

Manuscript of – Legendre, P. & H. J. B. Birks. 2012. Clustering and partitioning. Chapter 7, pp. 167-200 in: *Tracking Environmental Change using Lake Sediments, Volume 5: Data handling and numerical techniques*. H.J.B. Birks, A.F. Lotter, S. Juggins and J.P. Smol [eds.]. Springer, Dordrecht, The Netherlands. xi + 716 pp.

## 7. CLUSTERING AND PARTITIONING

PIERRE LEGENDRE (Pierre.Legendre@umontreal.ca)  
*Département de sciences biologiques,  
Université de Montréal,  
C.P. 6128, succursale Centre-ville, Montréal,  
Québec H3C 3J7, Canada.*

and

H. JOHN B. BIRKS (John.Birks@bio.uib.no)  
*Department of Biology and Bjerknes Centre for Climate Research,  
University of Bergen,  
Allégaten 41,  
N-5007 Bergen, Norway,  
and  
Environmental Change Research Centre,  
Pearson Building, Gower Street  
University College London,  
WC1E 6BT, London, UK.*

*Key words:* agglomerative clustering, constrained clustering, indicator species, two-way indicator species analysis, multivariate regression trees, partitioning

### Introduction

Hierarchical classification methods were developed as heuristic (empirical) tools to produce tree-like arrangements of large numbers of observations. The original intention of the biologists who developed them was to obtain a tree-like representation of the data, in the hope that it would reflect the underlying pattern of evolution. Hierarchical clustering starts with the calculation of a similarity or dissimilarity (= distance) matrix using a coefficient which is appropriate to the data and problem. The choice of an appropriate distance coefficient is discussed in Legendre and Birks (Chapter 7 this volume). Here we will briefly describe the algorithms most commonly used for hierarchical clustering.

Partitioning methods were developed within a more rigorous statistical frame. *K*-means partitioning, in particular, attempts to find partitions that optimise the least-squares criterion, which is widely and successfully used in statistical modelling, including regression, analysis of variance, and canonical analysis. Data may need to be transformed prior to *K*-means partitioning. This is the case, in particular, for community composition data. Please refer to Table 2 of Legendre and Birks (Chapter 7 this volume) for details of such transformations.

Palaeoecologists have long been interested in segmenting time series of data, such as sediment cores that represent depositional time-series. Several approaches have been proposed in the literature. They can all be seen as special cases of clustering or partitioning with constraints.

### Artificial example

Table 1 shows an artificial data dissimilarity and similarity matrix between 5 objects that will be used to illustrate various agglomerative clustering methods throughout this chapter. Clustering can be computed from either similarity (**S**) or dissimilarity (or distance) matrices (**D**); most software has preferences for either **S** or **D** matrices. Figure 1 shows the relationships among the five objects in the form of a two-dimensional principal

coordinate analysis (PCoA) ordination diagram (see Legendre and Birks: Chapter 7 this volume for details of PCoA). PCoA axis 1 accounts for 65.9% of the variation of the data while axis 2 accounts for 27.6%, leaving a mere 6.5% for axis 3. So the 5 data points are very well represented in 2 dimensions. We will cluster these 5 objects using various methods. Two of the interpoint distances are especially distorted in the two-dimensional ordination:  $D(1,2) = 1.389$  in the ordination instead of 1 in the original distance matrix (Table 1), and  $D(4,5) = 1.159$  instead of 2. The other pairwise distances in 2 dimensions are close to their original values (Table 1). We will see how the various clustering methods deal with these similarities or dissimilarities.

### Basic concepts in clustering

A cluster is a group of objects (observations, sites, samples, etc.) that are sufficiently similar to be recognised as members of the same group. Clustering results from an analysis of the similarities or dissimilarities among objects, calculated from the data of interest. The similarity and dissimilarity measures most commonly used by ecologists are described in Legendre and Birks (Chapter 7 this volume). A partition, such as produced by the *K*-means partitioning method, is a set of non-overlapping clusters covering the whole collection of objects in the study; some clusters may be of size 1 (singletons). Hierarchical clustering produces a hierarchy of nested partitions of objects. Numerical clustering algorithms will always produce a partition or a hierarchical clustering, whatever the data. So, obtaining a partition or a hierarchical set of partitions does not demonstrate that there are real discontinuities in the data. Most hierarchical clustering methods are heuristic techniques, producing a solution to the problem but otherwise without any statistical justification. A few methods are based on statistical concepts such as sums-of-squares.

Clustering methods summarise data with an emphasis on pairwise relationships. The most similar objects are placed in the same group, but the resulting dendrogram provides little information about among-group relationships. Ordination methods do the opposite: ordination diagrams depict the main trends in data but pairwise distances may be distorted. For many descriptive purposes, it is often valuable to conduct both forms of analysis (e.g. Birks et al. 1975; Birks and Gordon 1985; Prentice 1986).

The various potential uses of clustering and partitioning in palaeolimnology are summarised in Table 2. No attempt is made here to give a comprehensive review of palaeolimnological applications of clustering or partitioning. Emphasis is placed instead on basic concepts and on methods that have rarely been used but that have considerable potential in palaeolimnology.

Clustering with the constraint of spatial contiguity involves imposing that all members of a cluster be contiguous on the spatial map of the objects. Clustering with a one-dimensional contiguity constraint is often used on sediment cores to delineate sectors or zones where the core sections are fairly homogeneous in terms of their sediment texture, fossil composition, etc, and to identify transition zones (Birks and Gordon 1985). Cores can be seen as one-dimensional geographic (or temporal) data series, so the concept of clustering with a contiguity constraint can be applied to them. Other forms of constraint can be applied to the data to be clustered through the use of canonical analysis or multivariate regression trees.

The most simple form of clustering for multivariate data is to compute an ordination (principal component analysis (PCA), principal coordinate analysis (PCoA), correspondence analysis (CA) – see Legendre and Birks: Chapter 7 this volume), draw the points in the space of ordination axes 1 and 2, and divide the points into boxes of equal sizes. This will produce a perfectly valid partition of the objects and it may be all one needs for some purposes, such as the basic summarisation of the data. In other cases, one prefers to delineate groups that are separated from other groups by gaps in multivariate space. The clustering methods briefly described in this chapter should then be used.

## Unconstrained agglomerative clustering methods

Only the hierarchical clustering methods commonly found in statistical software will be described in this section. The most commonly used method is unweighted arithmetic average clustering (Rohlf 1963), also called UPGMA (for ‘Unweighted Pair-Group Method using Arithmetic averages’, Sneath and Sokal 1973) or ‘group-average sorting’ (Lance and Williams 1966, 1967). The algorithm proceeds by stepwise condensation of the similarity or dissimilarity matrix. Each step starts by the identification of the next pair that will cluster; this is the pair having the largest similarity or the smallest dissimilarity. This is followed by condensation of all the other measures of resemblance involving that pair, by the calculation of the arithmetic means of the similarities or dissimilarities.

The procedure is illustrated for similarities for the artificial data (Table 3). Objects 1 and 2 should cluster first because their similarity (0.8) is the highest. The similarity matrix is condensed by averaging the similarities of these 2 objects with all other objects in turn. Objects 4 and 5 should cluster during the second step because their similarity (0.6) is the highest in the condensed table. Again, the similarities of these two objects are averaged. During step 3, object 3 should cluster at  $S = 0.3$  with the group (4,5) previously formed. In UPGMA, one has to weight the similarities by the number of objects involved when calculating the average similarity:  $((1 \times 0.0) + (2 \times 0.1))/3 = 0.067$ . This *weighted average* is actually equivalent to calculating the *simple (unweighted) mean* of the 6 similarities between the objects 3, 4, and 5 in the first panel of the table:  $(0.0 + 0.0 + 0.0 + 0.0 + 0.2 + 0.2)/6 = 0.067$ . In that sense, the method is ‘unweighted’. The dendrogram representing the hierarchical clustering results is shown in Figure 2.

Weighted arithmetic average clustering (Sokal and Michener 1958), also called WPGMA (for ‘Weighted Pair-Group Method using Arithmetic averages’, Sneath and Sokal 1973), only differs from UPGMA in the fact that a *simple, unweighted mean* is computed at each step of the similarity matrix condensation. This is equivalent to giving *different weights to the original similarities* (first panel of Table 3) when condensing the similarities, hence the name “weighted”. For the data in our example, only the last fusion is affected; the similarity level of the last fusion is:  $(0.0 + 0.1)/2 = 0.05$ . Otherwise, the dendrogram is similar to that of Figure 2.

Unweighted centroid clustering (Lance and Williams 1967; UPGMC in Sneath and Sokal 1973) proceeds from a different paradigm. Imagine the objects in multidimensional space: as in UPGMA, the first two objects to cluster are chosen as the pair having the largest similarity or smallest dissimilarity or distance. Instead of averaging their similarities to all other objects, the two clustered objects are replaced by their centroid, or centre of mass, in multivariate space. This is illustrated in Figure 3a, a representation which is 2- instead of 3-dimensional. A UPGMC centroid is located at the centre of mass of all the objects that are members of a cluster.

In the weighted form of centroid clustering, called WPGMC (weighted centroid clustering, Gower 1967), a centroid is placed at the mid-point between the two objects of previously created centroids without regard for the number of objects in the cluster. Figure 3b shows the dendrogram corresponding to UPGMC of the 5 objects. The dendrogram for WPGMC only differs from that of Figure 3b by the position of the last fusion level, which is at  $S = 0.3$  instead of  $S = 0.317$ . The two forms of centroid clustering can lead to *reversals*. A reversal occurs when a later fusion occurs at a similarity value larger than that of the previous fusion. This phenomenon, which results from the geometric properties of centroid clustering, is explained in greater detail in Legendre and Legendre (1998: Section 8.6). Reversals are never large and can, most of the time, be interpreted as tri- or multi-furcations of the dendrogram represented by successive bifurcations.

Ward’s (1963) minimum-variance clustering minimises, at each agglomerative step, the sum of squared distances to the group centroids. This criterion, called “total error sum-of-squares” or TESS, is the same as used in analysis of variance and  $K$ -means partitioning. The example was calculated from a new distance matrix derived from Table 1 using the equation  $\mathbf{D} = [D/D_{\min}]$ . It is shown in Table 4 together with the matrix of squared

distances  $[(D/D_{max})^2]$  which will be used in the calculations. Ward's agglomerative clustering can be understood and computed in two different ways.

First, it can be computed in the same way as UPGMA clustering, by successive fusions of values in the matrix of squared distances  $[(D/D_{max})^2]$ . This is usually the strategy used in computer programs. The equation for the fusion of squared distances is given in textbooks describing cluster analysis, including Legendre and Legendre (1998: equation 8.10). Even though the cluster-fusion calculations are done using squared distances, it is useful to use the square roots of these fusion distances, as the scale for the dendrogram is then in the same units as the original distances (Table 1).

The second way of computing Ward's agglomerative clustering reflects the least-squares roots of the method. It is easier to understand but harder to compute; see Figure 3. The first two objects forming the first cluster are objects 1 and 2. The fusion of objects 1 and 2 produces a cluster containing unexplained variation; its value is calculated as the sum of the squared distances of objects 1 and 2 to their centroid. It turns out that this value can be computed directly from the matrix of squared distances (Table 4, top right), using the equation for error sum-of-squares (ESS):

$$ESS_k = (1/n_k) \sum D_{ij}^2 \quad (1)$$

where the values  $D_{ij}^2$  are the squared distances among the objects belonging to cluster  $k$  and  $n_k$  is the number of objects in that cluster. So for the first cluster,  $ESS_1 = 0.04/2 = 0.02$ . Since this is the only cluster formed so far, the total sum of squares is also equal to that value:  $TESS_1 = 0.02$ . To find the second cluster, the program has to search all possible fusions in turn and find the one that minimises TESS. As in UPGMA, the second cluster formed contains objects 4 and 5. The error sum of squares for that cluster is found using equation 1:  $ESS_2 = 0.16/2 = 0.08$ . Since there are now two clusters,  $TESS_2 = 0.02 + 0.08 = 0.10$ . The next cluster contains objects 3, 4, and 5. From equation 1,  $ESS_3 = (0.36 + 0.64 + 0.16)/3 = 0.38667$ . There are still only two clusters and  $TESS_3 = 0.02 + 0.38667 = 0.40667$ . The last fusion creates a cluster encompassing all five objects. ESS is found using equation 1:  $ESS_5 = 1.296$ . This is also the total sum of squares for all objects in the study,  $TESS_{max} = TESS_5 = 1.296$ .

