Ecological cluster analysis with Past

Øyvind Hammer, Natural History Museum, University of Oslo, 2011-06-26

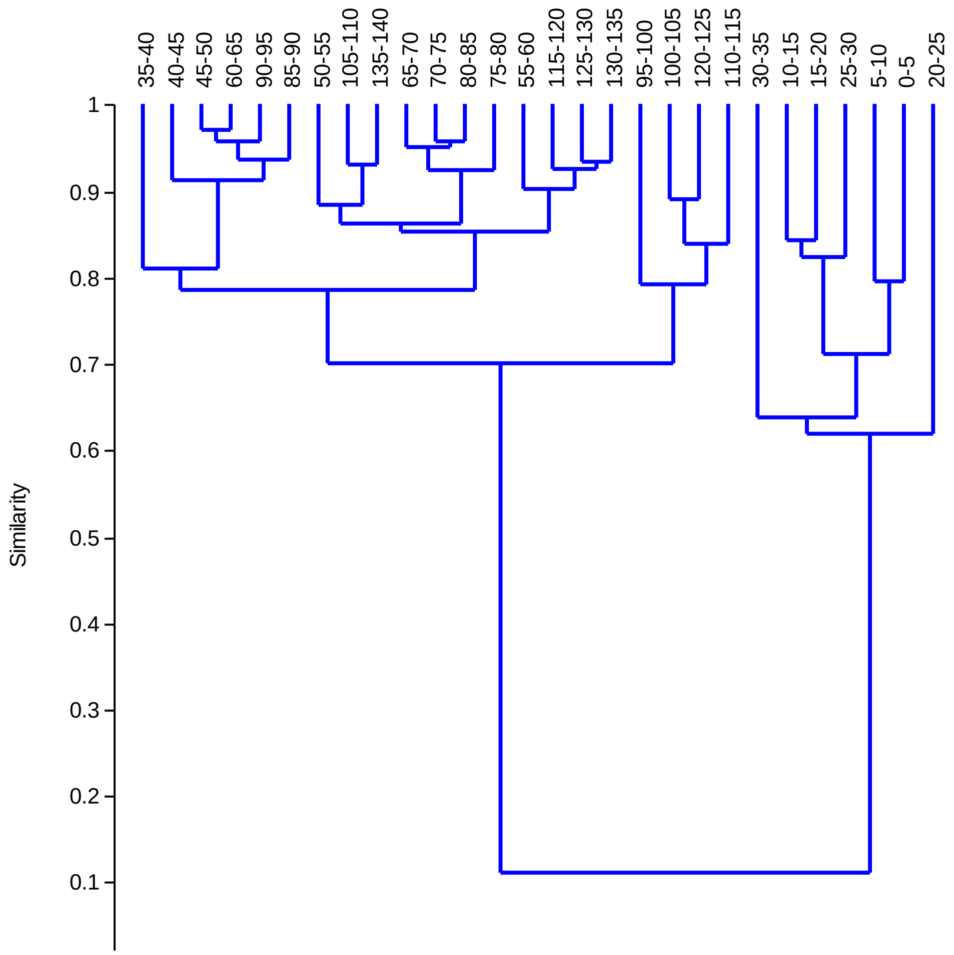
Introduction

Cluster analysis is used to identify groups of items, e.g. samples with counts or presence-absence of a number of taxa. Such groups may be interpreted in terms of biogeography, stratigraphy or environment. Cluster analysis is often criticized for imposing groups even when there are none, and it can be argued that it is better to use more "neutral" ordination techniques such as NMDS. Still, clustering is a very popular technique in paleontology.

Classical hierarchical cluster analysis

The most popular hierarchical cluster analysis methods are *agglomerative*. They are based on a similarity measure (e.g. Bray-Curtis or Dice, see the text on ordination), which must be selected by the user. The most similar pairs of samples are first joined into clusters. The most similar clusters are then joined into superclusters, and the process continues until all clusters are joined. This produces a tree called a *dendrogram*.

Open the file `LU10-05_red.dat`. This file contains counts of 17 taxa of (mainly) benthic foraminifera from a 140 cm long core in the Barents Sea. Select all, and run "Cluster analysis" in the Multivar menu.



The horizontal lines in the dendrogram are drawn at the levels of similarity between clusters. This means that well-separated clusters have long vertical "stems". In this case, the samples from 0-35 cm core depth form a distinct cluster separated from the lower part of the core.

