Precipitation Ranking by Located States

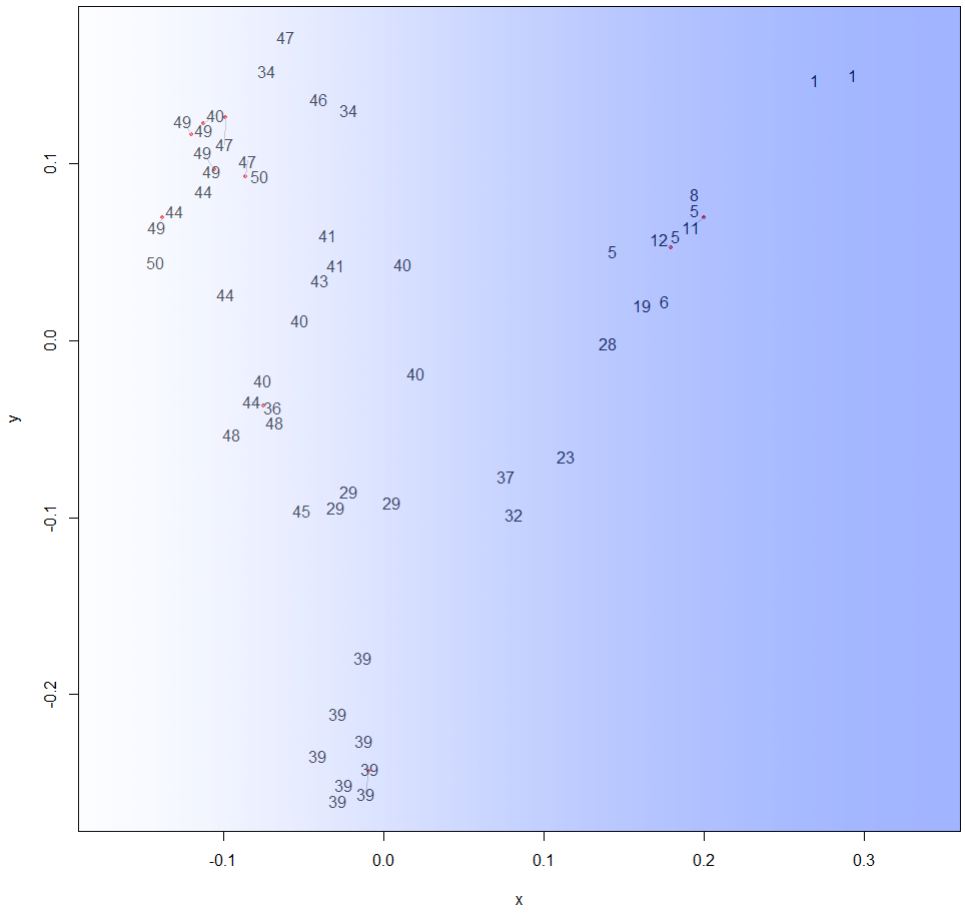