Depending on the computer program used, the results of Ward's agglomerative clustering may be presented using different scales (Figure 4): different programs may use the fusion distance, the squared fusion distance, the total sum of squared error statistic TESS, the fraction of the variance ( $R^2$ ) accounted for by the clusters formed at each partition level, etc. (see Grimm 1987).  $R^2$  is computed as  $(TESS_{max} - TESS_k)/TESS_{max}$ .

Linkage clustering is a family of methods in which objects are assigned to clusters when a user-determined proportion (connectedness,  $Co$ ) of the similarity links has been realised. The similarities (Table 1 right) are first rewritten in order of decreasing values (or the distances, Table 1 left, in order of increasing values). Clusters are formed as the program reads the list of ordered similarities. In *single linkage agglomerative clustering*, objects are placed in groups as soon as they have formed a single similarity link with at least one member of the group. For the example data, the highest similarity value is 0.8; it creates a link between objects 1 and 2 at level  $S = 0.8$ . The next pair, (4,5), is formed at  $S = 0.6$ . The next similarity value in the ordered list is 0.4; it attaches object 3 to the (4,5) cluster at  $S = 0.4$ . Finally, there are two similarity links, (2,4) and (2,5), formed at level  $S = 0.2$ . These links connect the previously-formed cluster (1,2) to the group (3,4,5) (Figure 5a). In *complete linkage agglomerative clustering*, all possible similarity links must be formed before an object is admitted into a previously-formed cluster or two clusters can be fused. For the example (Figure 5b), the pairs (1,2) and (4,5) are formed at the same levels as in single linkage since these clusters involve a single link. Incorporation of object 3 into cluster (4,5) must wait until the two possible similarity links (3,4) and (3,5) can be formed; this happens when the similarity level drops to  $S = 0.2$  (Table 1 right). Likewise, fusion of the clusters (1,2) and (3,4,5) has to wait until the 6 similarity links (1,3), (1,4), (1,5), (2,3), (2,4), and (2,5) are formed; this only happens at  $S = 0$ . In *proportional-link linkage agglomerative clustering*, the connectedness level is set at any value between  $Co = 0$

(single linkage) and  $Co = 1$  (complete linkage). Figure 5c shows the dendrogram obtained with  $Co = 0.5$ . The pairs (1,2) and (4,5) are formed again at the same levels as in single linkage clustering since these clusters involve a single link. Incorporation of object 3 into cluster (4,5) must wait until 50% of the two possible similarity links (3,4) and (3,5) are formed; in other words, the cluster (3,4,5) is formed as soon as one of the two links is formed. Link (3,4) is formed at  $S = 0.4$ , so object 3 can cluster with objects 4 and 5 at that level. The fusion of cluster (1,2) with cluster (3,4,5) must wait until 50% of the 6 similarity links between the two clusters, or 3 links, are formed; this only happens at  $S = 0$  (Table 1 right).

All the previously-described agglomerative clustering methods, including single and complete linkage but not proportional-link linkage with  $0 < Co < 1$ , can be computed using an algorithm described by Lance and Williams (1966, 1967). Different methods are obtained by specifying different combinations of four parameters called  $\alpha$ ,  $\alpha$ ,  $\beta$ , and  $\gamma$  by these authors. The algorithm of Lance and Williams, which is described in more detail in textbooks on data analysis (including Legendre and Legendre 1998: Section 8.5), is used in many computer packages that offer agglomerative clustering. The algorithm led Lance and Williams (1966, 1967) to propose a new family of methods called *flexible clustering*. In flexible clustering,  $\alpha = \alpha = (1 - \beta)/2$  and  $\gamma = 0$ . Varying  $\beta$  in the range  $-1 \leq \beta < 1$  produces solutions with dense groups separated by long branches (when  $\beta$  is near  $-1$ ), as in complete linkage, to loosely-chained objects as in single linkage clustering (when  $\beta$  is near  $+1$ ). No reversals can occur in flexible clustering.

It is often useful to compare dendrograms to the original similarity or distance matrix in order to determine which, among several clustering methods, has preserved the original information best. To accomplish that, we need to turn the dendrograms into numbers. A cophenetic matrix is a similarity (or dissimilarity) matrix representing a dendrogram (Table 5). To construct it, one simply has to read, on the dendrogram, the  $S$ - or  $D$ -level where two objects become members of the same cluster and write that value into a blank matrix. Different measures of goodness-of-fit can be used to compare the original matrix to the cophenetic similarities or dissimilarities. The most popular indices are the *matrix correlation* (also called *cophenetic correlation*), which is Pearson's  $r$  linear correlation coefficient computed between the values in the two half-matrices of similarities or dissimilarities, and the *Gower distance* which is the sum of the squared differences between the original and cophenetic values:

$$D_{\text{Gower}} = \sum (\text{original } S_i - \text{cophenetic } S_i)^2 \quad (2)$$

For the example data, the clustering method that best represents the original information, by these criteria, is UPGMA: the matrix correlation  $r$  is 0.946 (high values are better) while the Gower distance is 0.073 (small values are better).

A different problem is that of comparing classifications to one another. One can compute a consensus index for two classifications (reviewed in Rohlf 1982; Mickevich and Platnick 1989; Swofford 1991). Alternately, one can compute a consensus tree using a choice of rules (strict consensus, majority rule, Adams consensus, etc.) summarised in Swofford (1991); another rule, called 'average consensus', was described by Lapointe and Cucumel (1997). Legendre and Lapointe (2004) also described a way of testing the congruence among dissimilarity matrices derived from data-sets containing different types of variables about the same objects. If they are congruent, the data-sets can be used jointly in statistical analysis.

### **K-means partitioning**

The  $K$ -means problem was defined by MacQueen (1967) as that of partitioning a multivariate data-set (containing  $n$  objects and  $p$  variables) in Euclidean space into  $K$  non-overlapping groups in such a way as to

minimise the sum (across the groups) of the within-group sums of squared Euclidean distances to the respective group centroids. The function to be minimised is TESS, the same function that is used in Ward's agglomerative clustering.  $K$ -means will produce a single partition of the objects, not a hierarchy. The number of groups,  $K$ , to be found is determined by the user of the method. If one asks for several values of  $K$ , the  $K$  partitions produced may not be nested.

The search for the partition that minimises TESS is done by an iterative algorithm which begins with a starting configuration and tries to optimise it by modifying the group membership.

- A starting configuration is a preliminary partition of the objects into  $K$  groups given to the program. Depending on the program being used, one may have to provide the group membership for all objects, or the positions of the  $K$  cluster centroids in  $p$ -dimensional space. If a configuration is given as a hypothesis, one can use it as the starting point; the  $K$ -means algorithm will try to optimise this configuration, in the least-squares sense, by modifying the group membership if this results in a lower value for TESS. A second method is to restart the procedure several times, e.g. 50, 100, or 1000 times, using different random assignments of the objects to the  $K$  groups or random centroids as starting configurations. There are different ways of choosing random assignments of the objects or random centroids. A third method is to conduct agglomerative clustering, cut the dendrogram into  $K$  groups, find the positions of the group centroids in  $p$ -dimensional space, and use these as the starting configuration. Hand and Krzanowski (2005) advise against the use of this method, which has proved less efficient in simulations than random starts.
- Many different algorithms have been proposed to solve the  $K$ -means problem.  $K$ -means can even be computed from distance matrices. A simple alternating least-squares algorithm, used for instance in the SAS package, iterates between two steps: (1) compute the group centroids; they become the new group seeds; (2) assign each object to the nearest seed. One may start with either step 1 or step 2, depending on whether the initial configuration is given as an assignment of objects to groups or a list of group centroids. Such an algorithm can easily cluster tens of thousands of objects.
- Note that  $K$ -means partitioning minimises the sum of squared Euclidean distances to the group centroids (TESS). The important expression is *Euclidean distance*. Many of the data tables studied by ecologists should not be directly analysed using Euclidean distances. They need to be transformed first. This topic is discussed in detail in Legendre and Birks (Chapter 7 this volume: refer to Table 2 of that chapter for a summary). Physical variables may need to be standardised or ranged, while assemblage composition data may need to be subjected to the chord, chi-square, or Hellinger transformation, prior to PCA, redundancy analysis (RDA), or  $K$ -means analysis.
- If one computed  $K$ -means partitioning for different values of  $K$ , how does one decide on the optimal number of groups? A large number of criteria have been proposed in the statistical literature to decide on the correct number of groups in cluster analysis. Fifteen or so of these criteria, including C-H (see below), are available in the "cclust" library of the R computer language. A simulation study by Milligan and Cooper (1985) compared 30 of these criteria. The best one turned out to be the Calinski-Harabasz (1974) criterion (C-H), which we will describe here. C-H is simply the  $F$ -statistic of multivariate analysis of variance and canonical analysis:

$$\text{C-H}_x = [R_x / (K - 1)] / [(1 - R_x) / (n - K)] \quad (3)$$

where  $R_x = (\text{TESS}_{\text{min}} - \text{TESS}(K)) / \text{TESS}_{\text{min}}$ .  $\text{TESS}_{\text{min}}$  is the total sum of squared distances of all  $n$  objects to the overall centroid and  $\text{TESS}(K)$  is the sum of squared distances of the objects, divided into  $K$  groups, to their groups' own centroids. One is interested to find the number of groups,  $K$ , for which the Calinski-Harabasz criterion is maximum; this corresponds to the most compact set of groups in the least-squares sense. Even though C-H is constructed like an  $F$ -statistic, it cannot be tested for significance since there are no independent data, besides those that were used to obtain the partition, to test it. Another useful criterion, also found in the "cclust" library of the R language, is the Simple Structure Index (SSI, Dolnicar et al. 1999). It multiplicatively

combines several elements which influence the interpretability of a partitioning solution. The best partition is indicated by the highest SSI value.

The artificial example is too small for  $K$ -means partitioning. Notwithstanding, if we look for the best partition in two groups ( $K = 2$ ), a distance  $K$ -means algorithm<sup>†</sup> finds a first group with objects (1,2) and a second group with (3,4,5).  $R_{k=2} = 0.686$  (Table 4, step  $k = 3$ ) so that  $C-H_{k=2} = 6.561$ .

Indices are available to compare different partitions of a set of objects. They can be used to compare partitions obtained with a given method, for example,  $K$ -means partitions into 2 to 7 groups—or partitions across methods, for example, a 7-group partition obtained by  $K$ -means to the partition obtained by UPGMA at the level where 7 groups are found in the dendrogram. They can also be used to compare partitions obtained for different groups of organisms at the same sampling sites, for instance fossil diatoms and pollen analysed at identical levels in sediment cores.

Consider all pairs of objects in turn. For each pair, determine if they are (or not) in the same group for partition 1, and likewise for partition 2. Create a  $2 \times 2$  contingency table and place the results for that pair in one of the four cells of the table (Figure 6). When all pairs have been analyzed in turn, the frequencies  $a$ ,  $b$ ,  $c$ , and  $d$  can be assembled to compute the Rand index (1971), which is identical to the simple matching coefficient for binary data:

$$\text{Rand} = (a + d) / (a + b + c + d) \quad (4)$$

The Rand index produces values between 0 (completely dissimilar) and 1 (completely similar partitions). Hubert and Arabie (1985) suggested a modified form of this coefficient. If the relationship between two partitions is comparable to that of partitions chosen at random, the corrected Rand index returns a value near 0, which can be slightly negative or positive; similar partitions have indices near 1. The modified Rand index is the most widely used coefficient to compare partitions. Birks and Gordon (1985) used the original Rand index to compare classifications of modern pollen assemblages from central Canada with the modern vegetation-landform types from which the pollen assemblages were collected in an attempt to establish how well modern pollen assemblages reflected modern vegetation types. Birks et al. (2004) compared independent classifications of modern diatom, chrysophyte cyst, and chironomid assemblages and of modern lake chemistry on Svalbard using Hill's index of similarity between classifications (Moss 1985). This index is related to Rand's index.

