Michael Freeman

Tim Chartier

CSC 381

Clustering Golf Courses by Yardage Variance

**Introduction**

This paper will explore the relationship between variance in yardages of specific types of holes on golf courses and how highly the golf course is ranked by clustering the courses based on that data. For this project I looked at the top 70 rated golf courses in the United States and calculated the variance of the yardages of the par 3’s, par 4’s, and par 5’s. The hypothesis was that golf courses with higher variance of yardages have more variety of holes and therefore will be more enjoyable to play and more stimulating. The results could influence the thought process on how to design or redesign golf courses so that they would be better courses. To do this I used two methods of clustering on the combined variance of the par 3’s, 4’s, and 5’s. One method was the Singular Value Decomposition, which uses a matrix that contains the difference between the variance of the two corresponding courses in the column and row. The second method uses Fiedler’s method, which utilizes a matrix with only 1’s and 0’s, where a 1 signifies a connection between the two courses in the corresponding column and row. By using two different clustering methods the methods can be compared for accuracy and correlation to the ranking of the golf courses.

**Methods**

In order to cluster the data that was collected, I choose two methods: Fiedler’s Method and the Singular Value Decomposition (SVD). In both methods I needed to create a matrix that contains all of the relationships between all of the courses. Before I could create these two matrices and get into the methods, I needed to have the data in the correct form. I used MATLAB to create vectors that contained the variances of yardages on Par 3’s, 4’s, and 5’s. Then I added the vectors together to have a total variance number. This vector was what I used to create the matrices described below.

In Fiedler’s method I choose to create the connections by allowing a connection if the absolute value of the difference in total variance was less than the average variance of all the courses divided by four. Otherwise the corresponding value in the matrix would be a zero. This creates a matrix with only 1’s and 0’s in the entries. For example, if the average variance is 8 then the difference between two course’s variances must be 2 or less. So if course A has a variance of 10 and course B has a variance of 11, then a 1 is entered in the matrix in position (A, B) and (B, A). If course B had a variance of 13 then those same entries would be 0. The code used for class was adapted to cluster this data. The code first takes the adjacency matrix (A) that was created with the 0’s and 1’s and sums each column and places the sum along the diagonal, while all other values are zero to create a diagonal matrix (D). Then a Laplacian matrix (L) is formed by computing L = D – A. By Fiedler’s method we will find the eigenvector with the second smallest eigenvalue and then each entry in the column is sorted, so that the positive signs go together, and the negative signs go together. To get more clusters we used the eigenvector with the next biggest eigenvalue as well.

In the SVD method I needed entries in the matrix to represent the distance between each course. So, for example if course A has a variance of 10 and course B has a variance of 20, then the number 10 would go in the entries of the matrix (A, B) and (B, A). This creates a more accurate comparison between variances of each course. To start the clustering process, we divide each entry in a column by the column’s mean. Now that we have a matrix we can use the SVD to produce a matrix M where *M* = *UV*T . The columns are U are the left singular vectors and the columns and *V*T are the right singular vectors. Next we find the singular vector which corresponds to the largest singular value. Then each entry in the column is sorted, so that the positive signs go together, and the negative signs go together. For this project we used code written for class. In order get the 7 clusters shown in the data more than the largest singular value was used in clustering.

**Data**

The data for this project was the scorecards of the top 70 golf courses in the United States. From this data I found the variance of yardages for par 3’s, 4’s, and 5’s. The data and figures below are the results of the two methods of clustering. The first set of figures, tables, and graphs show the results of clustering using Fiedler’s method. Graph 1 shows all of the connections each of the courses have with the other courses after creating connections for courses with a similar variance.



Graph 1: Graph of the connections between the courses using Fiedler’s Method by plotting the points and connecting them



Figure 1: Graph of the connections between the courses using Fiedler’s Method using dots

A screenshot of a cell phone

Description automatically generated

(Table 1: The clusters produced by Fiedler’s method )

The following set of tables shows the results from clustering using the SVD. The tables contain all seven clusters. The outlier’s in this method are cluster’s 6 and 7. Olympic club gets its own cluster and Sahalee CC and Sea Island are together in cluster 7. The other clusters are larger and contain the courses that have more common variances.