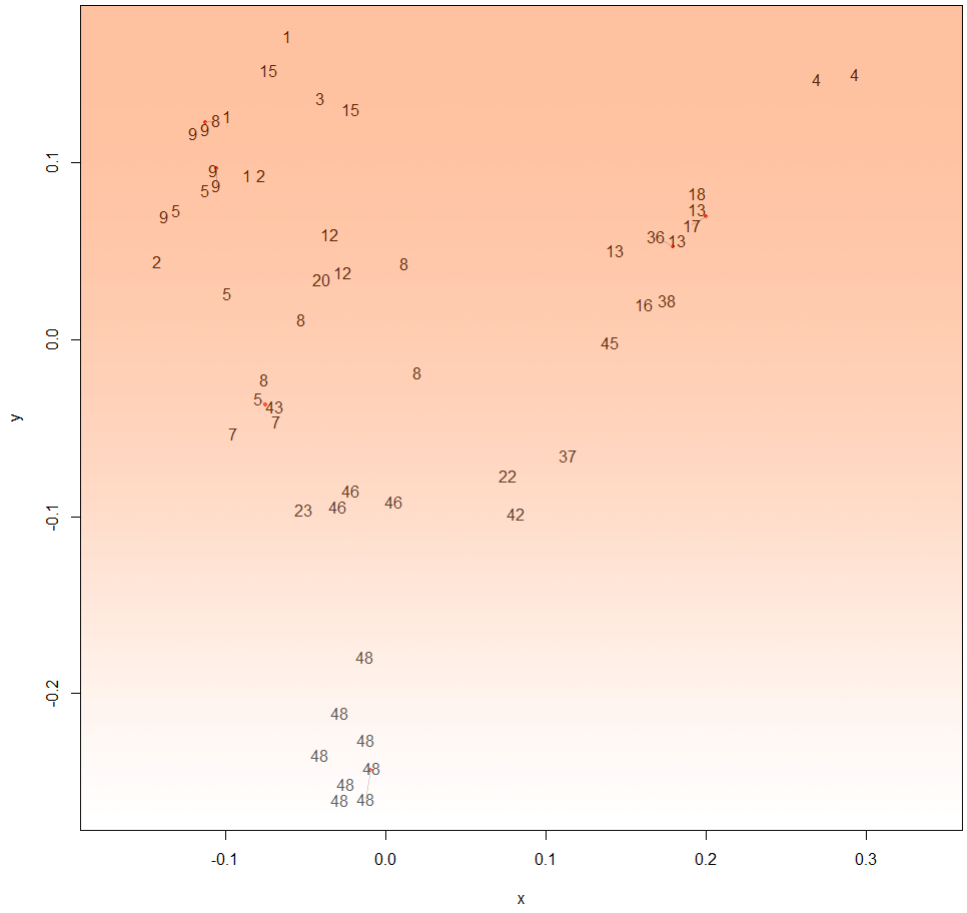


