Erosion Susceptibility of Boston Harbor Bluffs

Teaching Fellow - Minh Thu Bui Supervisor - Professor Masanao Yajima
Yingmai Chen, Maysen Pagan, Chang Shi, Yan Wang

Project Background and Objectives

Sarah Black is a PhD candidate in the Earth and Environment Department at Boston University's Graduate School of Arts and Sciences. The client is interested in the variability of erosion rates of the bluffs located in the Boston Harbor islands. The main objective of this project is to classify 31 Boston Harbor bluffs into different erosion vulnerability categories based on measured variables from the provided data. The client is also interested in which variables have the largest impact on erosion susceptibility.

The data provided by the client contains 10 columns and 35 rows representing 35 bluffs. The first column contains the names of each bluff. The variables that were used throughout our analysis include the orientation angle, retreat rate in meters per year, wave height and maximum wave height, mud composition, base and bluff elevation in meters, and a binary variable indicating the presence of a seawall. The orientation angle is a measure of the degree at which the bluff faces with north being 0 or 360 degrees and south being 180 degrees. The retreat rate of each bluff was determined using a Digital Shoreline Analysis system which uses satellite imaging to measure how far back the bluff moves. The maximum wave height was measured through simulating different scenarios of wind speed and wind direction, extracting the maximum simulated wave height of all scenarios measured. The wave height variable is the wave height simulated for each bluff for one specific scenario of NNE winds at 15 meters per second, which are typical winds leading up to a winter storm. The mud composition variable is a percentage of each bluff's sediment that contains mud based on samples taken. Not including the names of the bluffs, these 8 variables were used in our analysis to classify bluff erosion susceptibility and determine which variables are associated the most with erosion.

Exploratory Data Analysis

Data Cleaning and Organizing

The original data set contained 35 bluffs from the Boston Harbor. However, 4 bluffs contained missing values for the retreat rate variable as the satellite imaging was unclear for those bluffs. These 4 bluffs were removed from the data set and we continued our analysis with 31 bluffs. The data set also included a variable created by the client called `ErosionVulnerability` which declared the susceptibility of each bluff to erosion based on if the bluff's retreat rate was in the lower, middle, or upper third of the sorted retreat rates. This variable was not included in our analysis. The last step in preparing our data for the models was to scale the data. The method used is called Min-Max Scaling so that each feature scales the range of [0,1]. The formula to scale each feature is:

$$x^{\text{new}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x is each of the 8 variables.

Data Exploration

We begin the data exploration by observing the relationship and patterns between the orientation of the bluff and the presence of a seawall with the retreat rates of the bluffs. Figure 1 demonstrates this relationship on a

polar plot.