#### *Example: the SWAP-UK data*

The SWAP-UK data represent diatom assemblages comprising 234 taxa, from present-day surface samples from 90 lakes in England, Scotland, and Wales (Birks and Jones: Chapter 3 this volume; Stevenson et al. 1991). The diatom counts were expressed as percentages relative to the total number of diatom valves in each surface sample. This means that the counts have been transformed into relative abundances, following equation 8 of Legendre and Birks (Chapter 7 this volume), then multiplied by 100. They are thus ready for analysis using a method based on Euclidean distances.  $K$ -means partitioning was applied to the objects, with  $K$  values from 2 to 10 groups. The partition that had the highest value of the Calinski-Harabasz criterion was  $K = 5$  ( $C-H = 16.140$ ); that partition is the best one in the least-squares sense. The 5 groups comprised 20, 38, 20, 8, and 4 lakes, respectively. The 90 diatom assemblages are represented in a scatter plot of pH and latitude of the lakes; the

---

<sup>†</sup> A distance  $K$ -means algorithm had to be used here because the original data was a  $\mathbf{D}$  matrix (Table 1). Turning  $\mathbf{D}$  into a rectangular data matrix, followed by  $K$ -means partitioning, would not have yielded the same exact value for  $C-H$  because PCoA of  $\mathbf{D}$  produces negative eigenvalues (see Legendre and Birks: Chapter 7 this volume for a discussion of negative eigenvalues).

groups are represented by symbols as well as ellipses covering most of the points of each group (Figure 7). The graph shows that the five groups of lakes are closely linked to lake-water pH but not to latitude. They are not related to longitude either. This strong pH relationship reflects the overriding influence of lake-water pH on modern diatom assemblages in temperate areas (Smol 2008).

### Constrained clustering in one dimension

Palaeolimnologists have always been interested in detecting discontinuities and segmenting stratigraphical data (sediment cores), an operation called zonation in Bennett and Birks (Chapter 9 this volume). For univariate data, the operation can be conducted by eye on simple graphs, but for multivariate data like fossil assemblages, multivariate data analysis can be of help. One can, for instance, produce ordination diagrams from the multivariate data, using PCA or CA (see Legendre and Birks: Chapter 7 this volume), and detect by eye the jumps in the positions of the data points. Palaeolimnologists more often use constrained clustering, a family of methods that was first proposed by Gordon and Birks (1972, 1974) who introduced a constraint of temporal or stratigraphical contiguity into a variety of clustering algorithms to analyse pollen stratigraphical data (see also Birks and Gordon 1985; Birks 1986). The *constraint of temporal contiguity* simply means that, when searching for the next pair of objects to cluster, one considers only the objects (or groups) that are adjacent to each other along the stratigraphical or temporal series.

Several examples of zonation using this type of algorithm are given in Bennett and Birks (Chapter 9 this volume). One can use one of the stopping rules mentioned in the previous section, and in particular the Calinski-Harabasz (equation 3) and SSI criteria, to decide how many groups should be recognised in the time series.

Another approach is to use multivariate regression tree analysis (MRT), described in the last section of this chapter, to partition a multivariate data table representing a sediment core, for example, into homogeneous sections in the least-squares sense. A variable representing level numbers or ages in the core is used as the constraint. MRT finds groups of core levels whose total error sum-of-squares is minimum.

#### *Example: The Round Loch of Glenhead (RLGH) fossil data*

Another approach is the *chronological clustering* procedure of Legendre et al. (1985) who introduced a constraint of temporal contiguity into a proportional-link linkage agglomerative algorithm and used a permutation test as a stopping criterion to decide when the agglomeration of objects into clusters should be stopped. This method was applied to the RLGH fossil data, which consists of the counts of 139 diatom taxa observed in 101 levels of a Holocene sediment core from a small lake, The Round Loch of Glenhead, in Galloway, south-western Scotland (Jones et al. 1989; Birks and Jones: Chapter 3 this volume). The data series covers the past 10 000 years. Level no. 1 is the top one (most recent) while no. 101 is at the bottom of the core (oldest). The diatom counts were expressed relative to the total number of diatom valves in each level of the core. This means that the counts have been transformed into relative abundances, following equation 8 in Legendre and Birks (Chapter 7 this volume), where these data have also been analysed; principal coordinates of neighbour matrices (PCNM) analysis showed that their temporal structure was complex.

For the present example, Euclidean distances were computed among the levels, then turned into similarities using the equation  $S = 1 - D/D_{max}$ . Chronological clustering (module *Chrono* of THE R PACKAGE, Casgrain and Legendre 2004) produced 12 groups of contiguous sections, using  $Co = 0.5$  in proportional-link linkage agglomeration and the significance level  $\alpha = 0.01$  as the permutation clustering criterion (Table 6): levels 1-5, 6-14, 15-17, 18-36, 37-44, 45-53, 54-62, 63-65, 66-78, 79-90, 91-94, 95-101. Clustering was repeated on the diatom data detrended against level numbers to remove the linear trend present in the data, as described in Legendre and Birks (Chapter 7 this volume); the clustering results were identical. These 12 groups are almost



entirely compatible with the two dendrograms shown in Figure 1 in Bennett and Birks (Chapter 9 this volume); the position of a single object (level no. 90) differs. The difference is due to the use of proportional-link linkage clustering with  $Co = 0.5$  in this example, instead of CONISS (constrained incremental sum of squares (= Ward's) agglomerative clustering; Grimm 1987) or CONIIC (constrained incremental information clustering) in Bennett and Birks (Chapter 9 this volume). This partition in Table 6 will serve as the basis for indicator species analysis (see below).

MRT analysis was also applied to the RLGH fossil core data. The constraint in the analysis was a single variable containing the sample numbers 1 to 101. Cross-validation results suggest that the best division of the core was into 12 groups, but the groups differed in part from those produced by chronological clustering: levels 1-12, 13-17, 18-36, 37-44, 45-53, 54-62, 63-66, 67-81, 82-90, 91-95, 96-99, and 100-101. Only 5 division points between groups were identical in the results of MRT and chronological clustering.

### Constrained clustering in two dimensions

Caseldine and Gordon (1978) extended the concept of temporal contiguity constraints to that of spatial contiguity constraints to analyse surface pollen spectra from three transects across a bog. They showed that constraints can be applied to any data-set for which the graph-theory representation as a minimum spanning tree is such that removing any line joining pairs of adjacent samples divides the data into two connected groups (Gordon 1973). After this, time was ripe for the development of clustering procedures with the constraint of spatial contiguity, an idea that had been proposed by several other authors all at about the same time (e.g. Lebart 1978; Lefkovich 1978, 1980; Monestiez 1978; Roche 1978; Perruchet 1981; Legendre and Legendre 1984).

The constraint generally consists of a set of geographical contiguity links describing the points that are close to each other on the map. Several types of planar connection networks can be used to connect neighbouring points: for regular grids, one can choose from among different types of connections named after the movements of chess pieces (rook, bishop, king); for irregularly-spaced points, a Delaunay triangulation (Figure 9a shows an example), Gabriel graph, or relative neighbourhood graph can be used. These connection schemes are described in books on geographical statistics as well as in Legendre and Legendre (1998: Section 13.3). The connections between neighbouring objects are written as 1s in a spatial contiguity matrix (Figure 8); 0's indicate non-neighbours.

In constrained clustering, the similarity (or dissimilarity) matrix is combined with the matrix of spatial contiguity by a Hadamard product which is the cell-by-cell product of two matrices. The cells corresponding to contiguous objects keep their similarity values whereas the cells that contain 0s in the contiguity matrix contribute 0s to the constrained similarity matrix. From that point on, a regular clustering algorithm is applied: the highest value found in the constrained similarity matrix designates the next pair to cluster and the values of similarity of these two objects with all the other objects in the study are condensed in the similarity matrix (left in Figure 8), as in Table 3. The spatial contiguity matrix also has to be condensed: an object which is a neighbour of either of the two objects being clustered receives a 1 in the condensed spatial contiguity matrix. Figure 8 is a generalisation of constrained clustering in one dimension and applies to that case as well.

#### *Example: the SWAP-UK data*

The SWAP-UK data used to illustrate  $K$ -means partitioning will now be clustered with the constraint of spatial contiguity. A Delaunay triangulation (Figure 9a) was used to describe the neighbourhood relationships among lakes; the list of links was written to a file and passed to the constrained clustering program (module *Biogeo* in THE R PACKAGE, Casgrain and Legendre 2004). Figure 9b is a map showing 10 groups of lakes resulting from clustering with the constraint of spatial contiguity at level  $S = 0.648$ . Among the 90 lakes, 27 are not clustered at

that level and do not appear in Figure 9b. Contrary to the unconstrained clustering results (Figure 7), the partition is now clearly related to latitude, with most of the Scottish lakes forming a single group (empty circles). The interpretation of these constrained clustering results is unclear ecologically. The potential influence of geography and associated components of bedrock geology, climate, and land-use at the scale of the UK on modern diatom assemblages has not, to date, been explored. The idea of regionalisation or groupings of lakes with similar biological, chemical, and ecosystem properties within and between geographical regions is a topic of current research in applied freshwater science and is a research area where unconstrained and constrained clustering methods and a comparison of the resulting partitions could usefully be applied. An analysis, not of fresh waters but of Single Malt Scotch whiskies, showed that the organoleptic properties of these whiskies could be interpreted as reflecting the different geographical regions of Scotland (Lapointe and Legendre 1994). From an ecological viewpoint, there are strong theoretical reasons to hypothesise that broad-scaled geographically-structured processes may be important in controlling the structure of ecological assemblages (Legendre 1993). Much work remains to be done on the analysis of modern assemblages of diatoms and other organisms widely studied in palaeolimnology in relation to the range of possible processes that may determine their composition and structure (e.g. Jones et al. 1993; Weckström and Korhola 2001).

### Clustering constrained by canonical analysis

A more general form of constraint can be provided by canonical analysis (RDA or canonical correspondence analysis (CCA): see Legendre and Birks: Chapter 7 this volume). The idea is to extract the portion of the information of the response or biological data matrix  $\mathbf{Y}$  that can be explained by a table of explanatory or predictor variables  $\mathbf{X}$  and apply clustering or partitioning analysis to that part of the information.

Typically,  $\mathbf{Y}$  is a (fossil, recent) assemblage composition data table whereas  $\mathbf{X}$  may contain environmental, spatial, or historical variables. Figure 2 in Legendre and Birks (Chapter 7 this volume) shows that the first step of RDA is a series of multiple regressions. At the end of that step, which is also called multivariate linear regression (Finn 1974) and is available in statistical packages under that name, the fitted values of the regressions are assembled in a table of fitted values  $\hat{\mathbf{Y}}$ . RDA is obtained by applying PCA to that table, producing a matrix  $\mathbf{Z}$  of ordination scores in the space of the explanatory variables  $\mathbf{X}$ , called “Sample scores which are linear combinations of environmental variables” in the output of the CANOCO program (ter Braak and Šmilauer 2002). Computing Euclidean distances on either of these matrices,  $\hat{\mathbf{Y}}$  or  $\mathbf{Z}$ , will produce the same matrix  $\mathbf{D}$ . One can then apply cluster analysis to  $\mathbf{D}$  or  $\mathbf{S} = [1 - D/D_{max}]$ . An excellent example of combining cluster analysis with canonical correspondence analysis in plant geography and ecology to derive an integrated biogeographical zonation is given by Carey et al. (1995).