FIGURE 10. Sunshine Hours Ranking by Located States



As shown in Figure 9, along the x-axis, the precipitation generally increases as x increases. And as shown in Figure 10, along the y-axis, the sunshine hours generally increases. There are some rankings that are not following the general tendency very well. That could be caused by round-off errors from reducing the dimensions, or the data ranking by states not being representative.

The tendency of precipitation level and sunshine hours generally follows the graph. Also, precipitation and sunshine hours are important indicators of water and sunlight which are two crucial factors of life on earth. Different levels of precipitation and different sunshine hours could result in the existence of different species. The graph uncovers that the relationship between parks are due to gradual changes of the precipitation and sunshine hours.

Note: The goodness of fit (GOF) of this adjusted model is 0.21822.

Adjustment 2

Another adjustment will focus on the parks in the Northwest of the United States. This adjustment can give us a more accurate and also interesting result because the model contains significantly fewer parks that are relevant to our living area. These parks include Mount Rainier National Park, North Cascades National Park, and Olympic National Park in Washington, Crater Lake National Park in Oregon, Glacier National Park in Montana, and finally Grand Teton National Park and Yellowstone National Park in Wyoming. The locations of the parks are visualized on the map in the Figure 11 below.

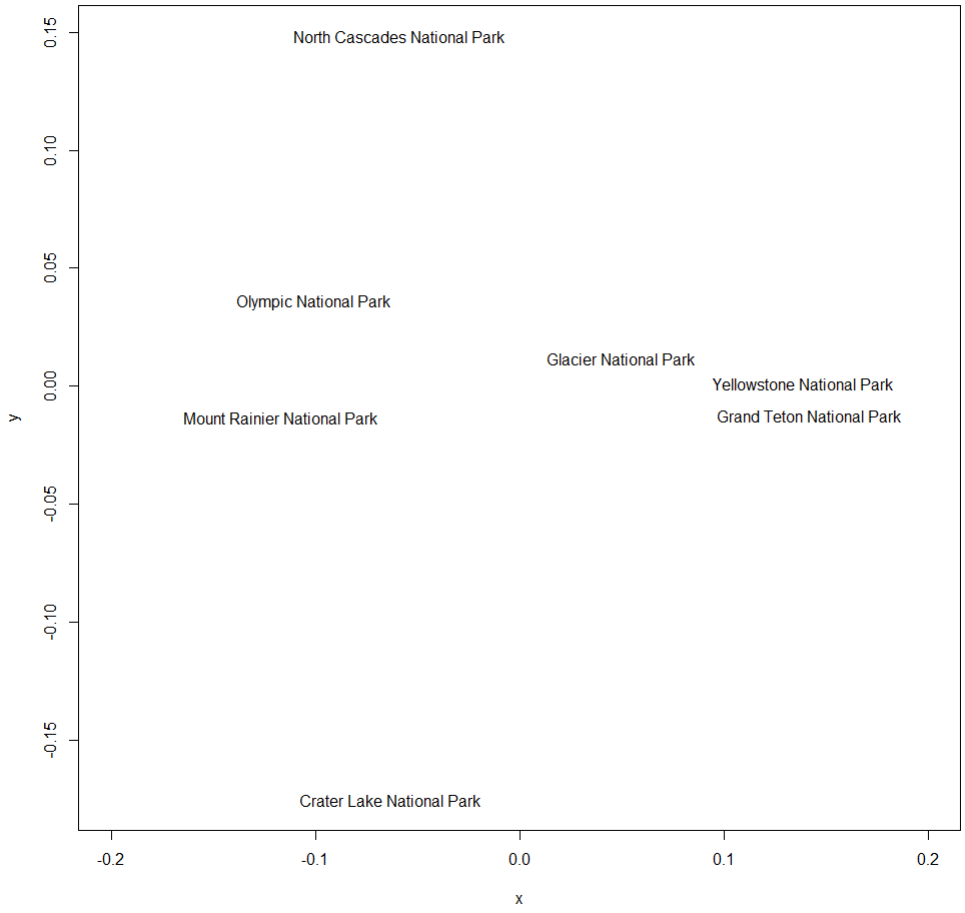
FIGURE 11. National Parks in the Northwest

The new set of data including those seven parks is run in the MATLAB codes we used for the original model to create a new distance matrix. The same MDS modeling process is run in R with the same distance formula (1) as the original model but with the new distance matrix. The original distance formula (1) we used is:

$$d_{i,j} = \left(1 - \frac{2c_{i,j}}{n_i + n_j}\right)^4 , \forall i, j \in P$$

The output figure is shown in Figure 12.

FIGURE 12. MDS Output of the Northwestern Parks



The goodness of fit (GOF) of this model is around 0.9 which shows how well our model is fit since it is very close to 1. Such a high GOF can be explained by the small size of data we used with seven parks. However, if we generate a random matrix of 7 by 7 dimension to apply MDS to, then the GOF of the random model gives 0.65 so our model is certainly better fit than random. We can compare Figure 11 and Figure 12, to conclude that x-axis of the MDS model is related to how close each park is to the Pacific coast. In other words, as x value increases, the distance of each park from the Pacific coast increases. The leftmost parks in the model are Mount Rainier and Olympic National Parks, which are placed in Washington, the closest to the Pacific coast. On the other hand, the rightmost parks in the model are Yellowstone and Grand Teton National Parks in Wyoming being the farthest away from the west coast. Glacier National Park is right before the two Wyoming parks on the horizontal axis, consistent with the fact that it is located in Montana, farther away from the west coast than Washington or Oregon. The Crater Lake and North Cascades National Parks, which are in Oregon and Washington respectively, have x values that are very close to

those of the two leftmost parks in Washington state, also being consistent with our definition of the x-axis.