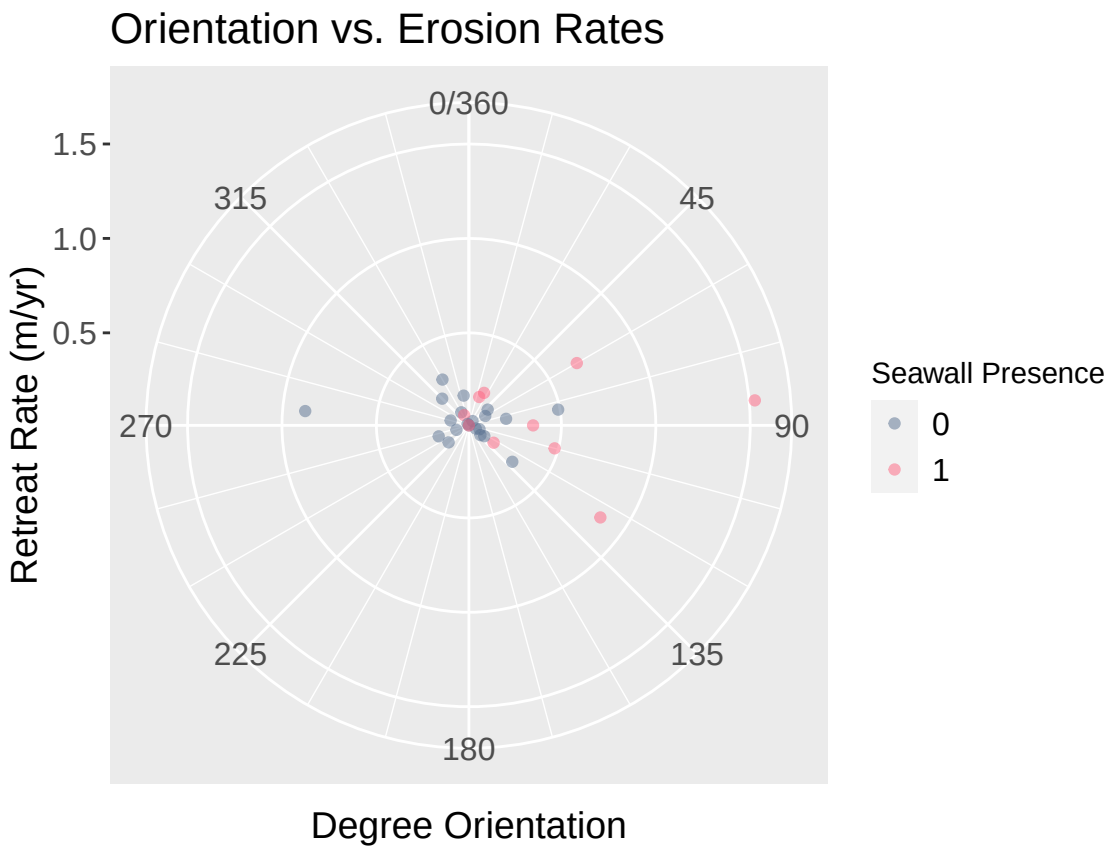

Figure 1: Polar plot of bluff orientation against retreat rate.

It appears that more bluffs that face East and therefore have lower degrees of orientation also have a seawall present. This intuitively makes sense as those bluffs with lower degrees of orientation face the mouth of the harbor and are more likely to have a seawall with the intentions of preventing erosion. We can also see that those bluffs with orientations less than 180 degrees tend to have higher retreat rates than those bluffs with orientations between 180 and 360 degrees.

Models

The method used to group the bluffs based on similar characteristics and compare susceptibility to erosion involves visualizing a multidimensional scaling plot and heatmap with hierarchical clusterings of the bluffs. These clusters are determined by the "distances" between each bluff.

Bluff Distances

Defining a distance between bluffs provides a measure of dissimilarity between each bluff. Variables for each of the bluffs included seven numerical variables as well as one binary or categorical variable (seawall presence). Not all of the variables are numeric and as a result, the Gower's distance is calculated between two bluffs which accounts for both numerical and categorical variables.

Gower's distance calculates a matrix of dissimilarities for each of the $\frac{n(n-1)}{2} = \frac{31(31-1)}{2} = 465$ pairs of bluffs. This method combines the Manhattan distance for numerical variables and Hamming loss for

categorical variables to get the total distance between two observations. For the seven numerical variables, a range-normalized Manhattan distance is calculated:

$$|\frac{T_{ik} - T_{jk}}{\text{range}_k}|$$

where $i, j = 1, 2, \ldots, 22$ and $k = 1, 2, \ldots, 7$. For the remaining categorical variable, the Hamming loss is calculated with the following formula:

$$I(T_{i,seawall} \neq T_{j,seawall})$$

where $i, j = 1, 2, \ldots, 22$. This indicator will equal 1 if $T_{i,seawall} \neq T_{j,seawall}$ and 0 if $T_{i,seawall} \neq T_{j,seawall}$. Gower's distance between two bluffs is calculated by summing these variable distances and dividing by 8, the total number of variables.

Multidimensional Scaling

Once we have the pairwise distances between bluffs from the data, multidimensional scaling provides one method of visualizing the clusters and their distances from each other on a two dimensional plot. Although the data for each bluff is multivariate, multidimensional scaling preserves the distances between pairwise observations when plotting. Multidimensional scaling takes in the calculated distances and returns a set of 31 points where the distances between each point is approximately equal to the dissimilarities between each point. These points are then plotted on a two dimensional scatter plot to visualize the "closeness" of the bluffs and can be viewed in Figure 2.

