Practical Stats Newsletter for December 2020

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

In this newsletter:

A. Register for Our Courses by March 31

On our online training site: https://practicalstats.teachable.com/

Our two online courses will be accepting registrations through March 31, 2021. Starting April 1, new course registration will be closed. All who have registered will continue to have complete access and support for one year from their signup date. See part C for more information.

Our Nondetects And Data Analysis (NADA) course is available online. It's a complete coverage of data analysis with nondetects and 'remarked data': summary statistics, regression, group testing, trend analysis and even some multivariate methods, all without substituting fabricated numbers like ½ the detection limit. One year's access to the materials costs $795. The R scripts included provide 37 new functions to make data analysis easier, and are a step forward from the NADA package in R.

Our Applied Environmental Statistics courses cover methods from simple statistics through trend analysis. They are also an introduction to using R software, the most widely used statistics software in the world. They are available in two parts, each $650 USD for a 1-year access for one person. Or get both courses together in a bundle for $1200 USD. See our online training site at the link above.

B. Multivariate Graphs 2: Nonmetric Multidimensional Scaling (NMDS)

How can one make sense of what's going on in a multivariate maze of observations? What patterns and structures are evident? Are there groupings of data that are important? These questions can be answered by looking at two-dimensional graphs that present information available in data of many more dimensions. This month we'll look at a second technique -- NMDS. See the October 2020 newsletter for the first technique, Principal Components Analysis.

Nonmetric Multidimensional Scaling (NMDS) arranges data onto a two-dimensional plot in a way that expresses multivariate data similarities and differences. The "nonmetric" in NMDS states that unlike PCA, the plot axes are not linear combinations of the axes of original data. Instead the distances between observations on the plot are in the same rank order as distances between observations in multivariate space. Though NMDS is not a slice through multivariate space and so is less of a physical analog to hyperspace, its benefit is that the distances in all k dimensions contribute to the distances between observations on the plot. There is no loss of information for distances in front of or behind the two-dimensional plot as with PCA. Distance between observations in all directions factor into the distances along the NMDS plot axes.

NMDS is a flexible method as it can be based on any of a number of different distance metrics such as Bray-Curtis (often used for species counts) or Euclidean (often used for physical/chemical variables). These metrics compute a k-dimensional distance between every pair of observations. Distances are then ranked. A random initial set of points on the plot are realigned using an optimization function until their pairwise distances are in the same rank order as the rank distances between observations.