### Indicator species analysis

Indicator species represent a classical problem in ecology (Hill et al. 1975; Hill 1979). One may be interested to find indicator species for groups known a priori, for example pH classes or geographical regions, or for groups obtained by clustering. Dufrêne and Legendre (1997) developed an operational index to estimate the indicator value of each species. The indicator value of a species  $j$  in a group of sites  $k$ ,  $IndVal_j$ , is the product of the specificity  $A_j$  and fidelity  $B_j$  of the species to that group, multiplied by 100 to give percentages. *Specificity* estimates to what extent species  $j$  is found only in group  $k$ . *Fidelity* measures in what proportion of the sites of group  $k$  species  $j$  is found. The indicator value of species  $j$  is the largest value found for that species among all groups  $k$  of the partition under study:

$$IndVal_j = \max[IndVal_{kj}] \quad (5)$$

The index is maximum (100%) when individuals of species  $j$  are found at all sites belonging to a group  $k$  of the partition and in no other group. A permutation test, based on the random reallocation of sites to the various groups, is used to assess the statistical significance of  $IndVal_j$ . A significant  $IndVal_j$  is attributed to the group  $j$  that has generated this value. The index can be calculated for a single partition or for all partitions of a hierarchical classification of sites. The INDVAL program is distributed by M. Dufrêne on the site <http://biodiversite.wallonie.be/outils/indval/home.html>. The method is also available in the package PC-ORD (MjM Software, P.O. Box 129, Gleneden Beach, Oregon 97388, USA: <http://home.centurytel.net/~mjm/pcordwin.htm>).

*Example: The Round Loch of Glenhead (RLGH) fossil data*

Indicator species analysis was conducted on the 12-group partition of the RLGH fossil data obtained by chronological clustering (see above), to identify the diatom taxa that were significantly related to the groups or zones. The diatom species with significant  $IndVal_j$  values are listed in Table 6 for each group of the partition. The number of statistically significant indicator species varies from 1 (group 7) to 12 (group 5); there were 139 taxa in the study. These indicator species highlight and summarise the differences in diatom composition between the groups or zones. This approach deserves wide use in palaeolimnology because it provides a simple and effective means of identifying the biological features of each group or zone of levels (Birks 1993). It provides a more rigorous approach to detecting groups of species that characterise or are indicative of particular sediment sections than the early attempts by Janssen and Birks (1994a).

### **Two-way indicator species analysis**

Two-way indicator species analysis (TWINSpan) (Hill et al. 1975; Hill 1979) is a partitioning method that was specifically developed for the simultaneous grouping of objects and their attributes in large, heterogeneous ecological data-sets. It has been widely used by plant community ecologists but, rather surprisingly, little used in palaeolimnology or palaeoecology (Grimm 1988). It is a polythetic divisive procedure. The division of the objects is constructed on the basis of a correspondence analysis (CA) of the data (see Legendre and Birks: Chapter 7 this volume). Objects are divided into those of the negative (left) side and those on the positive (right) side on the basis of the object scores on the first CA axis. The division is at the centroid of the axis. This initial division into two groups is refined by a second CA ordination that gives greater weight to those attributes that are most associated with one side of the dichotomy. The algorithm used is complicated but the overall aim is to achieve a partitioning of the objects based on the attributes (usually species) typical of one part of the dichotomy, and hence a potential indicator of the group and its underlying ecological conditions. The process is continued for four, eight, sixteen, etc. groups. The classification of the objects is followed by a classification of the attributes and the final structured table based on this two-way classification is constructed. Details of TWINSpan, the underlying algorithm, and questions of data transformation are given by Hill (1979), Kent and Coker (1992), and Lepš and Šmilauer (2003). The computer program TWINSpan has recently been modified, converted, and updated with a user-friendly interface to run under Microsoft Windows® (WinTWINS) by Petr Šmilauer and can be downloaded from <http://www.canodraw.com>. Despite its age and complex algorithm, TWINSpan remains a very useful and robust technique for classifying very large heterogeneous data-sets containing many zero values ('absences'), keeping in mind that the method assumes the existence of a single, strong gradient dominating the data and that the divisions between neighbouring groups may not always be optimal (Belbin and McDonald 1993). A classification resulting from TWINSpan can provide a useful starting configuration for  $K$ -means partitioning, particularly of large heterogeneous data-sets. Recent palaeolimnological applications of TWINSpan

include Brodersen and Lindegaard (1997), Brodersen and Anderson (2002), and Bennion et al. (2004) (see Simpson and Hall: Chapter 16 this volume).

*Example: the SWAP-UK data*

A two-way indicator species analysis of the SWAP-UK data (90 objects by 234 taxa) was implemented using WinTWINS 2.3. Eight pseudospecies were used with cut-levels of 0, 1, 2, 4, 8, 16, 32, and 64% (see Kent and Coker (1992) or Lepš and Šmilauer (2003) for an explanation of pseudospecies or conjoint coding). The classification into four groups and the associated indicator species are summarised in Figure 10. The four groups of lakes differ in their pH values, just as the 5 groups in the *K*-means partitioning do, with group medians of pH 5.0, 5.3, 6.3, and 6.8.

### **Multivariate regression trees**

Multivariate regression trees (MRT) produce a clustering of multivariate biological or ‘response’ data using a monothetic divisive approach, as explained below. The method combines data exploration and data interpretation (forecasting). MRT is related to regression in the sense that the explanation of the response data involves explanatory variables. It thus represents an alternative to multivariate explanatory methods such as RDA and CCA and it is part of the family of classification and regression trees discussed by Simpson and Telford (Chapter 8 this volume). MRT is a least-squares method, but it does not use simple or multiple regression.

Monothetic divisive classification methods base each split on a single variable. For each branching point of the tree, MRT chooses one of the explanatory variables, and a splitting point along it, that maximises the separation of two daughter groups in the multivariate space of the response variables (e.g. species assemblages). Group separation, or homogeneity, is maximum when the total sum-of-squares error statistic, called TESS in Ward’s clustering and *K*-means partitioning (see above), is minimised. MRT can be seen as a form of constrained clustering, the constraint being given by the environmental variables characterising each division step. The method was proposed by De’ath (2002) as an extension of univariate regression trees (Brieman et al. 1984; Simpson and Telford: Chapter 8 this volume) who also provided an R-language library, MVPART (De’ath 2007), implementing the method. Besides its emphasis on interpretation and forecasting, MRT is well-suited for the analysis of unbalanced ecological data (groups of different sizes), data containing missing values, or explanatory variables that are not necessarily related to the species in a linear or unimodal way. Davidson (2005) presents a palaeolimnological example of the use of MRT to relate cladoceran assemblage relationships to zooplanktivorous fish density and to submerged macrophyte abundance in shallow lakes in England and Denmark.

The result of MRT analysis is a hierarchical classification of the data represented by a tree, plus information about the explanatory (environmental) variables that best explain each split and the distribution of the response variables, which are often species in ecological applications, in each terminal group.

A somewhat similar approach is that of ter Braak (1986). This method starts with a hierarchical classification based on the species composition of the objects (in his case derived by two-way indicator species analysis, TWINSpan, Hill 1979) and finds by simple discriminant functions (Hill 1977) and accompanying environmental or predictor variables for the same set of objects those environmental variables (expressed as presence/absence, qualitative or nominal variables, ranks, quantitative, etc.) that optimally predict the classification of the objects into two groups, four groups, etc. Simple discriminant functions are constructed in a very simple way (Hill 1977, 1979). A presence/absence environmental variable is a possible group indicator if its frequency of occurrence is higher in the group than in the alternative group. The *n* variables with the highest absolute difference in frequency of occurrence are included in the discriminant function, where *n* is the smallest integer that minimises

the number of misclassifications. The great advantage of Hill's (1977, 1979) simple discriminant functions is their simplicity: the sign of a variable is given the same sign as that of the frequency difference, and the number of possible sets of indicator variables is restricted by ordering the variables on the basis of the absolute frequency difference. These restrictions avoid the need for optimisation by linear programming and facilitate ease of interpretation of the simple discriminants constructed (ter Braak 1986). Hill's simple discriminants make no assumptions about species–environmental relationships or about the underlying nature of the data. They simply consider if environmental variables differ in their frequency between groups. They are close in concept to Gower's (1974) maximal predictive classification as the emphasis is on classification prediction. With quantitative environmental variables, linear discriminant analysis or multiple discriminant analysis (= canonical variates analysis) can be used to discriminate between groups. For nominal environmental variables, correspondence analysis could be applied to a  $2 \times c$  table where the rows correspond to the two branches of the node and  $c$  is the total number of categories of the nominal variables. Linear discriminant functions may, however, be difficult to interpret and as one moves down the classification hierarchy the groups may contain so few objects that the coefficients of the discriminant functions cannot be reliably estimated, if at all. Using presence/absence data throughout for his simple discriminant functions, Hill (1977) circumvented these problems by proposing simple discriminant functions in which the coefficients can only take three values;  $-1$  and  $+1$  for variables that are characteristic for one group or another, and  $0$  for non-discriminating variables. Such functions are easy to interpret and both quantitative and nominal variables can easily be incorporated into this approach after careful recoding using disjoint or conjoint coding (ter Braak 1986). Janssen and Birks (1994b) applied this approach to stratigraphical pollen and plant macrofossil assemblages to detect predictor or indicator variables for constrained classifications of stratigraphical samples.

*Example: the SWAP-UK data*

A MRT was computed for the SWAP-UK data (90 sites, 234 diatom taxa) using the MVPART function (MVPART library) written by G. De'ath in the R computer language. Only three explanatory variables were available for these data: pH, latitude, and longitude of the lakes. Cross-validation is available in De'ath's function; it can be used to select the tree having the smallest predicted mean square error, or some other tree slightly longer or shorter than that. The cross-validated relative error criterion indicated that the partition in two groups (57 and 33 lakes, respectively) explained by pH was statistically the best; the Calinski-Harabasz criterion led to the same conclusion. For illustrative purposes only, we present a more developed tree containing 9 binary partitions and 10 terminal groups (Figure 11) along a scale representing the proportion of the species variation explained by each partition ( $R^2$ , as in Figure 4). pH is clearly the most important explanatory variable; this was also the case in the unconstrained partitions shown in Figure 7 and Figure 10. The mean value of each taxon in each group is available in De'ath's function output; it is not presented for this example because the taxa are too numerous. This example is presented here simply to illustrate the potential value of MRT in analysing modern sets of biological and associated environmental data, which are now such an important part of palaeolimnological research (Smol 2008). Multivariate regression trees provide a very powerful means of exploring complex biological – environmental relationships (De'ath 2002) and are of obvious wide application in a range of palaeolimnological problems. Simpson and Telford (Chapter 8 this volume) discuss the range of classification and regression trees of potential value in palaeolimnology.

Simple discriminants based on pH, latitude, and longitude coded as ranks and converted into +/- variables based on quartiles (ter Braak 1986) applied to the two-way indicator species analysis of the SWAP-UK data (Figure 10) similarly show the over-riding importance of pH in discriminating between the four TWINSpan groups, but with some influence of latitude apparent at the four-group level, in particular between the two left-hand groups on Figure 10.

## Summary

Hierarchical clustering methods and partitioning techniques such as *K*-means partitioning and two-way indicator species analysis are useful tools for summarising group structure within large, complex, multivariate data-sets that are increasingly common in palaeolimnology. The incorporation of one- or two-dimensional constraints in the clustering algorithms provides means of exploring group structure in temporal, stratigraphical data and in geographical modern data, respectively. Indicator species analysis with its associated permutation tests is a simple and effective means of detecting statistically significant indicator species for any grouping of a set of objects. The newly developed approach of multivariate regression trees combines partitioning and data exploration with regression and data interpretation and modelling.

Palaeolimnologists have largely concentrated on the use of classical ordination and canonical ordination techniques to explore patterns in their data. Modern classification and partitioning techniques along with indicator species analysis and multivariate regression trees are attractive and useful complementary tools for exploring and summarising large, complex, palaeolimnological data-sets (see Table 2). They deserve wider use than they currently receive within the palaeolimnological research community. Agglomerative clustering and *K*-means partitioning are available in most statistical software packages, as well as in the R computer language (functions HCLUST and KMEANS). Multivariate regression trees can be computed using the MVPART function of R.

