Chapter

8

Cluster analysis

8.0 A search for discontinuities

Humans have always tried to classify the animate and inanimate objects that surround them. Classifying objects into collective categories is a prerequisite to naming them. It requires the recognition of discontinuous subsets in an environment which is sometimes discrete, but most often continuous.

To cluster is to recognize that objects are sufficiently similar to be put in the same group and to also identify distinctions or separations between groups. Measures of similarity between objects (Q mode) or descriptors (R mode) have been discussed in Chapter 7. The present Chapter considers the different criteria that may be used to decide whether objects are similar enough to be allocated to a group; it also shows that different clustering strategies correspond to different definitions of a what a cluster is.

Few ecological theories predict the existence of discontinuities in nature. Evolutionary theory tells taxonomists that discontinuities exist between species, which are the basic units of evolution, as a result of reproductive barriers; taxonomists use classification methods to reveal these discontinuities. For the opposite reason, taxonomists are not surprised to find continuous differentiation at the sub-species level. In contrast, the world that ecologists try to understand is most often a continuum. In numerical ecology, methods used to identify clusters must therefore be more contrasting than in numerical taxonomy. Ecologists who have applied taxonomic clustering methods directly to data, without first considering the theoretical applicability of such methods, have often obtained disappointing results. This has led many ecologists to abandon clustering methods altogether, hence neglecting the rich potential of similarity measures, described in Chapter 7, and to rely instead on factor analysis and other ordination methods. These are not always adapted to ecological data and, in any case, they don’t aim at bringing out partitions, but gradients.

Given a sufficiently large group of objects, ecological clustering methods should be able to recognize clusters of similar objects while ignoring the few intermediates which often persist between clusters. Indeed, one cannot expect to find discontinuities when clustering sampling sites unless the physical environment is itself discontinuous, or unless sampling occurred at opposite ends of a gradient, instead of within the gradient (Whittaker, 1962: 88). Similarly, when looking for associations of species, small groups of densely associated species are usually found, with the other species gravitating around one or more of the association nuclei.

Typology

The result of clustering ecological objects sampled from a continuum is often called a *typology* (i.e. a system of types). In such a case, the purpose of clustering is to identify various *object types* which may be used to describe the structure of the continuum; it is thus immaterial to wonder whether these clusters are “natural” or unique.

For readers with no practical experience in clustering, Section 8.2 provides a detailed account of single linkage clustering, which is simple to understand and is used to introduce the principles of clustering. The review of other methods includes a survey of the main dichotomies among existing methods (Section 8.4), followed by a discussion of the most widely available methods of interest to ecologists (8.5, 8.7 and 8.8). Theoretical aspects are discussed in Sections 8.3 and 8.6. Section 8.9 deals with clustering algorithms useful in identifying biological associations, whereas Section 8.10 gives an overview of seriation, a method useful to cluster non-symmetric resemblance matrices. A review of clustering statistics, methods of cluster validation, and graphical representations, completes the chapter (Sections 8.11 to 8.13). The relationships between clustering and other steps of data analysis are depicted in Fig. 10.3.

Several, but not all statistical packages offer clustering capabilities. All packages with clustering procedures offer at least a Lance & Williams algorithm capable of carrying out the clustering methods listed in Table 8.8. Many also have a *K*-means partitioning algorithm. Few offer proportional-link linkage or additional forms of clustering. Some methods are available in specialized packages only: clustering with constraints of temporal (Section 12.6) or spatial contiguity (Section 13.3); fuzzy clustering (e.g. Bezdek, 1987); or clustering by neural network algorithms (e.g. Fausett, 1994). The main difference among packages lies in the list of resemblance coefficients available (Section 7.7). Ecologists should consider this point when selecting a clustering package.

While most packages nowadays illustrate clustering results in the form of dendrograms, some programs use “skyline plots”, which are also called “trees” or “icicle plots”. These plots contain the same information as dendrograms but are rather odd to read and interpret. The way to transform a skyline plot into a dendrogram is explained in Section 8.13.

Despite the versatility of clustering methods, one should remember that not all problems are clustering problems. Before engaging in clustering, one should be able to justify why one believes that discontinuities exist in the data; or else, explain that one has a practical need to divide a continuous swarm of objects into groups.

