3.1 Cluster analysis

The previous chapter has shown how to replace the original data matrix with pairwise similarities, chosen to reflect the particular aspect of community similarity of interest for that study (similarity in counts of abundant species, similarity in location of rare species etc). Typically, the number of pairwise similarities is large, *n(n–1)/2* for *n* samples, and it is difficult visually to detect a pattern in the triangular similarity matrix. Table 3.1 illustrates this for just part (roughly a quarter) of the similarity matrix for the Frierfjord macrofauna data *{F}*. Close examination shows that the replicates within site A generally have higher within-site similarities than do pairs of replicates within sites B and C, or between-site samples, but the pattern is far from clear. What is needed is a graphical display linking samples that have mutually high levels of similarity.

Table 3.1. Frierfjord macrofauna counts {F}. Bray-Curtis similarities, on $\sqrt{}\sqrt{}$-transformed counts, for every pair of replicate samples from sites A, B, C only (four replicate samples per site).

	A1	**A2**	**A3**	**A4**	**B1**	**B2**	**B3**	**B4**	**C1**	**C2**	**C3**	**C4**
A1	–											
A2	*61*	–										
A3	*69*	*60*	–									
A4	*65*	*61*	*66*	–								
B1	*37*	*28*	*37*	*35*	–							
B2	*42*	*34*	*31*	*32*	*55*	–						
B3	*45*	*39*	*39*	*44*	*66*	*66*	–					
B4	*37*	*29*	*29*	*37*	*59*	*63*	*60*	–				
C1	*35*	*31*	*27*	*25*	*28*	*56*	*40*	*34*	–			
C2	*40*	*34*	*26*	*29*	*48*	*69*	*62*	*56*	*56*	–		
C3	*40*	*31*	*37*	*39*	*59*	*61*	*67*	*53*	*40*	*66*	–	
C4	*36*	*28*	*34*	*37*	*65*	*55*	*69*	*55*	*38*	*64*	*74*	–

Cluster analysis (or *classification*, see footnote on terminology on page 1.2) aims to find natural groupings of samples such that samples within a group are more similar to each other, generally, than samples in different groups. Cluster analysis is used in the present context in the following ways.

a) Different sites (or different times at the same site) can be seen to have differing community compositions by noting that replicate samples within a site form a cluster that is distinct from replicates within other sites. This can be an important hurdle to overcome in any analysis; if

replicates for a site are clustered more or less randomly with replicates from every other site then further interpretation is likely to be dangerous. (A more formal statistical test for distinguishing sites is the subject of Chapter 6).

b) When it is established that sites can be distinguished from one another (or, when replicates are not taken, it is assumed that a single sample is representative of that site or time), sites or times can be partitioned into groups with similar community structure.

c) Cluster analysis of the *species* similarity matrix can be used to define species assemblages, i.e. groups of species that tend to co-occur in a parallel manner across sites.

Range of methods

Literally hundreds of clustering methods exist, some of them operating on similarity/dissimilarity matrices whilst others are based on the original data. Everitt (1980) and Cormack (1971) give excellent and readable reviews. Clifford & Stephenson (1975) is another well-established text from an ecological viewpoint.

Five classes of clustering methods can be distinguished, following the categories of Cormack (1971) .

1. *Hierarchical methods*. Samples are grouped and the groups themselves form clusters at lower levels of similarity.

2. *Optimising techniques*. A single set of mutually exclusive groups (usually a pre-specified number) is formed by optimising some clustering criterion, for example minimising a within-cluster distance measure in the species space.

3. *Mode-seeking methods*. These are based on considerations of *density* of samples in the neighbourhood of other samples, again in the species space.

4. *Clumping techniques*. The term ‘clumping’ is reserved for methods in which samples can be placed in more than one cluster.

5. *Miscellaneous techniques*.

Cormack (1971) also warned against the indiscriminate use of cluster analysis: “availability of … classification techniques has led to the waste of more valuable scientific time than any other ‘statistical’ innovation”. The ever larger number of techniques and their increasing accessibility on modern computer systems makes this warning no less pertinent today. The policy adopted here is to concentrate on a single technique that has been found to be of widespread utility in ecological studies, whilst emphasising the potential arbitrariness in all classification methods and stressing the need to perform a cluster analysis in conjunction with a range of other techniques (e.g. ordination, statistical testing) to obtain balanced and reliable conclusions.

Revision #13
Created 19 February 2022 08:49:51 by Arden
Updated 27 June 2023 23:36:21 by Arden