A close up of text on a white background

Description automatically generated

(Table 2: The clusters(1-4) produced from the SVD)

A close up of text on a white background

Description automatically generated

(Table 3: The clusters (5-7) produced from the SVD)

**Analysis**

The two different methods of clustering showed two very different results. One of them showed a strong correlation between a lower ranking and variance while the other showed there was almost none. The hope for each of these methods was for them to show there is a negative correlation between these two variables, which means as the variance gets bigger, the ranking of the course gets better.

Fiedler’s method, which used the unweighted connections, showed almost no correlation between the average rank of each cluster and the average combined variance of each cluster. The graph below shows all four of the clusters and how those two data points relate. The graph shows an almost horizontal trendline, which signifies no relation between the variables. To further confirm this fact the chart below the graph shows the correlation is a 0.028, a slightly positive number, which was the opposite of the desired outcome.

(Graph 2: Graph of Average Rank vs. Average Variance for the clusters using Fiedler’s Method)

**A screenshot of a cell phone

Description automatically generated**

(Table 4: Table of the average rank, average variance, and the correlation using Fiedler’s Method)

These results show that using Fiedler’s method of clustering is not the most effective in showing a relationship between these two variables. It can be seen in the lack disparity of the in the average variances. There are two pairs of points that have very similar y values as seen on Graph 2. This is part can be attributed to only having 4 clusters, which did not allow for the outliers of the data to be correctly clustered as they were in the SVD method. In conclusion this method was not an effective way to cluster this particular data and did not produce the desired results.

The SVD method was very successful in clustering the data and produced a high correlation between average variance and average rank within the clusters. This method produced seven clusters as seen in Graph 3. The trendline on the graph had a large negative slope which signals a strong negative correlation between average rank and average variance. This conclusion is supported by Table 5, which shows the correlation between the two to be -0.87. This means as variance goes up, the ranking will go down, which means the course will be ranked better. Since correlation is measured between 1 and -1, the results are significant. This correlation is definitely not the only reason behind the high ranking of golf courses. Other factors include course condition, surrounding nature and views, prestige, and difficulty. Therefore, given there are numerous other factors that contribute to a highly ranked course, yardage variance on par 3’s, 4’s, and 5’s makes a strong impact.

(Graph 3: Graph of Average Rank vs. Average Variance for the clusters using the SVD)

A screenshot of a cell phone

Description automatically generated

(Table 5: Table of the average rank, average variance, and the correlation using the SVD)

These results show the SVD is successful in clustering this data as the average variance is almost evenly spread out between the clusters, unlike what was seen in the clustering by Fiedler’s method. Having seven clusters instead of four in addition to more accurate clustering caused the SVD method to be more successful in finding the correlation between golf course ranking and the overall variance. This is because the SVD was able to capture the 3 main outliers, Olympic Club, which had the highest variance by about 6000, and Sahalee CC and Sea Island, which had the two lowest variances of the 70 courses. The SVD method of clustering did very well in that it accounted for the highest and lowest outliers, while dividing the very crowded middle into 5 distinct clusters.

**Summary**

Overall the project was able to produce an effect method of clustering and an ineffective method of clustering, while showing the predicted inverse relationship between overall variance and ranking. The SVD was successful because of the number of clusters created and accuracy of the clustering, while Fiedler’s method did not create precise clusters and therefore did not show the desired relationship between the two variables. The result of this project can be used by golf course designers, golf course rankings, and the casual golfer who may just be curious as to why one course might be better than another. Variance in yardage on par 3’s, 4’s, and 5’s is definitely not the only variable that goes into the quality of the golf course, but this project has shown that it has a significant effect. Future work on this topic might use a much higher quantity of courses to complete this project. The top 70 courses in the country are all quality golf courses and therefore there might not have been as much separation between the elite courses and the very good ones. An extension of this project could be to create a ranking system based off the variance statistic, the higher the better, and compare it to other rankings of golf courses. Finally adding another variable to cluster on that might explain why some courses with high variance aren’t as highly ranked would increase the accuracy of the clustering and explain more of the reasoning behind the rankings.