8.1 Definitions

Clustering
Partition

Clustering is an operation of multidimensional analysis which consists in partitioning the collection of objects (or descriptors) in the study. A *partition* (Table 8.1) is a division of a set (collection) into subsets, such that each object or descriptor belongs to one and only one subset for that partition (Legendre & Rogers, 1972). The classification of objects (or descriptors) that results from clustering may include a single partition, or several hierarchically nested partitions of the objects (or descriptors), depending on the clustering model that has been selected.

From this definition, it follows that the subsets of any level of partition form a series of *mutually exclusive* cells, among which the objects (or descriptors) are distributed. This definition *a priori* excludes all classification models in which classes have elements in common (overlapping clusters) or in which objects have fractional degrees of membership in different clusters (fuzzy partitions: Bezdek, 1987); these models have not been used in ecology yet. This limitation is such that a “hard” or “crisp” (*versus* fuzzy) partition has the same definition as a descriptor (Section 1.4). Each object is characterized by a state (its cluster) of the classification and it belongs to only one of the clusters. This property will be useful for the interpretation of classifications (Chapter 10), since any partition may be considered as a qualitative descriptor and compared as such to any other descriptor. A clustering of objects defined in this way imposes a discontinuous structure onto the data set, even if the objects have originally been sampled from a continuum. This structure results from the grouping into subsets of objects that are sufficiently similar, given the variables considered, and from the observation that different subsets possess unique recognizable characteristics.

Table 8.1 Example of hierarchically nested partitions of a group of objects (e.g. sampling sites). The first partition separates the objects by the environment to which they belong. The second partition, hierarchically nested into the first, recognizes clusters of sites in each of the two environments.

Partition 1	Partition 2	Sampling sites
Observations in environment A	Cluster 1	7, 12
	Cluster 2	3, 5, 11
	Cluster 3	1, 2, 6
Observations in environment B	Cluster 4	4, 9
	Cluster 5	8, 10, 13, 14

Clustering has been part of ecological tradition for a long time. It goes back to the Polish ecologist Kulczynski (1928), who needed to cluster ecological observations; he developed a method quite remote from the clustering algorithms discussed in the paragraphs to come. His technique, called seriation, consists in permuting the rows and columns of an association matrix in such a way as to maximize the values on the diagonal. The method is still used in phytosociology, anthropology, social sciences, and other fields; it is described in Section 8.10 where an analytical solution to the problem is presented.

Most clustering (this Chapter) and ordination (Chapter 9) methods proceed from association matrices (Chapter 7). Distinguishing between clustering and ordination is somewhat recent. While ordination in reduced space goes back to Spearman (factor analysis: 1904), most modern clustering methods have only been developed since the era of second-generation computers. The first programmed method, developed for biological purposes, goes back to 1958 (Sokal & Michener)*. Before that, one simply plotted the objects in a scatter diagram with respect to a few variables or principal axes; clusters were then delineated manually (Fig. 8.1) following a method which today would be called centroid (Section 8.4), based upon the Euclidean distances among points. This empirical clustering method still remains the best approach when the number of variables is small and the structure to be delineated is not obscured by intermediate objects between clusters.

Clustering is a family of techniques which is undergoing rapid development. In their report on the literature they reviewed, Blashfield & Aldenderfer (1978) mentioned that they found 25 papers in 1964 that contained references to the basic texts on clustering; they found 136 papers in 1970, 294 in 1973, and 501 in 1976. The number has been growing ever since. Nowadays, hundreds of mathematicians and researchers from various application fields are collaborating within 10 national or multinational *Classification Societies* throughout the world, under the umbrella of the *International Federation of Classification Societies* founded in 1985.

The commonly-used clustering methods are based on easy-to-understand mathematical constructs: arithmetic, geometric, graph-theoretic, or simple statistical models (minimizing within-group variance), leading to rather simple calculations on the similarity or dissimilarity values. It must be understood that most clustering methods are heuristic; they create groups by reference to some concept of what a group embedded in some space should be like, but without reference, in most case, to the processes occurring in the application field — ecology in the present book. They have been developed first by the schools of numerical taxonomists and numerical ecologists, later joined by other researchers in the physical sciences and humanities.