## References

- Belbin L, McDonald C (1993) Comparing three classification strategies for use in ecology. *J Veg Sci* 4:41-34
- Bennett KD, Birks HJB This volume. Chapter 9 Analysis of stratigraphical data. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2009. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht
- Bennion H, Fluin J, Simpson GL (2004) Assessing eutrophication and reference conditions for Scottish freshwater lochs using subfossil diatoms. *J Appl Ecol* 41:124-138
- Birks HJB (1986) Numerical zonation, comparison and correlation of Quaternary pollen-stratigraphical data. In: Berglund BE (ed) Handbook of Holocene Palaeoecology and Palaeohydrobiology. Wiley, Chichester, pp.743-774
- Birks HJB (1993) Quaternary palaeoecology and vegetation science – current contributions and possible future developments. *Rev Palaeobot Palynol* 79:153-177
- Birks HJB, Gordon AD (1985) Numerical methods in Quaternary pollen analysis. Academic Press, London
- Birks HJB, Jones, VJ This volume. Chapter 3 Data-sets. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2009. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht
- Birks HJB, Monteith DT, Rose NL, Jones VJ, Peglar SM (2004) Recent environmental change and atmospheric contamination on Svalbard as recorded in lake sediments – modern limnology, vegetation, and pollen deposition. *J Paleolimnol* 31:411-431
- Birks HJB, Webb T III, Berti AA (1975) Numerical analysis of pollen samples from central Canada: a comparison of methods. *Rev Palaeobot Palynol* 20:133-169
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Chapman & Hall, New York
- Brodersen KP, Anderson NJ (2002) Distribution of chironomids (Diptera) in low arctic West Greenland lakes: trophic conditions, temperature and environmental reconstruction. *Freshw Biol* 47:1137-1157
- Brodersen KP, Lindegaard C (1997) Significance of subfossil remains in classification of shallow lakes. *Hydrobiologia* 342/343:125-132
- Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. *Comm Stat* 3:1-27
- Carey PD, Preston CD, Hill MO, Usher MB, Wright SM (1995) An environmentally defined biogeographical zonation of Scotland designed to reflect species distributions. *J Ecol* 83:833-845
- Caseldine CJ, Gordon AD (1978) Numerical analysis of surface pollen spectra from Bankhead Moss, Fife. *New Phytol* 80:435-453
- Casgrain P, Legendre P (2004) The R Package for Multivariate and Spatial Analysis, Version 4.0 (development release 7) – User's Manual. Département de sciences biologiques, Université de Montréal. 148 pp. Manual and program available from <http://www.bio.umontreal.ca/casgrain/en/labo/R/index.html>
- Davidson T (2005) Zooplankton ecology and palaeoecology in nutrient enriched shallow lakes. PhD thesis, University College London
- De'ath G (2002) Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology* 83:1105-1117
- De'ath G. (2007). MVPART: multivariate partitioning. R package version 1.2-6. URL <http://cran.r-project.org/>.
- Dolnicar S, Grabler K, Mazanec JA (1999) A tale of three cities: perceptual charting for analyzing destination images. In: Woodside A (eds) Consumer psychology of tourism, hospitality and leisure. CAB International, New York, pp.39-62
- Dufrêne M, Legendre P (1997) Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol Monogr* 67:345-366
- Finn JD (1974) A General Model for Multivariate Analysis. Holt, Rinehart & Winston, New York
- Gordon AD (1973) Classification in the presence of constraints. *Biometrics* 29:821-827
- Gordon AD, Birks HJB (1972) Numerical methods in Quaternary palaeoecology. I. Zonation of pollen diagrams. *New Phytol* 71:961-979

- Gordon AD, Birks HJB (1974) Numerical methods in Quaternary palaeoecology. II. Comparison of pollen diagrams. *New Phytol* 73:221-249
- Gower JC (1974) Maximal predictive classification. *Biometrics* 30:643-654
- Grimm, E.C. 1987. CONISS: A FORTRAN 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. *Comp Geosci* 13:13-35
- Grimm EC (1988) Data analysis and display. In: Huntley B, Webb III T (eds) *Vegetation History*. Kluwer Academic Publishers, Dordrecht, pp.43-76
- Hand DJ, Krzanowski WJ (2005) Optimising *k*-means clustering results with standard software packages. *Computational Stat & Data Anal* 49:969-973
- Hill MO (1977) Use of simple discriminant functions to classify quantitative phytosociological data. In: Diday E, Lebart L, Pagés JP, Tomassone R (eds) *First International Symposium on Data Analysis and Informatics Volume 1*:181-189. Institut de Recherche d'Informatique et D'Automatique, Le Chesnay
- Hill MO (1979) TWINSpan – a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of individuals and attributes. Cornell University, Ithaca, New York, USA
- Hill MO, Bunce RGH, Shaw MW (1975) Indicator species analysis, a divisive polythetic method of classification and its application to a survey of native pinewoods of Scotland. *J Ecol* 63:597-613
- Hubert LJ, Arabie P (1985) Comparing partitions. *J Classification* 2:193-218
- Janssen CR, Birks HJB (1994a) Recurrent groups of pollen types in time. *Rev Palaeobot Palynol* 79:153-177
- Janssen CR, Birks HJB (1994b) Examples of recurrent groups of pollen and macrofossils in space and time delimited by simple discriminant functions. *Diss Bot* 234:337-351
- Jones VJ, Juggins S, Ellis-Evans JC (1993) The relationship between water chemistry and surface sediment diatom assemblages in maritime Antarctic lakes. *Antarc Sci* 5:339-348
- Jones VJ, Stevenson AC, Battarbee RW (1989) Acidification of lakes in Galloway, southwest Scotland: a diatom and pollen study of the post-glacial history of The Round Loch of Glenhead. *J Ecol* 77:1-23
- Kent M, Coker PD (1992) *Vegetation Description and Analysis – A Practical Approach*. Wiley, Chichester
- Lance GN, Williams WT (1966) A generalized sorting strategy for computer classifications. *Nature* 212:218
- Lance GN, Williams WT (1967) A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer J* 9:373-380
- Lapointe F-J, Cucumel G (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Syst Biol* 46:306-312
- Lapointe F-J, Legendre P (1994) A classification of pure malt Scotch whiskies. *Appl Stat* 43:237-257
- Lebart L (1978) Programme d'agrégation avec contraintes (C. A. H. contiguïté). *C Anal Données* 3:275-287
- Lefkovich LP (1978) Cluster generation and grouping using mathematical programming. *Math Biosci* 41:91-110
- Lefkovich LP (1980) Conditional clustering. *Biometrics* 36:43-58
- Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology* 74:1659-1673
- Legendre P, Birks HJB This volume. Chapter 7 From classical to canonical ordination. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2009. *Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques*. Springer, Dordrecht
- Legendre P, Lapointe F-J (2004) Assessing the congruence among distance matrices: single malt Scotch whiskies revisited. *Austral New Zeal J Stat* 46:615-629
- Legendre P, Legendre L (1998) *Numerical Ecology*. 2nd English edn. Elsevier, Amsterdam
- Legendre P, Legendre V (1984) Postglacial dispersal of freshwater fishes in the Québec peninsula. *Can J Fish Aquat Sci* 41:1781-1802
- Legendre P, Dallot S, Legendre L (1985) Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton. *Am Nat* 125:257-288
- Lepš J, Šmilauer P (2003) *Multivariate analysis of ecological data using CANOCO*. Cambridge University Press, Cambridge
- MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J (eds) *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. University of California Press, Berkeley, pp.281-297
- Mickevich MF, Platnick NI (1989) On the information content of classifications. *Cladistics* 5:33-47
- Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50:159-179
- Monestiez P (1978) Méthodes de classification automatique sous contraintes spatiales. In: Legay JM, Tomassone R (eds) *Biométrie et écologie*. Inst nat Rech agronomique, Jouy-en-Josas, pp.367-379
- Moss D (1985) An initial classification of 10-km squares in Great Britain from a land characteristic data bank. *Appl Geogr* 5:131-150
- Perruchet C (1981) Classification sous contrainte de contiguïté continue. In: *Classification automatique et perception par ordinateur*. Séminaires de l'Institut national de Recherche en Informatique et en Automatique (C 118), Rocquencourt, pp.71-92
- Prentice IC (1986) Multivariate methods for data analysis. In: Berglund BE (ed). *Handbook of Holocene Palaeoecology and Palaeohydrology*. Wiley, Chichester, pp.775-797
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 66:846-850
- Roche C (1978) Exemple de classification hiérarchique avec contrainte de contiguïté. Le partage d'Aix-en-Provence en quartiers homogènes. *C Anal Données* 3:289-305
- Rohlf FJ (1963) Classification of *Aedes* by numerical taxonomic methods (Diptera: Culicidae). *Ann Entomol Soc Am* 56:798-804
- Rohlf FJ (1982) Consensus indices for comparing classifications. *Math Biosci* 59:131-144
- Simpson GL, Hall RI This volume. Chapter 16 Human impacts – applications of numerical methods to evaluate surface-water acidification and eutrophication. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2009. *Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques*. Springer, Dordrecht
- Simpson GL, Telford RJ This volume. Chapter 8 Classification and regression trees. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2009. *Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques*. Springer, Dordrecht
- Smol JP (2008) *Pollution of lakes and rivers – A paleoenvironmental perspective*. Blackwell, Oxford
- Sneath PHA, Sokal RR (1973) *Numerical taxonomy – The principles and practice of numerical classification*. Freeman, San Francisco

- Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 38:1409-1438
- Stevenson AC, Juggins S, Birks HJB, et al. (1991) The Surface Waters Acidification Project Palaeolimnology Programme: modern diatom/lake-water chemistry data-set. Ensis Publishing, London
- Swofford DL (1991) When are phylogeny estimates from molecular and morphological data incongruent? in: Miyamoto MM, Cracraft J (eds) *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, Oxford, pp.295-333
- ter Braak CJF (1986) Interpreting a hierarchical classification with simple discriminant functions: an ecological example. In: Diday E, et al. (eds) *Data Analysis and Informatics 4*. North Holland, Amsterdam, pp.11-21
- ter Braak CJF, Šmilauer P (2002) *CANOCO Reference Manual and Canodraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5)*. Microcomputer Power, Ithaca, New York, USA
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58:236-244
- Weckström J, Korhola A (2001) Patterns in the distribution, composition, and diversity of diatom assemblages in relation to ecoclimatic factors in Arctic Lapland. *J Biogeogr* 28:31-45



*Table 1.* Matrices showing the dissimilarity (or distance, **D**, on the left) and similarity (**S**, on the right, with  $S_i = 1 - D_i/D_{max}$ ) relationships among 5 objects, numbered 1 to 5 (artificial data). In each matrix, the upper-triangular portion (not shown) is symmetric to the lower-triangular. In the distance matrix (**D**), the main diagonal (italics) contains 0's, whereas it contains 1's in the similarity matrix (**S**).

<b>D</b>	1	2	3	4	5
1	<i>0</i>				
2	1	<i>0</i>			
3	5	5	<i>0</i>		
4	5	4	3	<i>0</i>	
5	5	4	4	2	<i>0</i>

<b>S</b>	1	2	3	4	5
1	<i>1</i>				
2	0.8	<i>1</i>			
3	0.0	0.0	<i>1</i>		
4	0.0	0.2	0.4	<i>1</i>	
5	0.0	0.2	0.2	0.6	<i>1</i>

Table 2. Palaeolimnological uses of clustering and partitioning techniques (abbreviations are explained below and in the main text).

