Practical Stats Newsletter for May 2021

Subscribe and unsubscribe: http://practicalstats.com/news
Archive of past newsletters http://practicalstats.com/news/archive.html

In this newsletter:

A. Course registrations are closed
B. Cluster analysis, 8 years later
C. In the future

A. Course registrations are closed

Registration for all of our for-payment online courses is now closed. All who have registered will continue to have complete access to those materials and support from me for one year from their signup date or until March 31, 2022, whichever is first.

Our free webinar recording 'courses' are still available at no cost to anyone for streaming. You can see the listing of all free recordings at https://practicalstats.com/training/webinar.html and the recordings themselves are at our Online Training Center, https://practicalstats.teachable.com .

B. Cluster analysis, 8 years later

The multivariate course I taught for many years never became an online course. I never wrote a textbook on it, though an applied textbook on multivariate analysis for science applications is still sorely needed. Some information from the course is available by clicking the `Multivariate` button in our News Archive (https://practicalstats.com/news/archive.html) and reading the newsletters there dealing with multivariate methods. This newsletter will supplement the March 2013 article on cluster analysis - read that first, it provides the background for today's information.

There are three types of decisions that the data analyst must make when performing cluster analysis. First is the choice of a multivariate distance (or similarity) matrix between all pairs of observations (locations, species, etc.). A simple example is Pearson's r correlation coefficient. The two observations with the smallest distance (highest correlation) make up the first cluster. There are over 50 distance measures used; ecologists have specialized distance measures to deal with counts that have true zeros and other conditions. The most common measure for continuous data is not Pearson's r but the Euclidean distance, the straight-line distance through multivariate space. This would be commonly used with a series of concentrations or other chemical/physical measurements.

A second common measure is the Bray-Curtis distance, often used with counts of organisms. Bray-Curtis standardizes the Manhattan distance measure so that nonexistant species do not distort the distance -- correlation among marine organisms should not be affected by including tigers (which even I know are not found in the ocean!) in the list of species. Bray-Curtis also gives a zero similarity between two observations that have no overlap of species. This would not be the case for the Euclidean measure and many other distance measures on the same data.

The Euclidean and Bray-Curtis distance measures are 'standard practice' for continuous chemical/physical and biological count data, respectively. Other distance measures are used for specific applications, but these two work quite well in many cases for their respective data types. If you are not sure which distance measure to use, these two are a very good start.

The second decision is whether or not to standardize or transform data prior to computing distances. Anytime distances are computed using a diverse set of measurements (concentrations, geomorphic characteristics, temperature and rainfall, etc.) the diverse scales should lead you to standardize each variable. Otherwise variables with larger variances (usually those with the largest values) will have the largest effect on the total distance. Standardization subtracts from each observation the mean of all data for that variable and divides by the standard deviation of all data for that variable, resulting in each variable have a mean of zero and standard deviation of 1. Using the standardized values when computing distance measures gives each variable the same weight in determining distances between observations. This is done prior to cluster analysis.

A similar process is often performed when both sparse and frequently-observed species are present. If species present in much higher numbers should have that proportionate weight in determining distances between clusters, use the original count data. But if sparsely populated species should have a more equal influence in determining which groupings differ from others, transform the counts of the more frequently-observed species. This is often done by taking the square root of counts when those counts are 10 to 100 times the counts of the sparse species and taking a fourth root if the difference in counts is more than a factor of 100.

The third decision is to choose the linkage scale between clusters. The linkage determines the type (shape in multidimensional space) of cluster the method looks for. Again there are 'standard practice' linkages that work well in most applications. The first is called "average link" which is the Euclidean distance between clusters divided by the number of variables used. It is the standard linkage when using both Euclidean and Bray-Curtis distance measures. A second method usually similar in results to average link is called "Ward's linkage". It minimizes the variance within clusters and is optimal for spherical clusters.

Two methods common in software that do not work well for environmental data are "single link" and "complete link" (or "farthest link"). Single link uses the distance between the closest member of each cluster as the distance between clusters. This often produces clusters that are hard to interpret, especially when the data follow a gradient (Figure 1. Also see the March 2013 newsletter for another example). Complete link defines the distance between clusters using the distance between the points farthest from each other in different clusters. This results in smaller somewhat spherical clusters giving the opposite problem from single-link: "clear" clusters that don't follow the pattern of data (Figure 2). Figures 1 to 3 show the results for single, complete, and average linkages on the same data, designed to have two groups with a small 'bridge' in between them. Note how poorly and what types of problems were produced by single and complete linkages. Stay away from using either of these linkage methods.