* Historical note provided by Prof. F. James Rohlf: “Actually, Sokal & Michener (1958) did not use a computer for their very large study. They used an electromechanical accounting machine to compute the raw sums and sums of products. The coefficients of correlation and the cluster analysis itself were computed by hand with the use of mechanical desk calculators. Sneath did use a computer in his first study.”

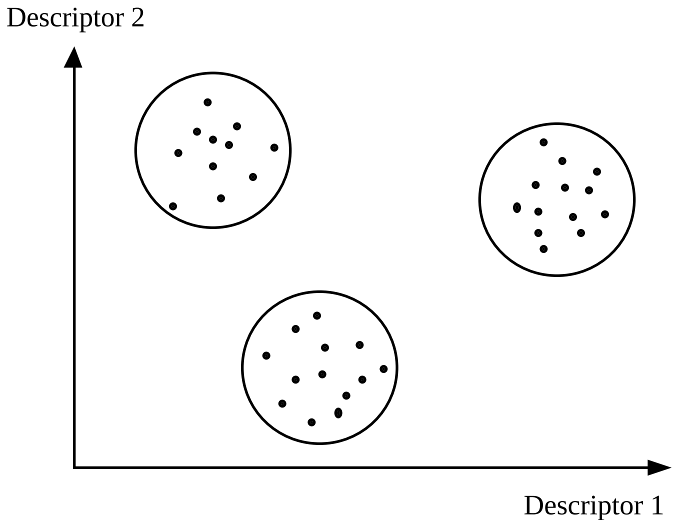


Figure 8.1 Empirically delineating clusters of objects in a scatter diagram is easy when there are no intermediate objects between the groups.

Clusters are delineated on the basis of statements such as: "$\mathbf{x}_1$ is closer to $\mathbf{x}_2$ than it is to $\mathbf{x}_3$", whereas other methods rest on probabilistic models of the type: "Chances are higher that $\mathbf{x}_1$ and $\mathbf{x}_2$ pertain to the same group than $\mathbf{x}_1$ and $\mathbf{x}_3$". In all cases, clustering models make it possible to link the points without requiring prior positioning in a graph (i.e. a metric space), which would be impractical in more than three dimensions. These models thus allow a graphical representation of other interesting relationships among the objects in the data set, for example the dendrogram of their interrelationships. Chapter 10 will show how it is possible to combine clustering and ordination, computed with different methods, to obtain a more complete picture of the data structure.

Descriptive, synoptic clustering

The choice of a clustering method is as critical as the choice of an association measure. It is important to fully understand the properties of clustering methods in order to correctly interpret the ecological structure they bring out. Most of all, the methods to be used depend upon the type of clustering sought. Williams *et al.* (1971) recognized two major categories of methods. In a *descriptive clustering*, misclassifying objects is to be avoided, even at the expense of creating single object clusters. In a *synoptic clustering*, all objects are forced into one of the main clusters; the objective is to construct a general conceptual model which encompasses a reality wider than the data under study. Both approaches are useful.

When two or more clustering models seem appropriate to a problem, ecologists should apply them all to the data and compare the results. Clusters that repeatedly come out of all runs (based on appropriate methods) are the robust solutions to the clustering problem. Differences among results must be interpreted in the light of the known properties of clustering models, which are explained in the following sections.

8.2 The basic model: single linkage clustering

For natural scientists, a simple-to-understand clustering method (or *model*) is *single linkage* (or *nearest neighbour*) clustering (Sneath, 1957). Its logic seems natural, so that it is used to introduce readers to the principles of clustering. Its name, *single linkage*, distinguishes it from other clustering models, called complete or intermediate linkage, detailed in Section 8.4. The algorithm for single linkage clustering is sequential, agglomerative, and hierarchical, following the nomenclature of Section 8.3. Its starting point is any association matrix (similarity or distance) among the objects or descriptors to be clustered. One assumes that the association measure has been carefully chosen, following the recommendations of Section 7.6. To simplify the discussion, the following will deal only with objects, although the method is equally applicable to an association matrix among descriptors.

The method proceeds in two steps:

- First, the association matrix is rewritten in order of decreasing similarities (or increasing distances), heading the list with the two most similar objects of the association matrix, followed by the second most similar pair, and proceeding until all the measures comprised in the association matrix have been listed.