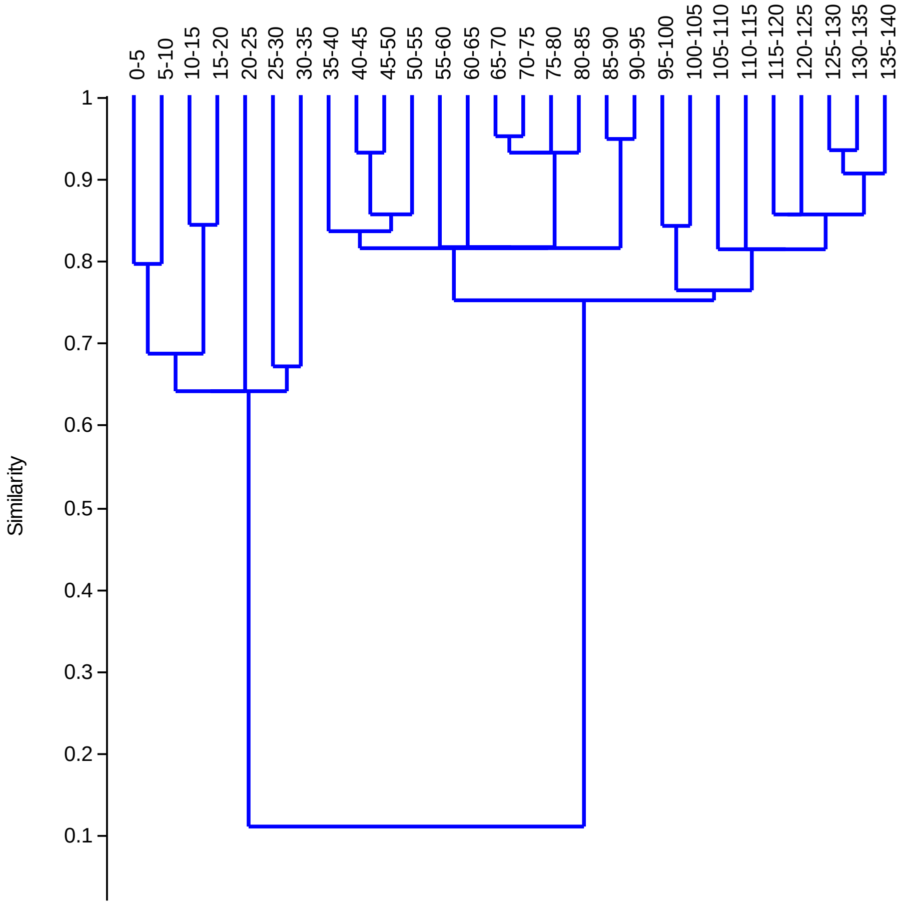
The "cophenetic coefficient" indicates how well the dendrogram reflects the structure of similarity in the original data. The value of 0.97 is very high.

Bootstrapping

Set the "Boot N" value to 1000 and press Enter. This will repeat the clustering 1000 times, based on random selections of columns. For each cluster, the percentage of random replicates where the cluster is still supported (containing the same set of taxa) is given at the root. This gives an impression of the robustness of the clusters.

Constrained clustering

It is possible to force the algorithm to only join samples that are adjacent in the data matrix. This can be useful for straigraphy or clustering along a gradient. Select the “Constrained” option in the clustering window.

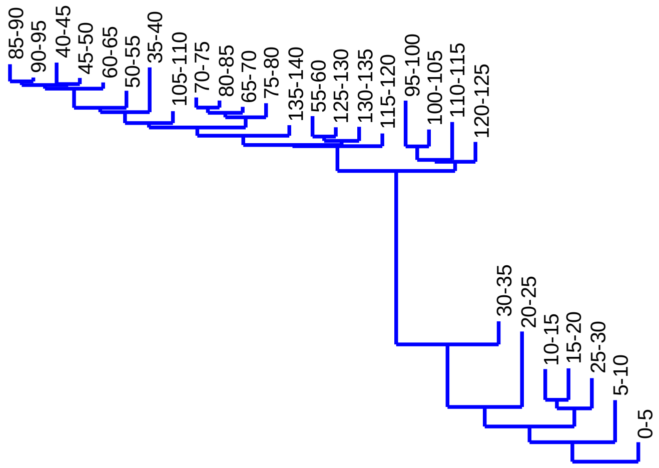


The labels above the dendrogram are now in stratigraphic order. The faunal break at 35 cm core depth is still obvious., marking the abrupt change from glacioproximal sediments to the thin Holocene cover. A smaller break may be defined at 95 cm depth, with a slightly less cold fauna below.

Neighbour joining

A curious property of the dendrogram is that from each node, the sum of branch lengths along any branch up to the top of the tree is the same (this is known as the ultrametric property). A newer method called neighbour joining (NJ) does not have this constraint. It is now a popular clustering method in molecular systematics, but may also be useful in ecology.

The NJ tree is not rooted, and can be oriented with any node at the bottom of the tree. Below is shown the NJ tree with “outgroup rooting”, i.e. with the first row in the table as the root.



