Project 2: Cluster Analysis

This goal of this project is to assess your ability to perform the cluster analyses we have covered in class as well as test the significance of groups/clusters recovered from the cluster analyses.

I. K-means clustering of the Darlingtonia data set. MRPP and ANOSIM of resulting groups.

Please answer the following questions based on K-means analysis of the Darlingtonia data set. Log transform (log1p) the variables tube\_diam, keel\_diam, wing2\_length, hoodarea, wingarea, and tubearea and standardize the data set using z-scores.

1. What is the optimal clustering solution (i.e., number of clusters, k) for the Darlingtonia data set according to a scree plot? (1pt)

2 clusters seems to be the inflection point for the scree plot.

1. What is the optimal clustering solution (i.e., number of clusters, k) for the Darlingtonia data set according to the average silhouette width? (1pt)

2 has the highest silhouette width (.276).

1. Conduct a PCA on the Darlingtonia data set (this should be familiar) and plot a biplot showing the optimal number of groups from the K-means analysis in color. Submit the plot with your worksheet. (1pt)
2. Which PCA axis best separates your clusters? (1pt)

Axis 1.

1. Based on the PCA axes, which variable best separates your clusters and what is the mean of that variable for each cluster? (1pt)

Axis 1 seems to be a fairly even composite of wingarea, hoodarea, wing1\_length, mouth\_diam, and wing2\_length. I’ll go with wingarea, because that has the highest contribution to the axis (0.4, compared to the others of 0.36-0.388).

The mean of wingarea for cluster 1 is 30.2 and cluster 2 is 13.3.

6)  Run a Multiple Response Permutation Procedure (MRPP) on the clusters obtained from the optimal K-means cluster solution. Report the p-value and expected delta from the MRPP. What do you conclude based on the MRPP results? (1pt)   
  
The p-value is 0.00099; the expected delta is 4.14, and the actual delta is 3.48. The clusters are significantly different.

7)  Run an Analysis of Similarity (ANOSIM) on the clusters obtained from the optimal K-means cluster solution. Report the p-value and R statistic from the ANOSIM. What do you conclude based on the ANOSIM results? (1pt)

The p-value is 0.000999 and R is 0.5159. The clusters are significantly different.

II. Polythetic Agglomerative Hierarchical Clustering of the Dune vegetation data set (dune.csv).

This dataset contains species presences/absence for 30 species across 20 dune meadow sites. Use the Jaccard index to create your dissimilarity matrix.

8)  Run cluster analysis and build dendrograms using the 6 fusion methods we discussed in class. Submit the dendrogarm using the “Average-Linkage” (UPGMA). (1pt)

9)  Calculate the cophenetic correlation coefficient for each fusion method. Which method most accurately depicts the original dissimilarity matrix? (1pt)

The average linkage – it has the highest cophenetic coefficient, 0.88.

10)  Calculate the agglomerative coefficient for each fusion method. Which method has the most cluster structure? (1pt)

The Ward has the most cluster structure, 0.73.

11)  Run a bootstrap randomization of the Average-Linkage dendrogram (method.dist="binary" is the same as jaccard). How many clusters/groups are identified that have a multi-scale bootstrap probability (au) > 0.95 at the highest level in the hierarchical tree (i.e., don’t count clusters that are part of larger significant clusters)? (1pt)

Two.

III. Polythetic Divisive Hierarchical Clustering of the Dune vegetation data set (dune.csv).

Use the Jaccard index to create your dissimilarity matrix.

12)  Run cluster analysis and build a dendrogram using Diana. Submit the dendrogarm. (1pt)

13)  Calculate and report the cophenetic correlation coefficient.   
0.804.

14)  Calculate and report the divisive coefficient.   
0.558

15)  Based on these two coefficients, does diana do a better job representing the original distance matrix and defining cluster structure than Polythetic Agglomerative Hierarchical Clustering using average-linkage? (1pt)   
The PAHC did slightly better at capturing the original distance matrix (cophenetic coefficient of 0.88 vs. 0.804), and had more cluster structure.

\*Extra credit:

What is the name of the structure in the diana dendrogram describing the relationship between site 1, 6, and the other two large clusters in the tree (i.e., the lack of bifurcation at the base of the tree)? (1pt)

Paraphyly?