Standard Practices

If you haven't got a lot of experience with specific applications of cluster analysis, stick to the 'standard practices'.

1. Standardize and possibly transform chemical/physical variables before clustering. The goals are to equalize the effect of variables measured on diverse scales and produce an even spread of data across scatterplots to lower the influence of outliers.

2. For species counts, transform variables with larger numbers of counts if the goal is to equalize the influence of sparse versus abundant species.
3. Use the Bray-Curtis distance measure on species variables and the Euclidean distance for chemical/physical variables. This can be modified as you learn more about different distances measures and their impact in specific situations.
4. Use Ward's or average linkages (Figure 3) unless you know exactly what you are doing.

How Many Clusters?

Cluster analysis itself does not determine how many clusters exist in the data. The analyst for years simply had to choose the number they thought best fit the data. More recently some attempts at determining the 'significance' of distances between clusters have been developed. One is the Calinski criterion, available in the vegan package of R. It looks for the maximum sum of squares within clusters versus between clusters similar to an analysis of variance approach. A second is the Bayesian Information Criterion (BIC) which determines by maximum likelihood methods the quality of fit for each 'model' (set of clusters). The third is the SIMPROF (similarity profile) method in the clustsig package of R. It is a permutation (nonparametric) approach that compares the strength of similarity within clusters to what would be expected by chance using random permutations. I favor the latter because it is unaffected by outliers (unlike Calinski) and doesn't assume a model (unlike BIC).

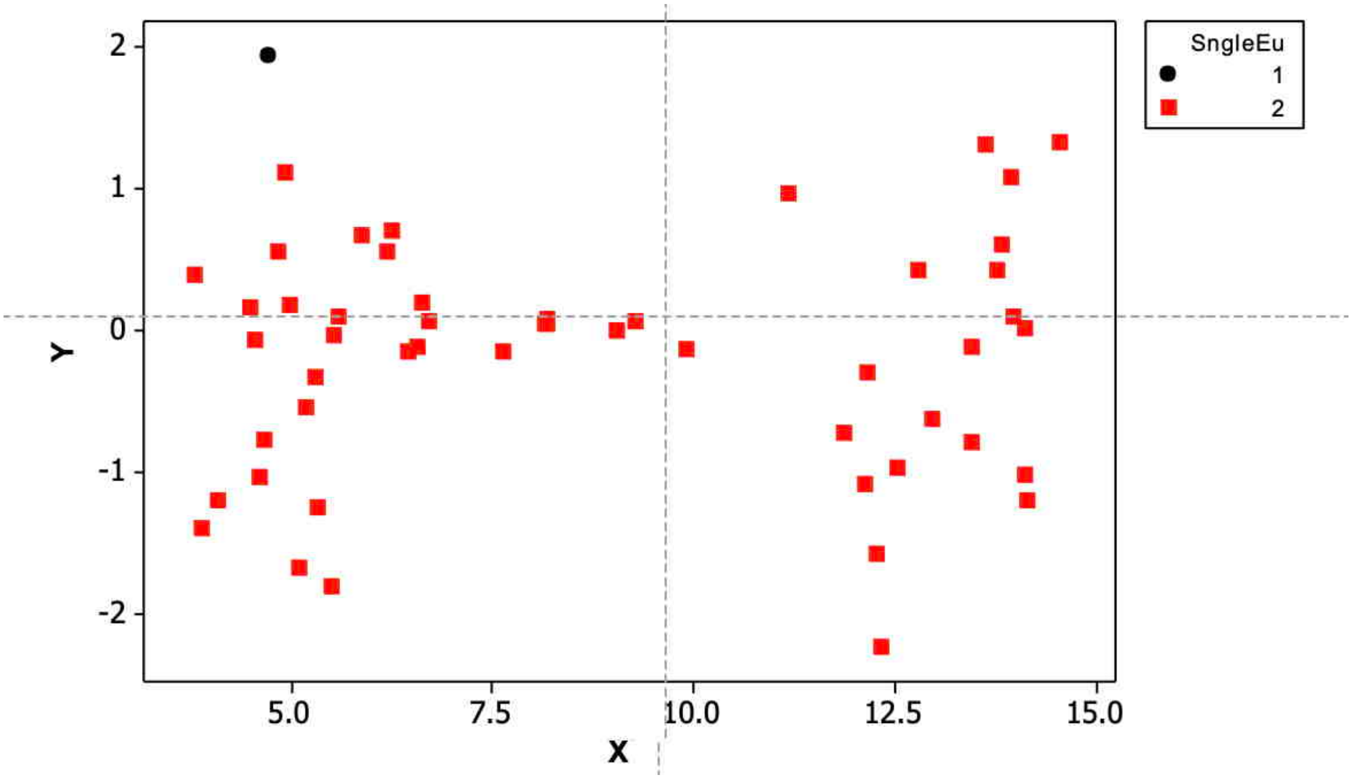

Figure 1. Single linkage Euclidean distance clustering. Does not discriminate between the two groups at all.

