BEE 4310/6310: Multivariate Statistics for Environmental Applications
Assignment #5 (10 points)

Include responses to everything below in bold, and make sure your final assignment is well organized (i.e., numbered responses, full sentences, code attached at the end). This makes it easier to grade (and easier to give partial credit).

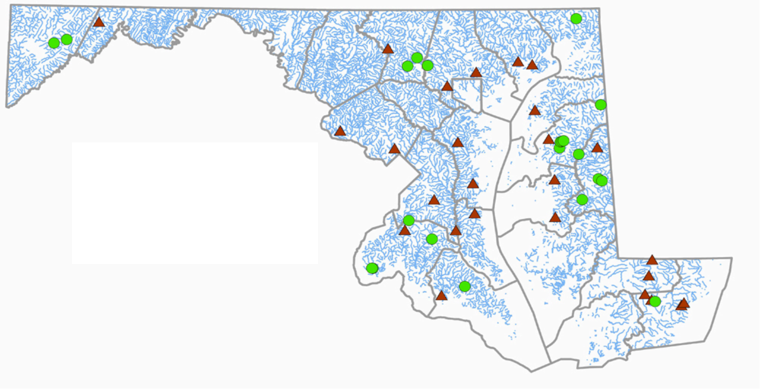

Figure 1. Favorable (green) and unfavorable (brown) brook trout streams in Maryland. From Smith and Sklarew (2012).

In this exercise, you will use clustering and classification methods to explore how you can use stream reach properties to determine whether brook trout are likely to reside within a given stream reach. You will use field data collected in Maryland between 2000-2008 with records of land use and in-stream parameters for 84 different stream reaches across the state (Fig 1, Smith and Sklarew, 2012). Note that multiple sample sites can be associated with each point in Fig 1. The reach properties include the logarithmic distance (log-meters) from the sample site to the nearest road (LOG_RD), % agricultural land cover in the watershed upstream of the sampling site (X.Ag), spring water temperature in °C (TEMP_FLD), riffle quality – a measure of suitable shallow areas in the channel for habitat - (RIFFQUAL), and dissolved oxygen in mg/L (DO_FLD). At each site, we also have a measure of brook trout presence/absence (bkt=1/0). This data can be found on Canvas in the excel file “Maryland.Trout”.

1. Hierarchical Clustering to Explore Partitioning

a) To begin, you will explore whether the reach properties show any tendency to cluster across the different reaches. If they do, this might indicate that other characteristics of those reaches, like adequate brook trout habitat, might also follow similar patterns.

   To begin, calculate a distance matrix on the scaled reach properties (X.Ag, LOG_RD,TEMP_FLD,RIFQUAL,DO_FLD) using the dist() function.

```
d <- dist(scale.covariates)
```

   Here, scale.covariates is a matrix of the data after centering and scaling each variable. Remember, this scaling is necessary so that no single variable dominates the distance measure simply because it has a large variance.

b) **(1 pt) Using the hclust() function, generate 4 different dendrograms based on the distance matrix, using four different methods for clustering groups with more than one member (“centroid”, "single linkage", “complete linkage”, and “average”). Plot these four dendrograms using the plot function.**

c) **(1 pts) Discuss the dendrograms, focusing on their interpretability in terms of chaining and reversals. Which dendrogram do you find least interpretable? Why? Which dendrogram do you find most interpretable? Why?**

d) Using the complete-linkage dendrogram and the cutree function(), cut the tree to allocate different observations to two different clusters.

e) **(2 pts) Report the mean of the five reach characteristics across the two different clusters and discuss these results. Then, present a 2X2 contingency table that shows how trout presence/absence is distributed across the clusters (hint: using the table() function for this is helpful). What can you conclude from these results?**

2. K-Means Clustering to Explore Partitioning

f) You will now attempt a similar exercise, but using K-Means clustering. Using the same scaled covariates, develop clusters for k=1,…10 groups using the K-Means function kmeans().

Set nstart=10 in order to run the algorithm 10 times and take the best result (i.e., the result that maximizes explained variance). This will smooth out any influence from the arbitrary starting positions of the groups.

g) **(1 pt) Plot the variance explained by the clusters against the number of clusters for each of the different values of K tested.** Explained variance is equal to the ratio of the between and total variations in the data, both of which are outputs of the kmeans() function. **Based on the results in the 'elbow' plot, why might K=2 be a reasonable number of clusters for further analysis?**

h) **(2 pts) Based on K=2 and similar to (e) above, show how the mean characteristics of reaches differ across the two clusters. Also similar to (e) above, present a table that shows how trout presence/absence is distributed across the clusters. What can you conclude from these results, and how do they compare to the results using the hierarchical clustering approach?**

3. Discrimination and Classification

i) The clustering methods above are a useful way to explore the data. However, now you'll need to build a model that can use stream reach properties to actually predict whether trout are present in a new reach, assuming you don't have any trout information available for that reach. You will use discrimination/classification analysis to develop this model.

First, split the original dataset into a training and testing dataset. Let the training dataset be composed of the first 60 observations, and the testing data be composed of the remaining 24 observations.

j) Next, determine the discrimination function. Split the training data into two groups, one with and one without the presence of trout. Then create variables for:

- the number of observations associated with each group.
- The pooled covariance matrix across the two groups.
- The discrimination function (i.e., alpha, α).

k) **(3 pts)** Finally, you need to determine how to classify observations in the testing dataset into one of the two groups based on whether the discriminant function values, δ, for those observations are closer to $\alpha^T \bar{x}_1$ or $\alpha^T \bar{x}_2$. **Create a 2X2 table that shows the number of correct and incorrect classifications for the testing data that fall into 1 of 4 categories (again, you can use the table() function):**

   a. 1: No trout present, and you estimate no trout present
   b. 2: No trout present, and but you estimate trout present
   c. 3: Trout present, but you estimate no trout present
   d. 4: Trout present, and you estimate trout present

   Discuss the skill of your proposed classification model in terms of accurate estimates of brook trout presence and absence and both false positives and false negatives.

Reference

Smith, A.K., and Sklarew, D. (2012), A stream suitability index for brook trout (Savelinus fontinalis) in the Mid-Atlantic United States of America, Ecological Indicators, 23, 242-249.