**Modern biological assemblages (e.g. diatoms, chironomids)**

- Detect groups of samples with similar biological composition - AHC, TWINSPAN, *K*-means
- Detect groups of samples with similar biological composition along a single environmental gradient (e.g. altitude, pH) - CC1
- Detect groups of geographically contiguous samples with similar biological composition – CC2
- Detect indicator species for groups of samples - ISA, TWINSPAN
- Comparison of groupings based on different groups of organisms - AHC, TWINSPAN, or *K*-means followed by CC
- Detect groups of samples that can be overlain on an ordination of the same samples - AHC, TWINSPAN, *K*-means

**Modern environmental data (e.g. lake-water chemistry)**

- Detect groups of lakes with similar environmental variables - AHC, *K*-means
- Detect groups of geographically contiguous lakes with similar environmental variables – CC2
- Comparison of groupings based on different types of environmental variables - AHC or *K*-means followed by CC
- Detect groups of samples that can be overlain on an ordination of the same samples - AHC, *K*-means

**Fossil biological assemblages (e.g. diatoms, chironomids)**

- Detect groups of samples (zones) with similar fossil composition - CC1, MRT
- Detect indicator species for groups of samples (zones) with similar fossil composition - ISA
- Comparison of groups of samples (zones) based on different fossil groups - CC1, MRT, CC
- Detect groups of samples (zones) with similar fossil composition that can be overlain on an ordination of the same samples - CC1, MRT

**Down-core non-biological data (e.g. geochemistry, magnetics)**

- Detect groups of samples with similar geochemical composition, magnetic properties, etc - CC1, MRT
- Comparison of clusterings of samples based on different set of variables - CC1, MRT followed by CC
- Detect groups of samples (zones) with similar chemical composition or magnetic properties that can be overlain on an ordination of the same samples - CC1, MRT

**Modern and fossil biological assemblages (e.g. diatoms, chironomids)**

- Detect similarities between modern and fossil samples ('analogues') as an aid to interpreting fossil samples - AHC, TWINSPAN, *K*-means

**Modern biological assemblages and modern environmental data (e.g. diatoms and lake-water chemistry)**

- Detect environmental variables that characterise clusters or partitions of modern samples - MRT, DA
- Integrated clustering of sites on the basis of both biological and environmental data - CCC followed by *K*-means
- Comparison of groups of samples of similar biological composition with groups of lakes based on environmental variables - AHC, TWINSPAN, or *K*-means followed by CC

**Modern biological assemblages, modern environmental data, and fossil biological assemblages (e.g. diatoms and lake-water chemistry)**

- Detect groups of similar modern and fossil samples on basis of similar composition and then relate the groups of modern samples to modern environmental variables - TWINSPAN with fossil samples declared passive in WinTwins, followed by DA using modern environmental data of TWINSPAN modern groups

**Fossil biological assemblages and palaeoenvironmental variables (e.g. diatoms, occurrences of volcanic tephtras, stable-isotope data)**

- Relate fossil assemblage changes to palaeoenvironmental variables - MRT or CC1 of biological data followed by DA
- Detect groups of samples with similar fossil composition ('zones') and palaeoenvironmental variables - CCC followed by CC1

**Fossil biological assemblages from many sites**

- Detect groups of sites with similar fossil biological assemblages at a particular time - AHC, *K*-means, TWINSPAN
- Detect groups of geographically contiguous sites with similar fossil biological assemblages at a particular time – CC2

Abbreviations used in this table

AHC: agglomerative hierarchical clustering  
 TWINSPAN: two-way indicator species analysis  
*K*-means: *K*-means partitioning  
 CC1: constrained clustering in one dimension  
 MRT: multivariate regression tree

CC2: constrained clustering in two dimensions  
 ISA: indicator species analysis  
 DA: simple discriminant function analysis  
 CC: clustering (or partitioning) comparison (Rand's index, etc)  
 CCC: clustering constrained by canonical ordination results

Table 3. Stepwise condensation of the similarity matrix from Table 1 during UPGMA agglomerative clustering. At each step, the highest similarity value is identified (italic boldface value in the previous step); then the values corresponding to the similarities of these 2 objects or groups, with all other objects or groups, are averaged in turn.

Objects	1	2	3	4	5
1	—				
2	<b>0.8</b>	—			
3	0.0	0.0	—		
4	0.0	0.2	0.4	—	
5	0.0	0.2	0.2	0.6	—
Cluster objects 1 and 2 at $S = 0.8$					
1-2		—			
3		0.0	—		
4		$(0+0.2)/2=0.1$	0.4	—	
5		$(0+0.2)/2=0.1$	0.2	<b>0.6</b>	—
Cluster objects 4 and 5 at $S = 0.6$					
1-2		—			
3		0.0	—		
4-5		$(0.1+0.1)/2=0.1$	$(0.4+0.2)/2=0.3$	—	
Cluster object 3 with group (4,5) at $S = 0.3$					
1-2		—			
3-4-5		$(1 \times 0.0) + (2 \times 0.1) / 3 = 0.067$	—		
Cluster group (1,2) with group (3,4,5) at $S = 0.067$					

Table 4. Matrices of normalised distances  $[D/D_{max}]$  and squared normalised distances  $[(D/D_{max})^2]$  used in Ward's agglomerative clustering.

$D/D_{max}$	1	2	3	4	5
1	0				
2	0.2	0			
3	1.0	1.0	0		
4	1.0	0.8	0.6	0	
5	1.0	0.8	0.8	0.4	0

$(D/D_{max})^2$	1	2	3	4	5
1	0				
2	0.04	0			
3	1.00	1.00	0		
4	1.00	0.64	0.36	0	
5	1.00	0.64	0.64	0.16	0

Clustering step $k$	New cluster formed	Fusion $D_i$	Fusion $D_i$	ESS <sub><math>i</math></sub>	TESS <sub><math>i</math></sub>	$R_i$
0	5 separate objects	0	0	0	0	1
1	(1,2)	0.04	0.2	0.02	0.02	0.985
2	(4,5)	0.16	0.4	0.08	0.10	0.923
3	(3,4,5)	0.613	0.783	0.387	0.407	0.686
4	(1,2,3,4,5)	1.779	1.334	1.296	1.296	0

Table 5. Cophenetic matrices for the UPGMA (left) and Ward's dendrograms (right) representing the dendrograms shown in Figures 2 and 4, respectively.

UPGMA <b>S</b>	1	2	3	4	5
1	<i>1</i>				
2	0.800	<i>1</i>			
3	0.067	0.067	<i>1</i>		
4	0.067	0.067	0.300	<i>1</i>	
5	0.067	0.067	0.300	0.600	<i>1</i>

Ward's <b>D</b>	1	2	3	4	5
1	<i>0</i>				
2	0.200	<i>0</i>			
3	1.334	1.334	<i>0</i>		
4	1.334	1.334	0.783	<i>0</i>	
5	1.334	1.334	0.783	0.400	<i>0</i>

Table 6. Results of indicator species analysis for the partition of the RLGH core data into 12 groups by chronological clustering. The diatom taxa *j* with significant *IndVal* values (significance level: 0.05) are listed for each group of the partition. Indicator values are given in parentheses. Group 1 is the most recent, group 12 the oldest of the 10000-year fossil data series. Diatom nomenclature follows Stevenson et al. (1991).

Group No.	Membership (level no.)	Significant indicator diatoms (taxon names)
1	1-5	<i>Eunotia bactriana</i> (69), <i>Navicula pupula</i> (52), <i>Tabellaria quadrisepata</i> (46), <i>Suirella delicatissima</i> (44), <i>Cymbella aequalis</i> (39), <i>Navicula hoefleri</i> (37), <i>Tabellaria binalis</i> (34), <i>Eunotia exigua</i> (25), <i>Eunotia pectinalis</i> var. <i>minor</i> (16)
2	6-14	<i>Navicula cumbriensis</i> (42), <i>Eunotia tenella</i> (19)
3	15-17	<i>Neidium bisulcatum</i> (38), <i>Eunotia naegelii</i> (36), <i>Eunotia denticulata</i> (33), <i>Pinnularia microstauron</i> (26), <i>Peronia fibula</i> (26), <i>Suirella biseriata</i> (25), <i>Navicula leptostriata</i> (25), <i>Eunotia incisa</i> (16)
4	18-36	<i>Eunotia pectinalis</i> var. <i>minor</i> form <i>impressa</i> (26), <i>Achnanthes pseudoswazi</i> (25), <i>Tabellaria flocculosa</i> (19), <i>Achnanthes austriaca</i> (17), <i>Navicula mediocris</i> (16)
5	37-44	<i>Achnanthes umara</i> (56), <i>Navicula minima</i> (48), <i>Cymbella microcephala</i> (47), <i>Cyclotella kuetzingiana</i> agg. (38), <i>Navicula pupula</i> (38), <i>Nitzschia perminuta</i> (32), <i>Navicula arvensis</i> (30), <i>Nitzschia fonticola</i> (28), <i>Achnanthes flexella</i> (27), <i>Navicula minuscula</i> var. <i>muralis</i> (27), <i>Achnanthes minutissima</i> (25), <i>Navicula radiosa</i> var. <i>tenella</i> (22)
6	45-53	<i>Aulacoseira lirata</i> var. <i>lacustris</i> (48), <i>Eunotia diodon</i> (24)
7	54-62	<i>Navicula indifferens</i> (32)
8	63-65	<i>Eunotia vanheurckii</i> (32), <i>Frustulia rhomboides</i> var. <i>saxonica</i> (14), <i>Fragilaria elliptica</i> (43)
9	66-78	<i>Semiorbis hemicyclus</i> (38), <i>Brachysira</i> [sp.1] (28), <i>Aulacoseira perglabra</i> var. <i>floriniae</i> (24), <i>Fragilaria vaucheriae</i> (24)
10	79-90	<i>Navicula tenuicephala</i> (56), <i>Cymbella</i> [PIRLA sp.1] (37), <i>Eunotia iatriaensis</i> (31), <i>Navicula bremensis</i> (24)
11	91-94	<i>Navicula seminuloides</i> (57), <i>Aulacoseira distans</i> var. <i>tenella</i> (47), <i>Navicula seminulum</i> (42), <i>Navicula impexa</i> (35), <i>Aulacoseira</i> [cf. <i>distans distans</i> ] (35), <i>Aulacoseira perglabra</i> (23), <i>Aulacoseira lirata</i> (26)
12	95-101	<i>Navicula hassiaca</i> (56), <i>Cymbella perpusilla</i> (50), <i>Stauroneis anceps</i> form <i>gracilis</i> (38), <i>Pinnularia subcapitata</i> var. <i>hilseana</i> (28), <i>Navicula angusta</i> (26), <i>Gomphonema acuminatum</i> var. <i>coronatum</i> (25), <i>Brachysira vitrea</i> (20), <i>Fragilaria virescens</i> var. <i>exigua</i> (19), <i>Achnanthes marginulata</i> (15)

Figure 1. Principal coordinate analysis (PCoA) ordination computed from the distance matrix of the artificial data (Table 1) showing the relationships among the 5 objects.

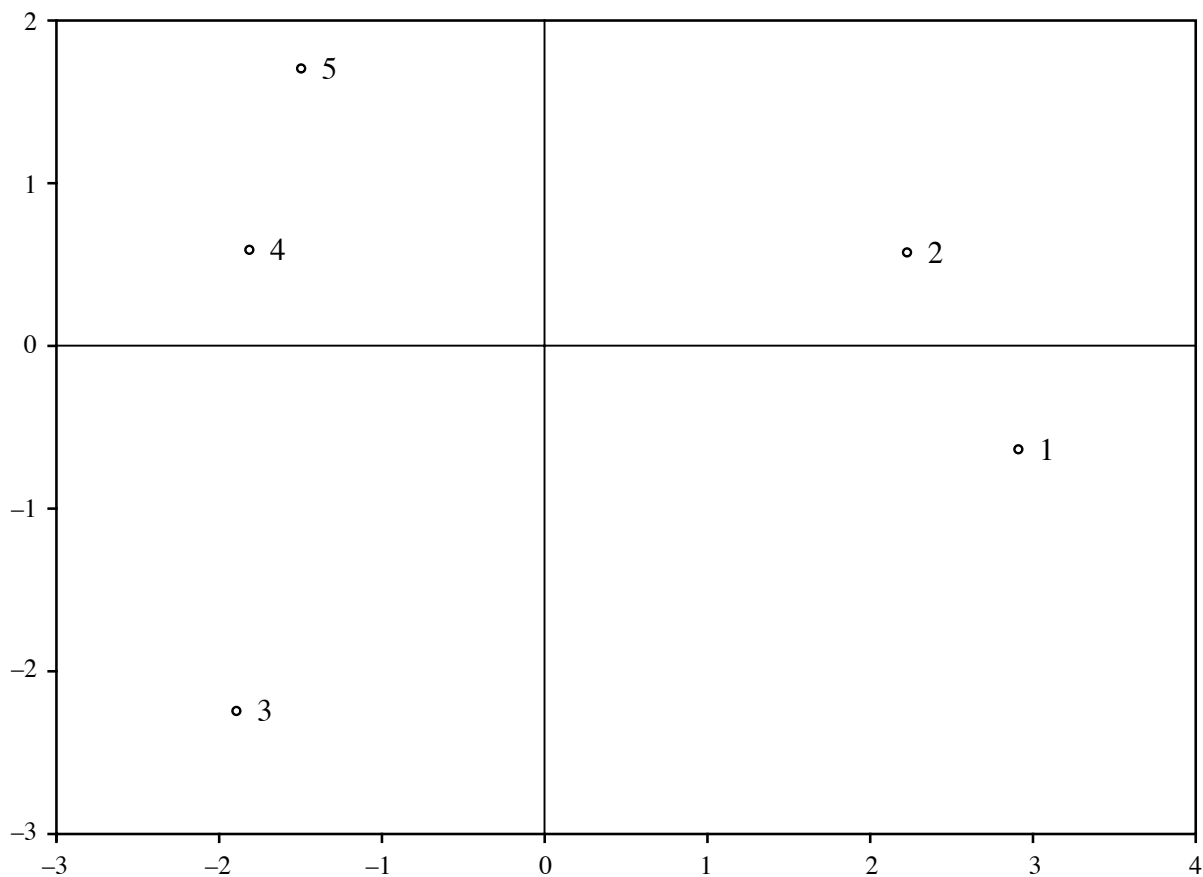


Figure 2. UPGMA agglomerative clustering of the objects of the artificial example.

