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

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

   3

3) 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)

4) Which PCA axis best separates your clusters? (1pt)

   component 1

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

   wing area, group 1 = 0.23, group 2 = 2.76, group 3 = -0.70

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)

   p = 0.000999 , expected delta = 4.11. We rejected the null model of no difference among groups, so these clusters are distinct

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)

   p = 0.000999, R = 0.5491. We rejected the null model of no difference among groups by dissimilarity ranks. Our clusters are distinct.

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)

   average linkage

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

    ward

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)

    0

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.

    0.7722845

14) Calculate and report the divisive coefficient.

    0.4584207

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

    diana does a slightly better job making clusters than average-linkage. However, the discrepancy between the correlation coefficient for average-linkage and diana is much higher than the former comparison and favors average-linkage, so average-linkage is a better approach for this data overall.

***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)