The vertical axis is subject to question because of an outlier of Glacier National Park. We can, however, cautiously conclude that the vertical axis of the model represents the latitude of the parks. As we can see again from Figure 11 and Figure 12, the y values of the model generally agree with the latitude of each park on the visual map. The topmost park in the model is North Cascades National Park, while the bottommost park is Crater Lake National Park, which is consistent with their latitudes being the highest and lowest among the parks, respectively. The rest of the parks also demonstrate this trend as their y values belong to the middle of the model, except the previously mentioned outlier of Glacier National Park, which is supposed to have a high y value with its high latitude. Overall, this MDS model of Northwestern National Parks in Figure 12 shows a plot that is very similar to the physical map in Figure 11. This was an unexpected result, but it is comprehensible because the locations of the parks indeed have influences on the kinds of species each park accommodates for climatic and geographical reasons.

Conclusion

Through examining multidimensional scaling of the species in National Parks we were able to gain insight into the relationship between species in National Parks and the environment each of the National Parks fall under. Our ultimate goal in this project was to visually represent information about 56 National Parks in the United States and the species they have in common and to try to determine the factors that contribute to the biological population of each park. MDS is a useful method of analyzing and visualizing relationships, but it has its drawbacks in that a subjective judgement must be made in determining how the information presented corresponds to the system behavior. From our analysis we were able to conclude that the factors most likely to lead to similarities and differences in species between each park are the average precipitation and temperature. As discussed previously, this result is to be expected since certain species are only able survive under specific climates, thus it is expected that there will be similar species in National Parks located in areas with similar precipitation and temperature, and less species in common between areas with drastically different climates. Under a different distance formula, however, the plot led us to believe sunshine hours were a greater contributing factor over temperature. Due to the subjective nature of MDS other factors could potentially match our data set and a low goodness of fit could lead to an incorrect interpretation of the system. Suggestions on how this could be improved include using formulas that lead to a better goodness of fit, changing the dimensions of the data set, using regional versus statewide data for a better image of the climate at the specific location of the park, and using analytical methods, such as multiple regression techniques, to interpret the dimensions and determine whether the variables judged to match the dimensions are an accurate portrayal of the system. In looking at a smaller subset of our data we were able to draw completely new conclusions about species relation in parks. We found that when looking at parks only in the Pacific Northwest, temperature, sunshine hours, and precipitation were less influential factors, and instead the proximity to the Pacific Ocean, and the latitude, or distance from the equator more closely matched our axes. This reveals one of the great benefits to using MDS: multiple conclusions can be drawn from the same data by analyzing its subsets. It allows us to mathematically model similarity and dissimilarity and visualize data that might otherwise be overwhelming. MDS combines both the psychological (what does it mean for two things to be similar?) and the mathematical (how do we define this similarity in a mathematical way?) and is, therefore, a useful tool for analyzing the real world.

References

[1] CurrentResults, ed. *Average Annual Precipitation by State*. Accessed November 28, 2016. URL: https://www.currentresults.com/Weather/US/average-annual-state-precipitation.php.

[2] CurrentResults, ed. *Average Annual Sunshine by State*. Accessed November 28, 2016. URL: https://www.currentresults.com/Weather/US/average-annual-state-sunshine.php.

[3] CurrentResults, ed. *Average Annual Temperature for each US State*. Accessed November 28, 2016. URL: https://www.currentresults.com/Weather/US/average-annual-state-temperatures.php.

[4] Patrick J.F. Groenen and Ingwer Borg. *The Past, Present, and Future of Multidimensional Scaling*. Ed. by Econometric Institute Report EI 2013-07. Accessed November 30, 2016. URL: repub.eur.nl/pub/39177/EI2013-07.pdf.

[5] National Resource Stewardship Report no. 525. U.S. Department of the Interior and National Park Service. 17. Science Natural Resource Technical Report, eds. *Integrated Upland Monitoring in Arches National Park*. Accessed November 14, 2016. URL: https://irma.nps.gov/DataStore/DownloadFile/443652.

[6] N. C. Kenkel and L. Orlci. *Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies: Some New Results.* Ed. by no. 4 (1986): 919-28. doi:10.2307/1939814. Ecology 67. Accessed November 14, 2016. URL: https://www.jstor.org/stable/1939814?seq=1#page_scan_tab_contents.

[7] R Reese. *List of Latitudes and Longitudes for Every State*. Accessed November 30, 2016. URL: https://inkplant.com/code/state-latitudes-longitudes.