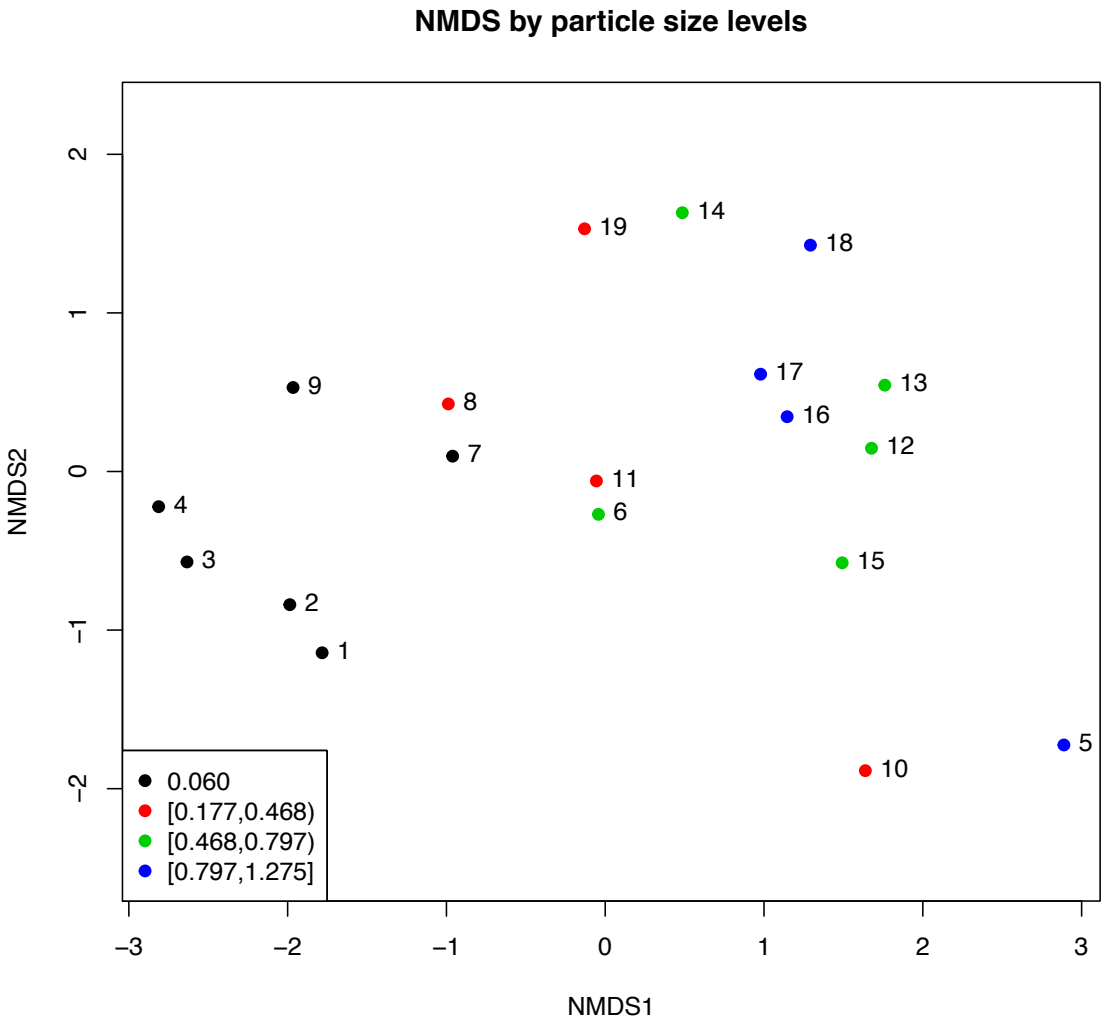

Figure 1. NMDS plot of physical/chemical characteristics at 19 sampling locations. Points are color coded by quartiles of particle size. One fourth of particle sizes are between 0 and 0.06, for example.

Figure 1 is a NMDS plot of 19 sampling locations in an estuary (Warwick, 1971) used in our October newsletter, based on their site characteristics. The site positions on the NMDS plot are quite similar to their positions on the PCA biplot computed for our October newsletter. Site numbers are shown along with colors representing the quartile levels for one of the six characteristics, particle size diameter of bed sediment. Sites 5 and 4 are the furthest apart on the plot and so are the most different in their six characteristics corporately. Sites to the left have smallest diameters (0 to 0.06) and increase in size going to the right (diameters 0.797 to 1.275). Particle size is strongly related to the NMDS1 horizontal axis. Sites 12, 13 and 15 of the 3rd quartile (green) group are to the right of sites 16 to 18 of the 4th quartile (blue) group, showing that other characteristics are also influencing position on the NMDS plot. The

second NMDS2 axis is related to an interstitial salinity gradient (Figure 2). Sites at the top of the plot have generally higher interstitial salinities (blue and green groups) than those toward the bottom.