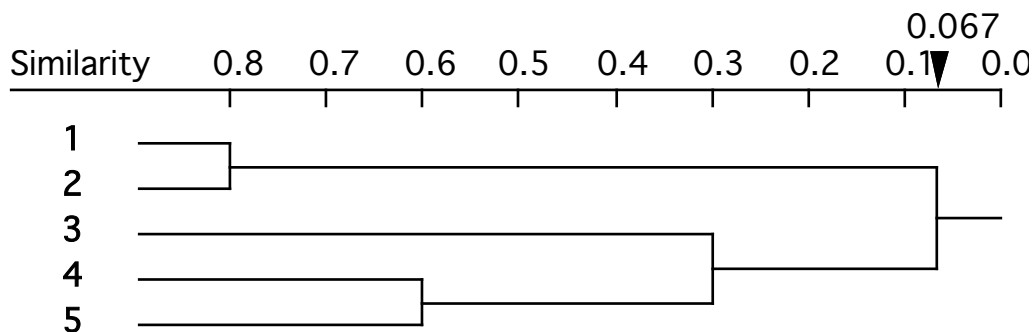


Figure 3. (a) In PCoA ordination space, each UPGMC clustering step leads to the replacement of objects (circles) by their centroid (squares), computed as the centre of mass of all the objects members of a group. (b) The UPGMC dendrogram.

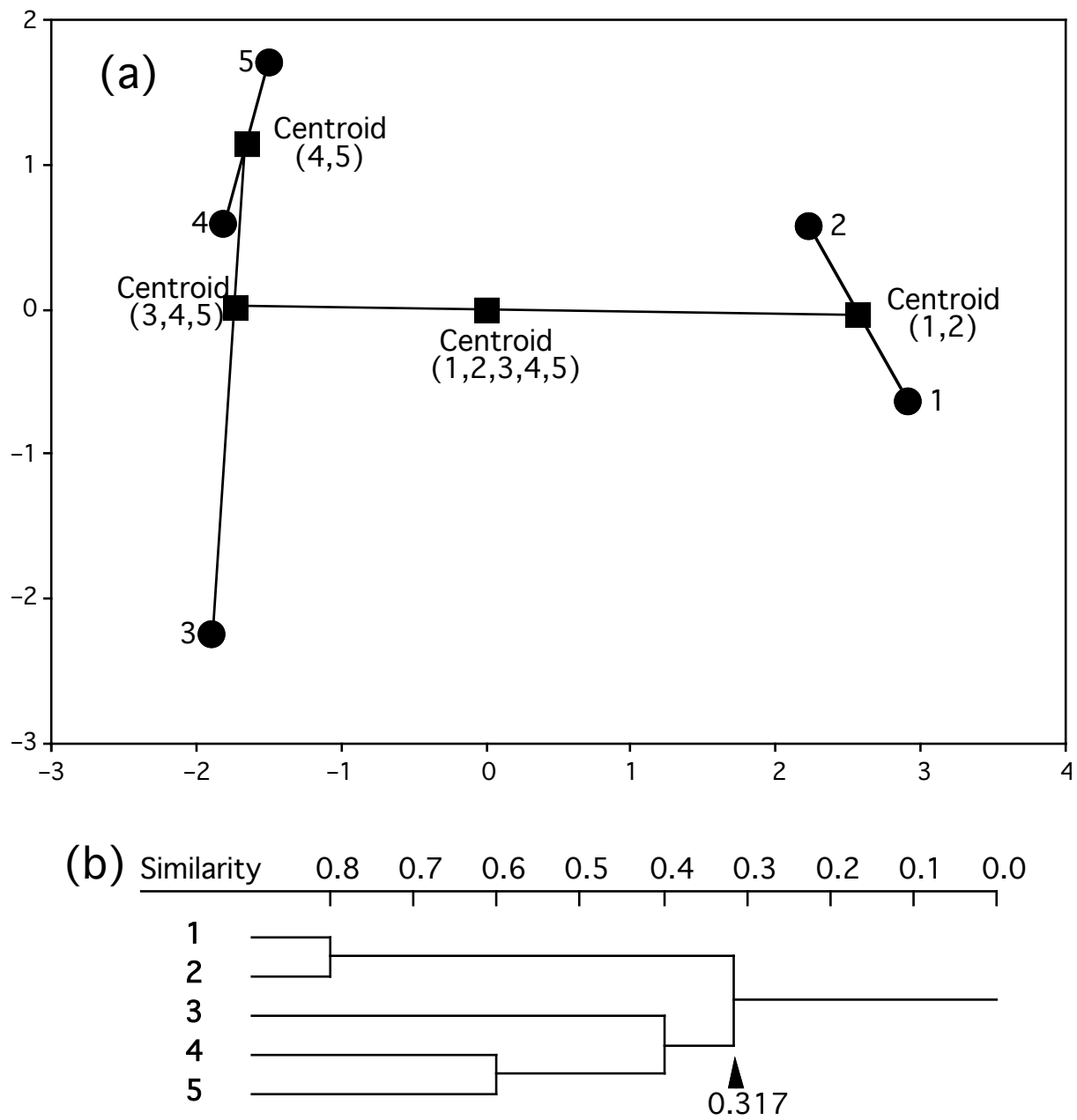






Figure 5. Linkage agglomerative clustering of the objects of the artificial example. (a) Single linkage ( $C_0 = 0$ ). (b) Complete linkage ( $C_0 = 1$ ). (c) Proportional-link linkage ( $C_0 = 0.5$ ).

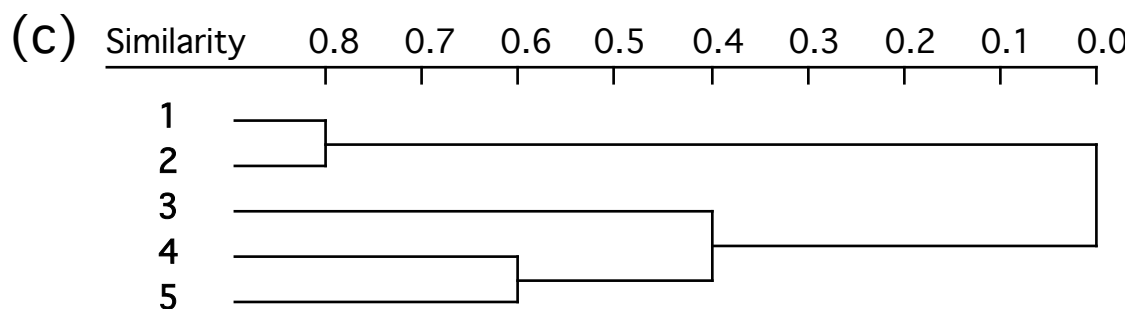
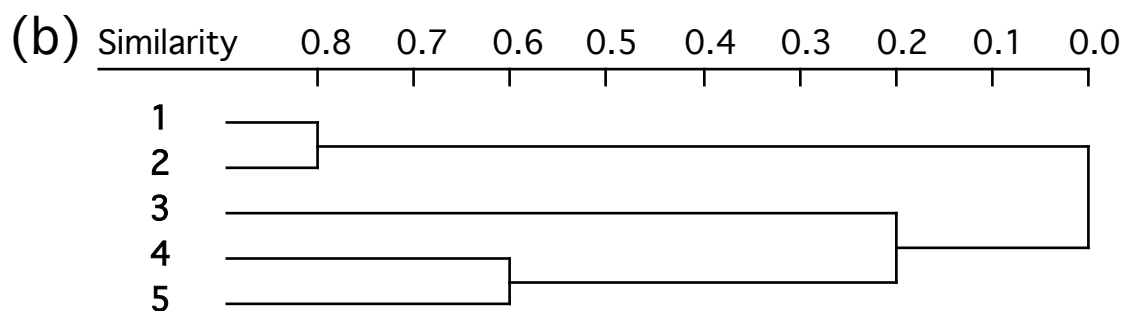
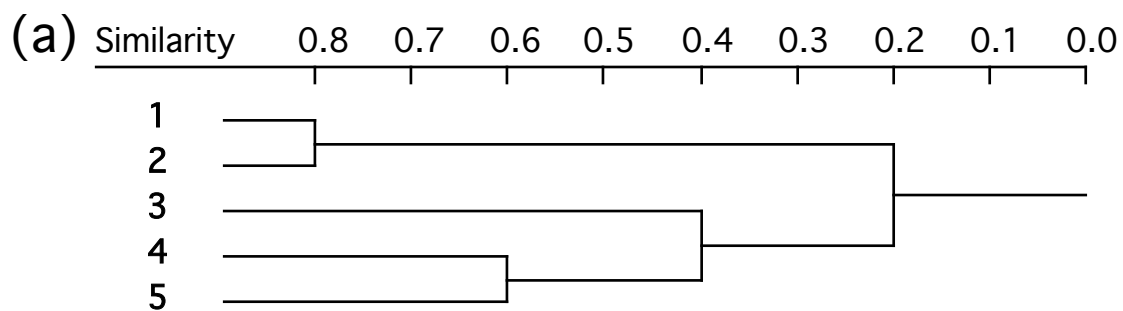


Figure 6. Contingency table comparing two partitions of a set of objects.  $a$ ,  $b$ ,  $c$ , and  $d$  are frequencies.

		Second partition	
		Same group	Different groups
First partition	Same group	a	b
	Different groups	c	d

Figure 7. K-means partitioning of the SWAP-UK data into  $K = 5$  groups. The diatom assemblages are represented in a scatter diagram of pH and latitude. The ellipses summarise the extent of the 5 groups in the graph.