[8] United States. National Park Service. National Parks Service., ed. Accessed November 18, 2016. URL: https://irma.nps.gov/NPSpecies/.

Appendixes

The following table provides the index i, the abbreviation, and the name of every US national park in the data set.

Table of US National Parks

Index *(i)*	Abbreviation	Name
1	ACAD	Acadia National Park
2	ARCH	Arches National Park
3	BADL	Badlands National Park
4	BIBE	Big Bend National Park
5	BISC	Biscayne National Park
6	BLCA	Black Canyon of the Gunnison National Park
7	BRCA	Bryce Canyon National Park
8	CANY	Canyonlands National Park
9	CARE	Capitol Reef National Park
10	CAVE	Carlsbad Caverns National Park
11	CHIS	Channel Islands National Park
12	CONG	Congaree National Park
13	CRLA	Crater Lake National Park
14	CUVA	Cuyahoga Valley National Park
15	DENA	Denali National Park and Preserve
16	DEVA	Death Valley National Park
17	DRTO	Dry Tortugas National Park
18	EVER	Everglades National Park
19	GAAR	Gates Of The Arctic National Park and Preserve
20	GLAC	Glacier National Park
21	GLBA	Glacier Bay National Park and Preserve
22	GRBA	Great Basin National Park
23	GRCA	Grand Canyon National Park
24	GRSA	Great Sand Dunes National Park and Preserve
25	GRSM	Great Smoky Mountains National Park
26	GRTE	Grand Teton National Park
27	GUMO	Guadalupe Mountains National Park
28	HALE	Haleakala National Park
29	HAVO	Hawaii Volcanoes National Park
30	HOSP	Hot Springs National Park
31	ISRO	Isle Royale National Park
32	JOTR	Joshua Tree National Park
33	KATM	Katmai National Park and Preserve
34	KEFJ	Kenai Fjords National Park
35	KOVA	Kobuk Valley National Park
36	LACL	Lake Clark National Park and Preserve
37	LAVO	Lassen Volcanic National Park
38	MACA	Mammoth Cave National Park
39	MEVE	Mesa Verde National Park
40	MORA	Mount Rainier National Park
41	NOCA	North Cascades National Park
42	NPSA	National Park of American Samoa
43	OLYM	Olympic National Park
44	PEFO	Petrified Forest National Park
45	REDW	Redwood National Park
46	ROMO	Rocky Mountain National Park
47	SAGU	Saguaro National Park
48	SHEN	Shenandoah National Park
49	THRO	Theodore Roosevelt National Park
50	VIIS	Virgin Islands National Park
51	VOYA	Voyageurs National Park
52	WICA	Wind Cave National Park
53	WRST	Wrangell - St Elias National Park and Preserve
54	YELL	Yellowstone National Park
55	YOSE	Yosemite National Park
56	ZION	Zion National Park

Here are some information about the machine, and software versions.

Machine Information:

