Reducing the species set

Of the 110 species, many occur only in one or two replicates, often as singleton individuals, so that whilst display of the whole matrix is perfectly possible, it will be cumbersome and less effective than viewing species which account for a non-negligible percent of the total number of individuals in each sample, and the Matrix display wizard dialog box default is (✓Reduce species set)>(Keep most important: 50). This concept was seen in Section 3 in the routine to **Select>Variables** >(•Use those that contribute at least 3 %), say. This would exclude any species which never (in any of the 24 samples) account for 3% or more of the total count for each sample. If that is run here it reduces the matrix to a set of 39 species. A weaker threshold criterion for elimination would be species not accounting for at least 1% of the total count somewhere, which leaves in 67 species. If phrased in terms of number of species retained, as in the option to **Select>Variables**>(•Use n-most important where n is 50) then the percentage threshold is manipulated until exactly 50 species are retained (this happens here if the % threshold is exactly 2%). This is the condition which **Matrix display** uses, and there is no flexibility to do other than change that threshold number of 50, but replicating the individual routines making up the wizard would allow a more flexible set of selection criteria, including Select>Variables>(•In at least n samples where n is). That you will need to do some species selection in large matrices is inevitable and often beneficial. It was stressed earlier that for sample analyses, all species can usually be retained (unless the resemblance measure involves a species standardisation, such as Gower or chi-squared distance) – the random nature of rare species occurrences in low numbers is given little weight in effective biological measures such as Bray-Curtis. But species analyses, defining similarities among species in their response over all samples – the idea of which was introduced in Section 5 – raise entirely different problems, and it is such species analyses that are the main topic of the remainder of this section. As seen above, in **Matrix display**, species similarities are used to cluster the species and/or re-order them in a way which optimises a seriation criterion and it is helpful if the rarer species, which cannot produce sensible assessments of similarity with other species – values will swing wildly between 0 and 100 – are deselected at the outset. Note, though, that it is an underlying principle of **Matrix display**, and thus preferably of direct **Shade Plot** runs, that where ordering/clustering of the samples is involved, it should be based on sample similarities from all species, not just those viewed in the shade plot.

Revision #3
Created 5 August 2024 21:24:22 by Arden
Updated 29 January 2025 04:51:17 by Abby Miller