- Second, the clusters are formed hierarchically, starting with the two most similar objects, and then letting the objects clump into groups, and the groups aggregate to one another, as the similarity criterion is relaxed. The following example illustrates this method.

Ecological application 8.2

Five ponds characterized by 38 zooplankton species have been studied by Legendre & Chodorowski (1977). The data were counts, recorded on a relative abundance scale from 0 = absent to 5 = very abundant. These ponds have already been used as example for the computation of Goodall's coefficient (S_{23}, Chapter 7; only eight zooplankton species were used in that example). These five ponds, with others, have been subjected to single linkage clustering after computing similarity coefficient S_{20} with parameter $k = 2$. The symmetric similarity matrix is represented by its lower triangle. The diagonal is trivial because it contains 1's by construct.

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.600	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—

The first clustering step consists in rewriting these similarities in decreasing order:

S_{20}	Pairs formed
0.600	212-214
0.500	431-432
0.300	233-431
0.214	214-432
0.200	233-432
0.071	214-233
0.063	214-431
0.000	212-233
0.000	212-431
0.000	212-432

Link

As the similarity levels drop, pairs of objects are formed. These pairs are called "links"; they serve to link the objects or groups into a chain, as discussed below.

Connected subgraphs are one of the many possible graphical representations of cluster formation (Fig. 8.2a). As the similarity decreases, clusters are formed, following the list of links in the table of ordered similarities above. Only the similarity levels at which clusters are modified by addition of objects are represented here. The first link is formed between ponds 212 and 214 at $S = 0.6$, then between 431 and 432 at $S = 0.5$. Pond 233 joins this second cluster nucleus at $S = 0.3$. Finally these two clusters merge at $S = 0.214$ by a link which is formed between ponds 214 and 432. The clustering may stop at this point since all ponds now belong to the same cluster. If the similarity criterion was relaxed down to $S = 0$, links would form between members of the cluster up to a point where all ponds would be linked to one another. This part of the clustering is of no interest in single linkage clustering, but these links will be of interest in the other forms of linkage clustering below.

Dendrogram

Edge

Node

A dendrogram (Fig. 8.2b) is another, more commonly-used representation of hierarchical clustering results. Dendrograms only display the clustering topology and object labels, not the links between objects. Dendrograms are made of branches ("edges") that meet at "nodes" which are drawn at the similarity value where fusion of branches takes place. For graphical convenience, vertical lines are used in Fig. 8.2b to connect branches at the similarity levels of the nodes; the lengths of these lines are of no consequence. Branches could be directly connected to nodes. The branches furcating from a node may be switched ("swivelled") without affecting the information contained in a dendrogram.