K-means clustering

In K-means clustering (KmC), the number of clusters is set by the user, and the samples are assigned to these clusters without a hierarchy. In our case, we could assume the existence of three “biozones”.

The usual K-means algorithm uses Euclidean distance between samples. This is usually inappropriate for ecological data. However, using principal coordinates analysis, we can move the data points into a configuration where their Euclidean distances reflect e.g. their Bray-Curtis distances.

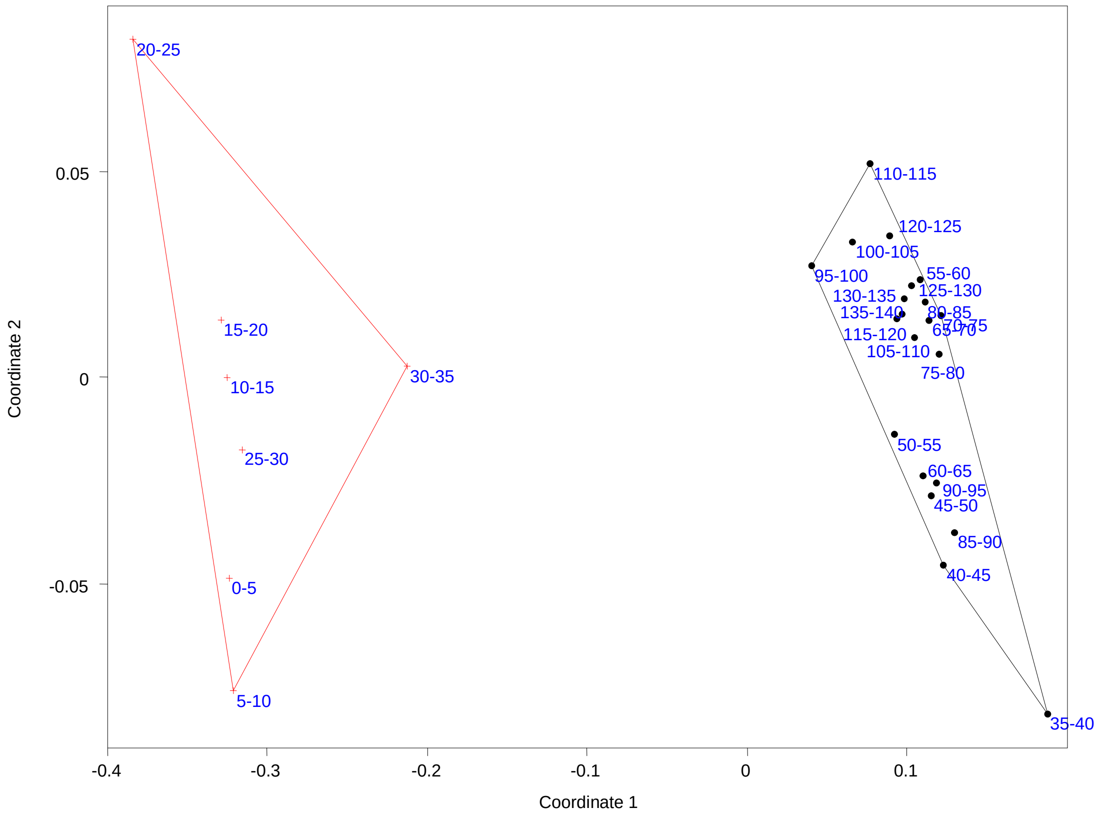
Select all, run “Principal coordinates” from the Multivar menu, and select the Bray-Curtis index. Click the “View scatter” button, and then the “View numbers” button in the scatter plot window. This gives a table of all the PCoA scores. Click the “Copy data” button (lower right). Then select the “Edit labels” box above the main spreadsheet, so that data can be pasted in from the top row. Place the cursor in the top left (empty) cell, and paste (ctrl-V). The original data are now replaced with the PCoA scores.

Select all again, and run “K-means clustering” from the Multivar menu. Specify 3 clusters. The samples are now classified into the clusters 1-3:

Item	Cluster
0-5	3
5-10	3
10-15	3
15-20	3
20-25	3
25-30	3
30-35	3

35-40	2
40-45	2
45-50	2
50-55	2
55-60	1
60-65	2
65-70	1
70-75	1
75-80	2
80-85	1
85-90	2
90-95	2
95-100	1
100-105	1
105-110	2
110-115	1
115-120	1
120-125	1
125-130	1
130-135	1
135-140	1

Finally, it is of interest to compare the clustering with ordination, here NMDS with the Bray-Curtis measure and the two main biozones marked with polygons:


Coordinate 2
Coordinate 1