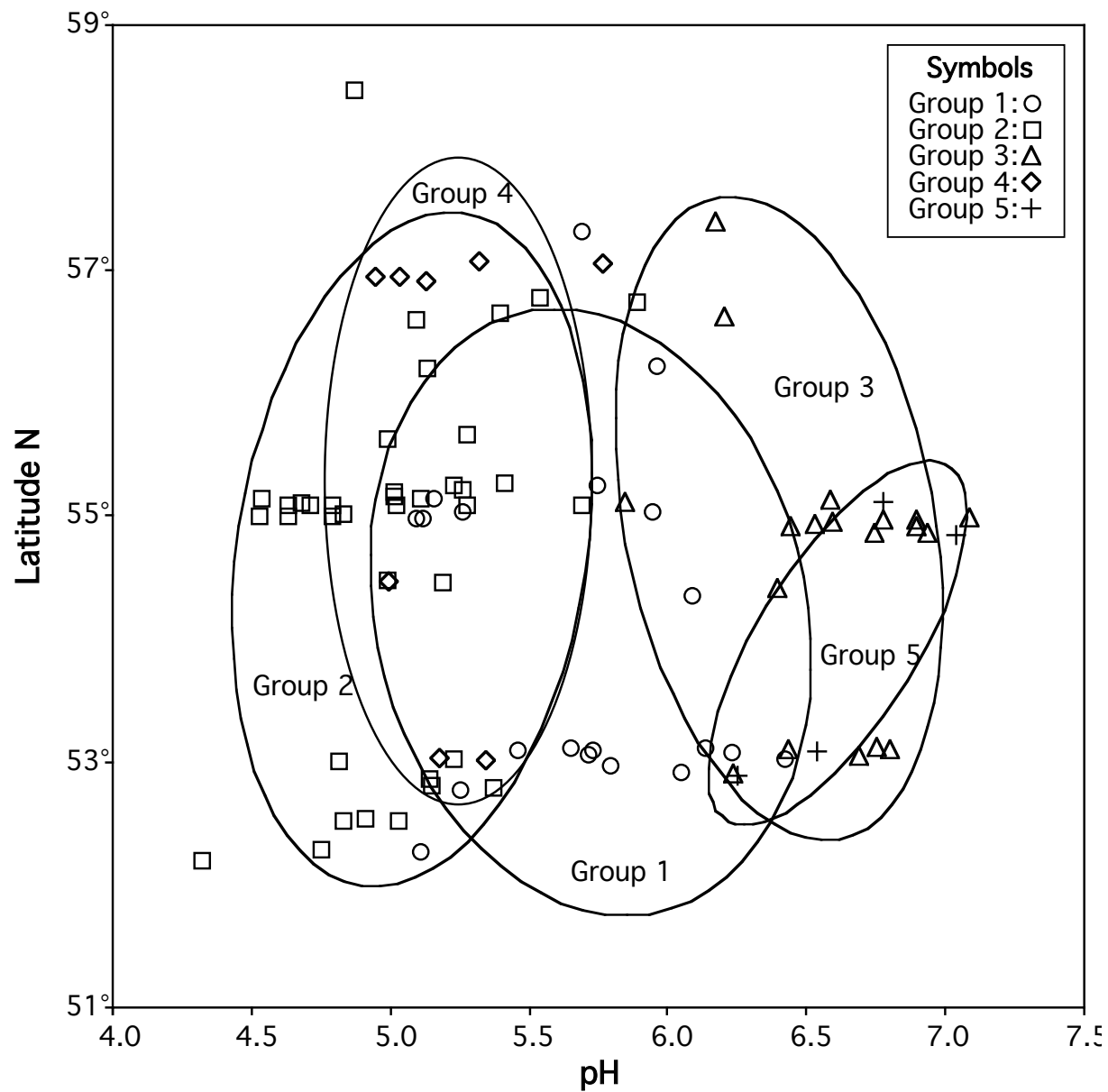


Figure 8. The constrained similarity matrix used in constrained clustering is the Hadamard product (cell-by-cell product) of the similarity with the matrix of spatial contiguity. Only the upper triangular portion of each matrix is shown. Nine objects are used in this illustration.

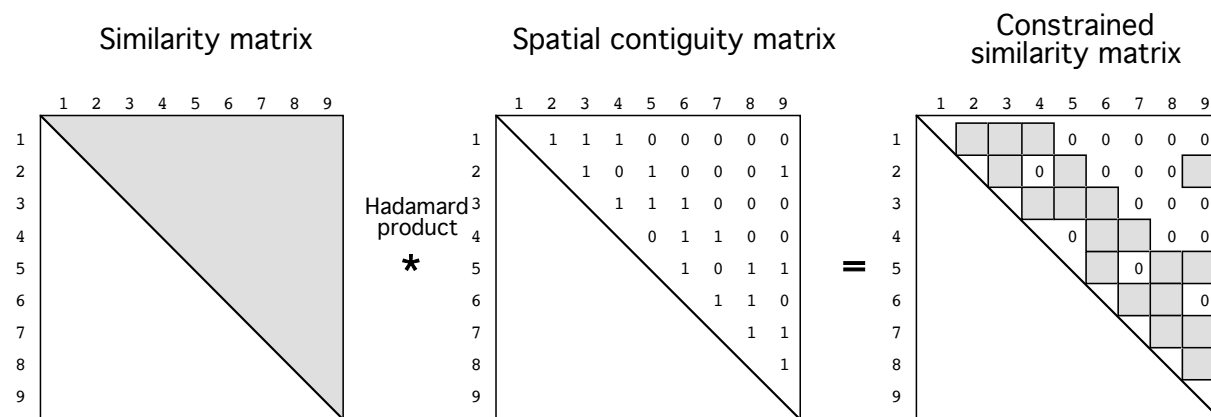


Figure 9. The SWAP-UK data set. (a) Delaunay triangulation. The edges of the graph served as constraints in constrained clustering. (b) Results of spatially-constrained clustering for 63 lakes among 90: 10-group partition of the lakes (symbols). All members of a group are connected through adjacency links in the triangulation.

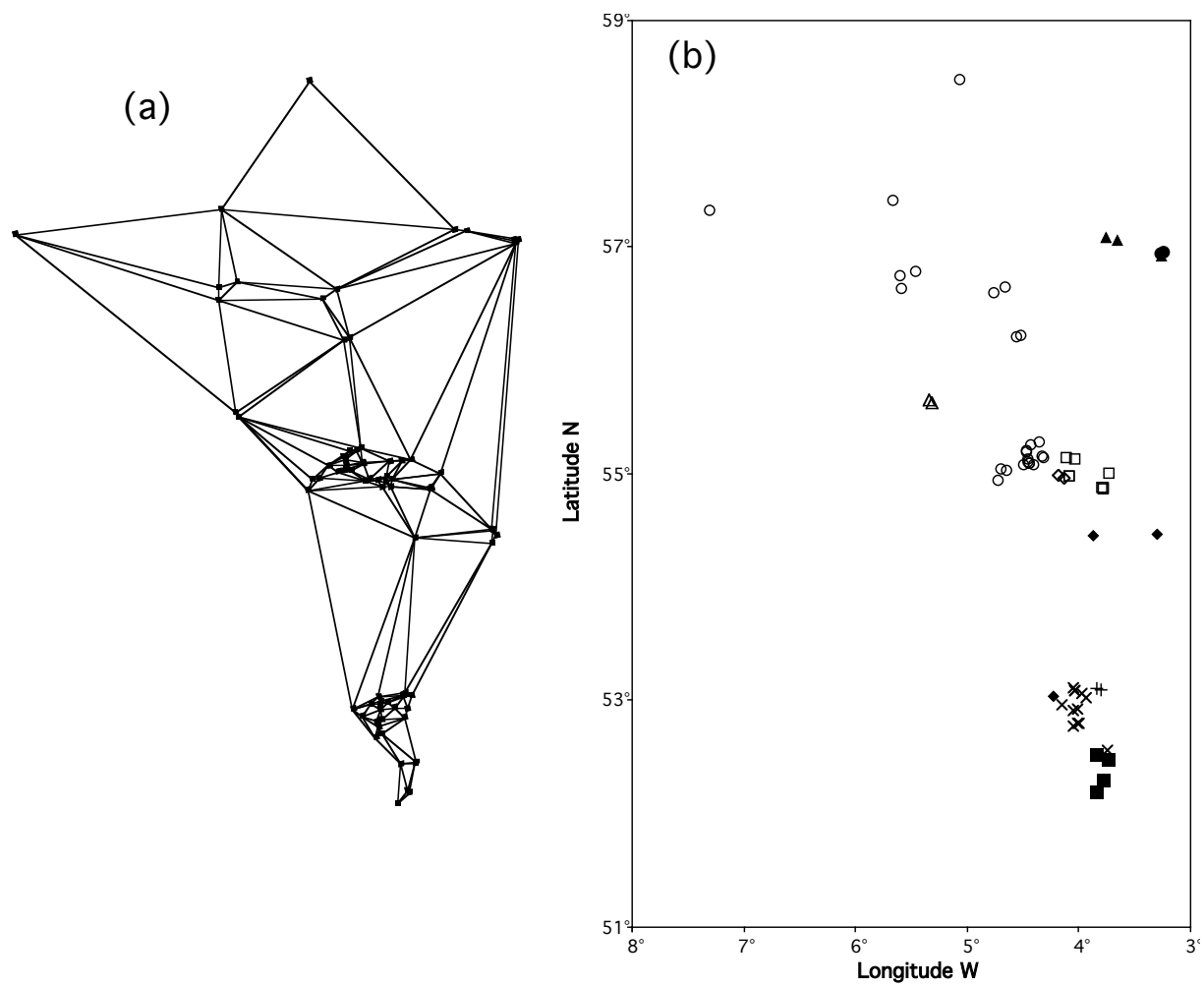


Figure 10. Two-way indicator species analysis (TWINSPAN) of the SWAP-UK data (90 lakes, 234 diatom species). The numbers of lakes in the total data-set, the second-level (2 groups), and the third-level (4 groups) divisions are shown in squares. The indicator species for each division are shown, with their pseudospecies values in parentheses (1 = <1%, 2 = 1-2%, 3 = 2-4%, 4 = 4-8%, 5 = 8-16%). The median pH and range of pH values for the lakes in the four groups are also shown.

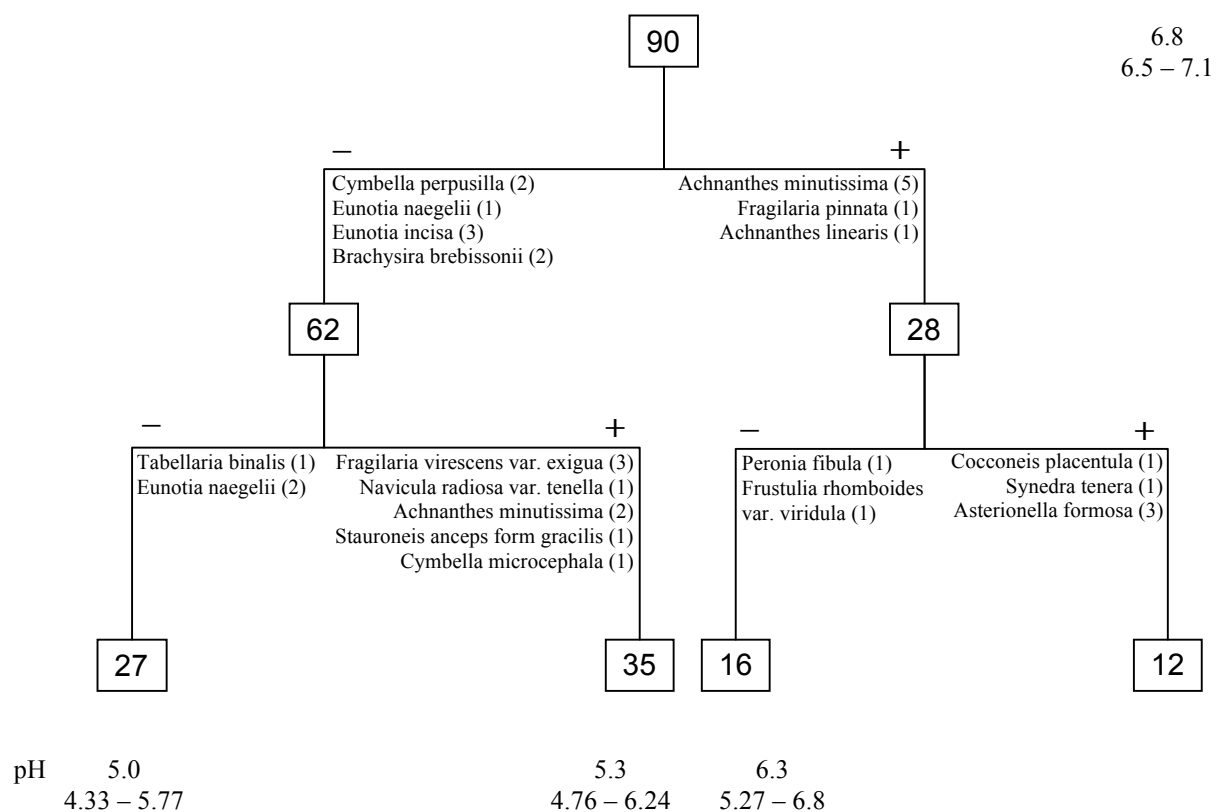


Figure 11. Multivariate regression tree (MRT) for the SWAP-UK data-set (90 lakes, 234 diatom taxa) using three explanatory variables: pH, longitude (Lon W) and latitude (Lat N) of the lakes. The best partition is two groups.  $n_k$ : number of lakes in group  $k$ .