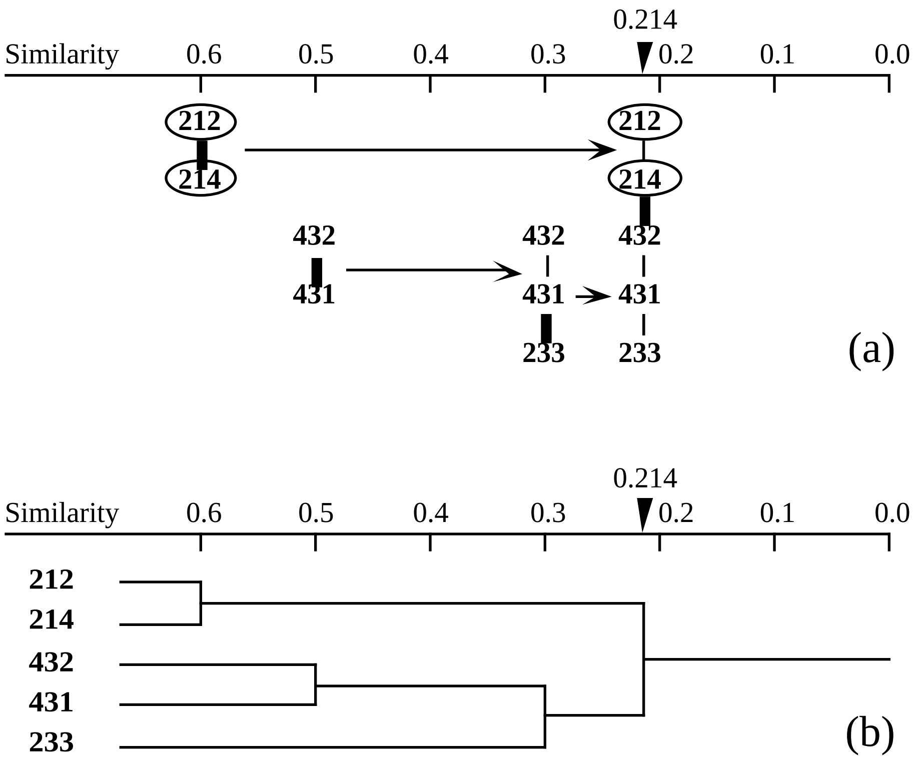


Figure 8.2 Illustrations of single linkage agglomerative clustering for the ponds of the example. (a) Connected subgraphs: groups of objects are formed as the similarity level is relaxed from left to right. Only the similarity levels where clusters are modified by addition of objects are represented. New links between ponds are represented by heavy lines; thin lines are used for links formed at previous (higher) similarity levels. Circled ponds are non-permanent; the others are permanent. (b) Dendrogram for the same cluster analysis.

The clustering results have been interpreted by Legendre & Chodorowski (1977) with respect to the conditions prevailing in the ponds. In their larger study, all non-permanent ponds (including 212 and 214) formed a cluster, while the permanent ponds (including 233, 431 and 432) formed a distinct group, on the basis of zooplankton assemblages.

Single linkage rule

From this example, it should be clear that the rule for assigning an object to a cluster, in single linkage clustering, requires an object to display a similarity at least equal to the considered level of partition with *at least one object already member of the cluster.* In complete linkage hierarchical clustering, the assignment rule differs and requires the object to display the given level of similarity with *all* the objects already members of the cluster. The chaining rule used in single linkage clustering may be stated as follows: at each level of partition, two objects must be allocated to the same subset if their degree of similarity is equal to or higher than that of the partitioning level considered. The same rule can be formulated in terms of dissimilarities (distances) instead: two objects must be allocated to the same subset if their dissimilarity is less than or equal to that of the partitioning level considered.

Estabrook (1966) discussed single linkage clustering using the language of graph theory. The exercise has didactic value. A cluster is defined through the following steps:

Link

a) For any pair of objects $\mathbf{x}_1$ and $\mathbf{x}_2$, a *link* is defined between them by a relation G_c:

$$\mathbf{x}_1\ G_c\ \mathbf{x}_2 \text{ if and only if } S(\mathbf{x}_1, \mathbf{x}_2) \geq c$$
$$\text{or equally, if } D(\mathbf{x}_1, \mathbf{x}_2) \leq (1 - c)$$

Index c in clustering relation G_c is the similarity level considered. At a similarity level of 0.55, for instance, ponds 212 and 214 of the example are in relation $G_{0.55}$ since $S(212, 214) \geq 0.55$. This definition of a link has the properties of symmetry ($\mathbf{x}_1\ G_c\ \mathbf{x}_2$ if and only if $\mathbf{x}_2\ G_c\ \mathbf{x}_1$) and reflexivity ($\mathbf{x}_\mathrm{i}\ G_c\ \mathbf{x}_\mathrm{i}$ is always true since $S(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{i}) = 1.0$). A group of links for a set of objects, such as defined by relation G_c, is called an *undirected graph.*

Undirected graph

Chain
Chaining

b) The chaining which characterizes single linkage clustering may be described by a G_c-chain. A G_c-chain is said to extend from $\mathbf{x}_1$ to $\mathbf{x}_2$ if there exist other points $\mathbf{x}_3$, $\mathbf{x}_4$, …, $\mathbf{x}_\mathrm{i}$ in the collection of objects under study, such that $\mathbf{x}_1\ G_c\ \mathbf{x}_3$ and $\mathbf{x}_3\ G_c\ \mathbf{x}_4$ and … and $\mathbf{x}_\mathrm{i}\ G_c\ \mathbf{x}_2$. For instance, at similarity level $c = 0.214$ of the example, there exists a $G_{0.214}$-chain from pond 212 to pond 233, since there are intermediate ponds such that 212 $G_{0.214}$ 214 and 214 $G_{0.214}$ 432 and 432 $G_{0.214}$ 431 and 431 $G_{0.214}$ 233. The number of links in a G_c-chain defines the *connectedness* of a cluster (Subsection 8.11.1).