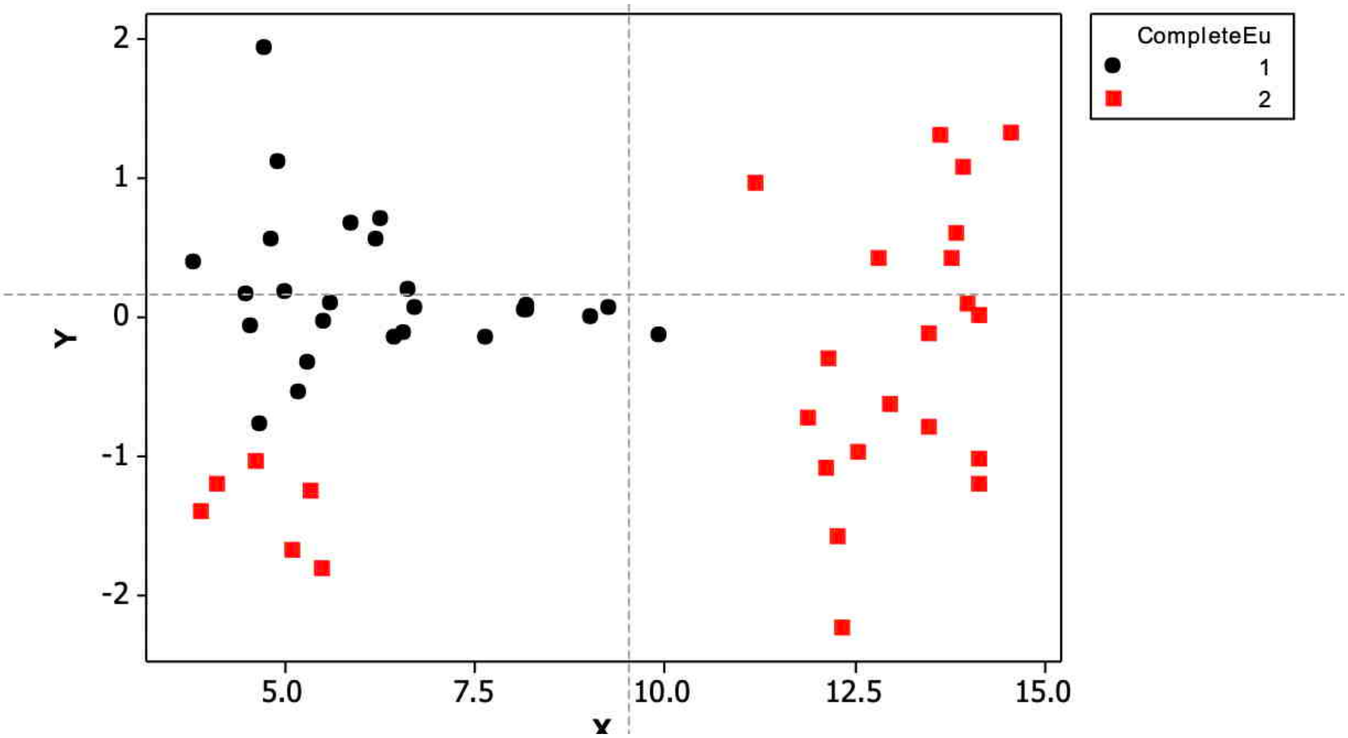

Figure 2. Complete linkage Euclidean distance clustering. Produces a circular cluster that doesn't make much sense.