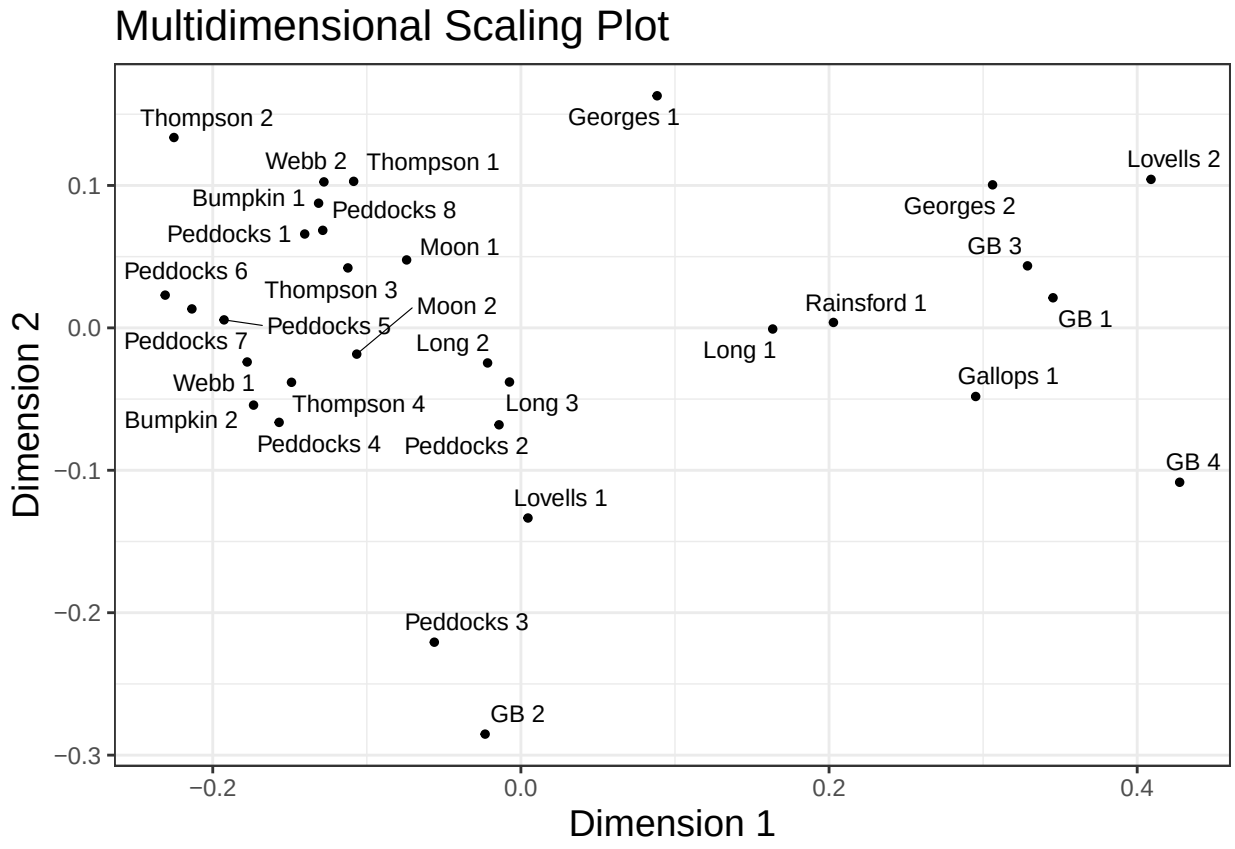

Figure 2: Multidimensional scaling plot of 31 bluffs.

Heatmap

We can also use a heatmap to visually identify similarities and dissimilarities between bluffs. In a heatmap cells are color-coded to quickly compare one row or blff to another. The heatmap also allows one to aggregate the rows in clusters based on the distances calculated from @sec-distances. Figure 3 below is the heatmap generated for the 31 bluffs using Gower's distance.

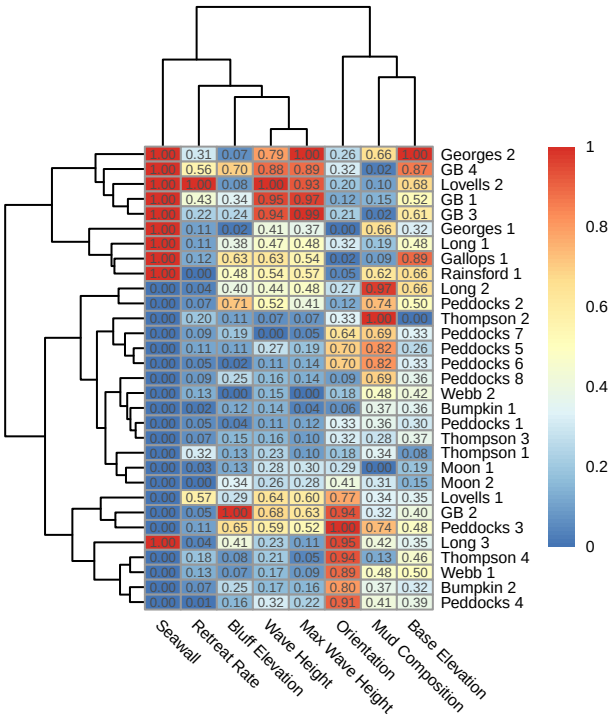

Figure 3: Clustered heatmap of 31 bluffs.

Analysis and Conclusion

Appendix