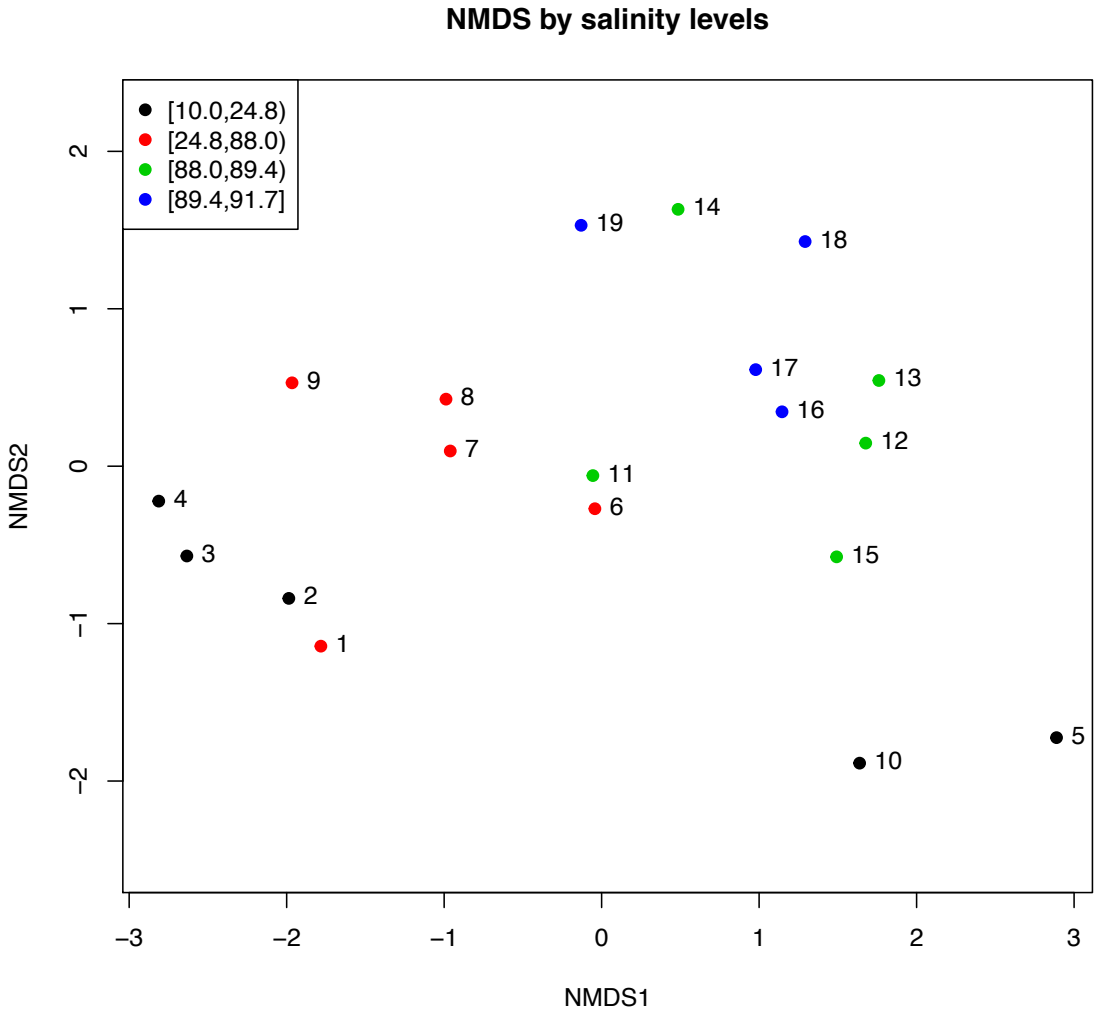

Figure 2. NMDS plot of physical/chemical characteristics at 19 sampling locations. Points are color coded by quartiles of interstitial salinity. One fourth of salinities are between 10 and 24.8, for example.

As with PCA the assignment of + and – signs to axes of an NMDS plot are arbitrary -- these may be reversed to form four equally-valid views of the plot. Different software may choose a different vantage point and so plot observations switched in algebraic signs. In addition, the scales along the components are arbitrary; not all software even shows the numeric scales. Scale values are a function of rank distances but are not directly relatable to scales of the original variables. What can be said is that two observations closer in straight line distance on the plot are closer in the multidimensional space defined by all of the variables used than are two observations farther from each other on the plot.

NMDS is not a parametric nor a linear or metric method, so there is no need to take transformations of the data except to adjust the relative weights of the variables. If the values for one variable go over several log scales (50 to 10,000) while a second variable has lesser range (50 to 100) the variable with higher range will have a greater effect on the outcome. This may be appropriate or not depending on the

situation. To equalize the effect of variables a transformation of the variable(s) with higher range by taking a log or square root, etc. can be used prior to computing distances between observations. This is a common procedure when counts of organisms are the variables used in visualizing community structure.

After reading both October's and this month's newsletters you should be better able to decide which method, NMDS or a PCA biplot, displays in two dimensions the information available in many dimensions so that groupings and relative differences can be shown. I often plot both.

Reference:
Warwick, R.M., 1971, *Nematode associations in the Exe estuary*: Journal of the Marine Biological Association of the United Kingdom, v. 51, no. 2, p. 439–454, https://doi.org/10.1017/S0025315400031908.

C. What is the March 31 Deadline All About?

As my colleague Ed Gilroy used to say, "Old statisticians never die, they just get less significant." On April 1, 2021 I will be limiting my work to only supporting the two online courses I teach, Applied Environmental Statistics (AES) and Nondetects And Data Analysis (NADA). Each attendee will continue to have access to all materials and to my coaching expertise for one year. After all students have completed their one year's access I will fully retire from my career in applied statistics. At the beginning of 2020 I stopped all consulting work. This year I have supported the two online courses, with my coauthors put the finishing touches on the modern version of our Statistical Methods in Water Resources textbook, and am currently finishing up the NADA2 package for R software. That package, containing all the routines that have been available as part of my NADA online course, will be completed and posted to the CRAN website in the first quarter of 2021.

As an instruction my goal has always been to quickly respond to students' email questions. People who register for our courses by March 31, 2021 will now have my full attention for their year's training -- no other work distractions. If you have been planning to take one of these courses, sign up before course registration closes on April 1. Please also notify your colleagues who may be considering registration.

'Til next time,

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