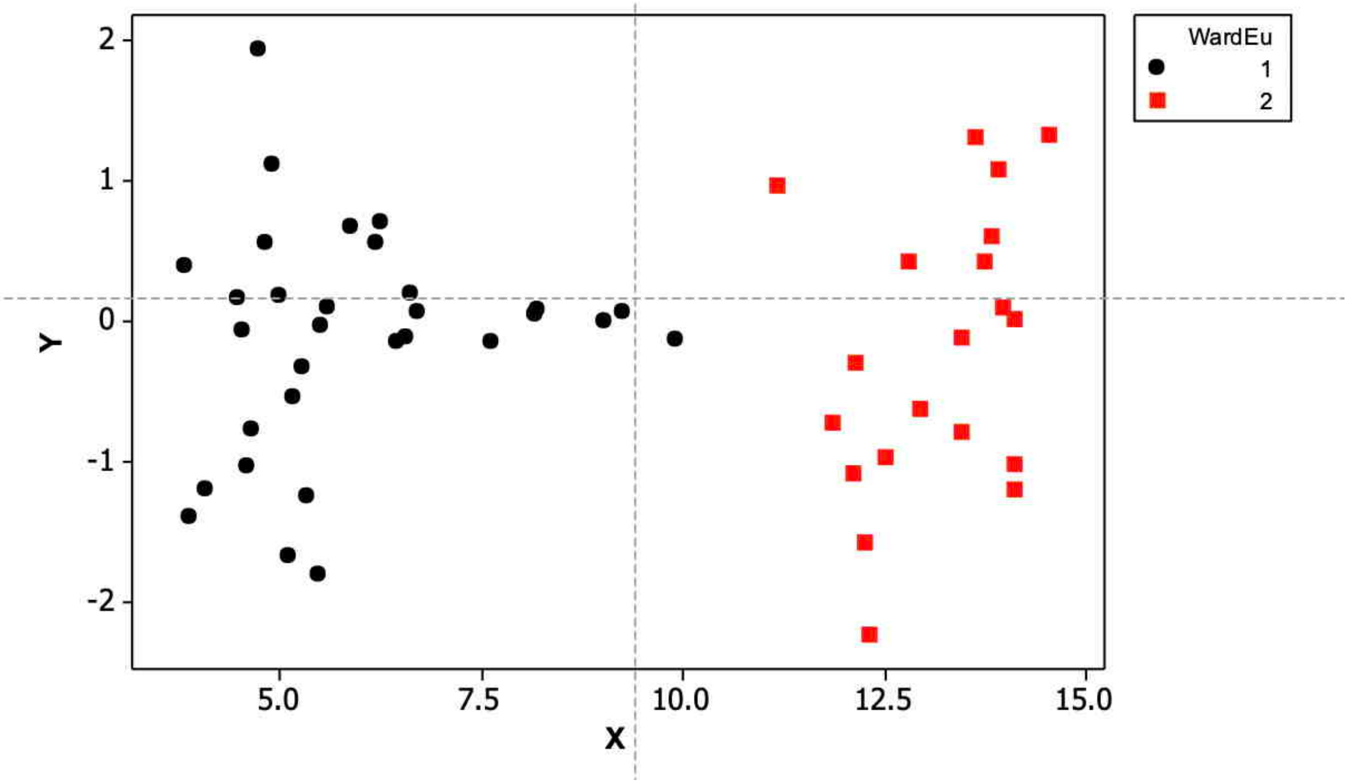

Figure 3. Ward linkage Euclidean distance clustering. Differentiates between the two groups well.

C. In the future

My plans for winding down Practical Stats is coming into focus. I expect to publish one or two more newsletters outlining the details. All free webinars can still be streamed (see part A) from their current locations until April 2022. By that time I expect to have found a more permanent location for those. All newsletters (essentially my blog) will still be available until April 2022 for downloading at https://practicalstats.com/news/archive.html .
I'm not sure they are worth keeping available after that, so download them now if you think you'll want them later.

There are alternatives to the for-payment course materials: The Applied Environmental Statistics (AES) course information is around 85% contained in the free pdf textbook "Statistical Methods in Water Resources" (2020 edition), available from the US Geological Survey at https://doi.org/10.3133/tm4A3 . A hardback book edition is also available from the USGS for under $25.

Perhaps 75% of the material contained in our Nondetects And Data Analysis (NADA) online course can be obtained by a combination of two sources: the textbook "Statistics for Censored Environmental Data Using Minitab and R" (2011), published by Wiley, and the reference manual (free pdf) for the NADA2 package for R software found at https://cran.r-project.org/package=NADA2 .

In the next newsletter I'll provide an update on what will continue online after March 2022.

'Til next time,

Dennis Helsel
ask@practicalstats.com
Practical Stats LLC
-- Make sense of your data