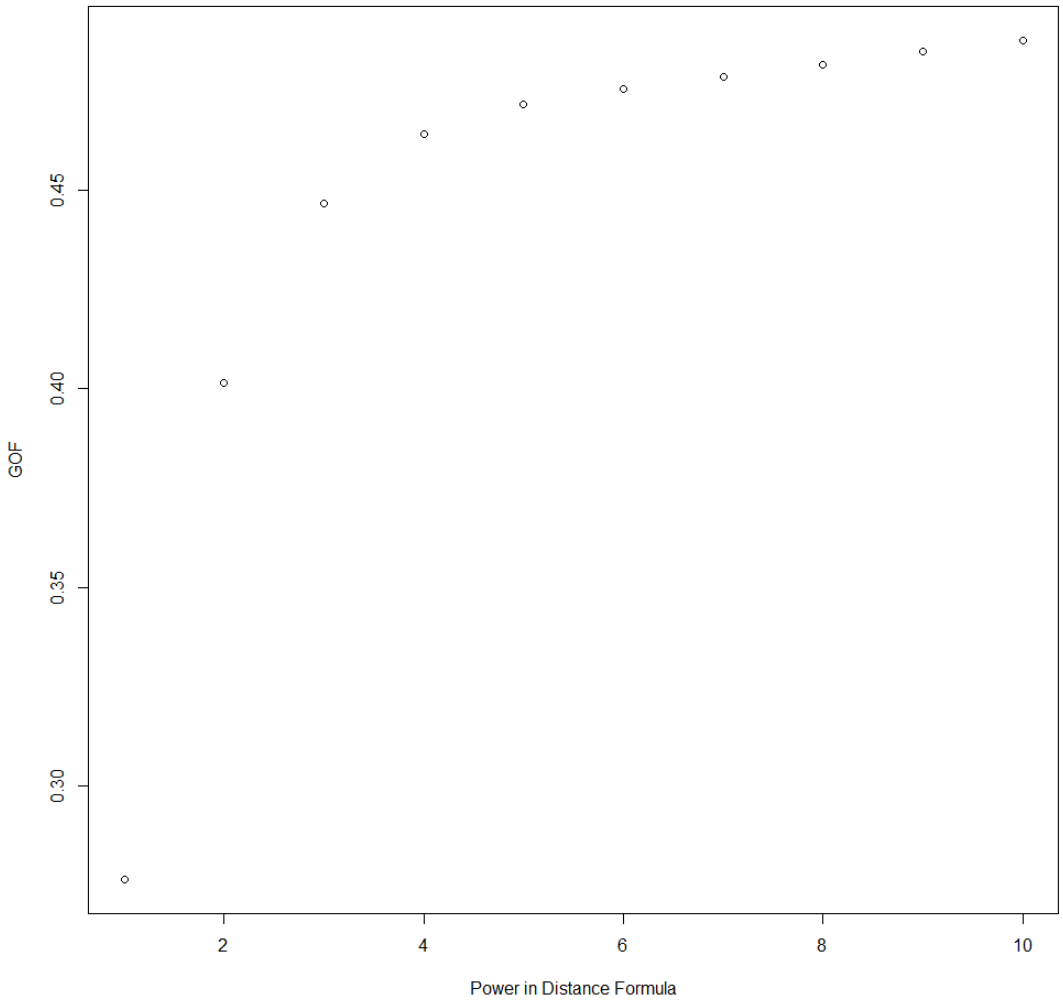
```
Operating system  : Windows 10 Pro (C) 2016 Microsoft Corporation
System type       : 64-bit Operating System, x64-based processor
Processor         : Intel(R) Core(TM) i7-4712HQ CPU @2.3GHz
Memory            : 16.0GB
```

Software Information:

```
MATLAB version    : Matlab R2015a (8.5.0.197613) 64-bit (win64)
                    (C) 1984-2015 The MathWorks, Inc.
R version         : 3.3.2 (2016-10-31) 64-bit
                    Copyright (C) 2016
                    The R Foundation for Statistical Computing
RStudio version   : Version 1.0.44  (C) 2009-2016 RStudio, Inc.
```

The following Figure 13 is a graph of the GOF of the original model against the power in distance formula (1).

FIGURE 13. GOF of the Model for Different Powers in Formula (1)



Files and data folder can be found in the folloing link.

https://goo.gl/V1k2Lk

Backup link:

https://drive.google.com/drive/folders/0BwRnkw8WUGUTbFIwLVpxY29YRlE

List of files or folders:

(1) 56 National Parks Data (by Abbreviation) Folder
This folder contains the original data downloaded from the NPS Offical website. Files are in format of `xlsx`, and are named by abbreviation of parks.

(2) 56 National Parks Data (by Index) Folder
This folder contains the original data downloaded from the NPS Offical website. Files are in format of `xlsx`, and are re-named by indecies from 1-56.

(3) 56 US National Parks Map
This is a link a map of the 56 US national parks based on Google Map. Each marker represents a park and its location. Clicking on the marker shows the abbreviation, index, and full name of the park.