c) There only remains to delineate the clusters resulting from single linkage chaining. For that purpose, an equivalence relation R_c ("member of the same cluster") is defined as follows:

$$\mathbf{x}_1\ R_c\ \mathbf{x}_2 \text{ if and only if there exists a } G_c\text{-chain from } \mathbf{x}_1 \text{ to } \mathbf{x}_2 \text{ at similarity level } c.$$

In other words, $\mathbf{x}_1$ and $\mathbf{x}_2$ are assigned to the same cluster at similarity level c if there exists a chain of links joining $\mathbf{x}_1$ to $\mathbf{x}_2$. Thus, at level $S = 0.214$ in the example, ponds 212 and 233 are assigned to the same cluster (212 $R_{0.214}$ 233) because there exists a $G_{0.214}$-chain from 212 to 233. The relationship "member of the same cluster" has the following properties: (1) it is reflexive ($\mathbf{x}_\mathrm{i}\ R_c\ \mathbf{x}_\mathrm{i}$) because G_c is reflexive; (2) the G_c-chains may be reversed because G_c is symmetric; as a consequence, $\mathbf{x}_1\ R_c\ \mathbf{x}_2$ implies that $\mathbf{x}_2\ R_c\ \mathbf{x}_1$; and (3) it is transitive because, by G_c-chaining, $\mathbf{x}_1\ R_c\ \mathbf{x}_2$ and $\mathbf{x}_2\ R_c\ \mathbf{x}_3$ implies that $\mathbf{x}_1\ R_c\ \mathbf{x}_3$. Each cluster thus defined is a connected subgraph, which means that the objects of a cluster are all connected in their subgraph; in the graph of all the objects, distinct clusters (subgraphs) have no links attaching them to one another.

Connected subgraph

Single linkage clustering provides an accurate picture of the relationships between pairs of objects but, because of its propensity to chaining, it may be undesirable for ecological analysis. This means that the presence of an object midway between two compact clusters, or a few intermediates connecting the two clusters, is enough to turn them into a single cluster. Of course, clusters do not chain unless intermediates are present; so, the occurrence of chaining provides information about the data. To describe this phenomenon, Lance & Williams (1967c) wrote that this method contracts the reference space. Picture the objects as laying in A-space (Fig. 7.2). The presence of a cluster increases the probability of inclusion, by chaining, of neighbouring objects into the cluster. This is as if the distances between objects were smaller in that region of the space; see also Fig. 8.23a.

Section 10.1 will show how to take advantage of the interesting properties of single linkage clustering by combining it with ordination results, while avoiding the undue influence of chaining on the clustering structure.

Chain of primary connections

Minimum spanning tree

Ecologists who have access to several computer programs for single linkage clustering should rely on those that permit recognition of the first connection making an object a member of a cluster, or allowing two clusters to fuse. These similarities form a *chain of primary* (Legendre, 1976) or *external connections* (Legendre & Rogers, 1972), also called *dendrites* by Lukaszewicz (1951). They are very useful when analysing clusters drawn in an ordination space (Section 10.1). Dendrites are also called a *network* (Prim, 1957), a *Prim network* (Cavalli-Sforza & Edwards, 1967), a *minimum spanning tree* (Gower & Ross, 1969), a *shortest spanning tree*, or a *minimum-length tree* (Sneath & Sokal, 1973). If these dendrites are drawn on a scatter diagram of the objects, one can obtain a non-hierarchical clustering of the objects by removing the last (weakest) similarity links. Such graphs are illustrated in Figs. 10.1 and 10.2, drawn on top of an ordination in a space of principal coordinates; they may also be drawn without any reference space.

8.3 Cophenetic matrix and ultrametric property

Any classification or partition can be fully described by a cophenetic matrix. This matrix is used for comparing different classifications of the same objects.

1 — Cophenetic matrix

Cophenetic similarity

The *cophenetic similarity* (or *distance*) of two objects $\mathbf{x}_1$ and $\mathbf{x}_2$ is defined as the similarity (or distance) level at which objects $\mathbf{x}_1$ and $\mathbf{x}_2$ become members of the same cluster during the course of clustering (Jain & Dubes, 1988), as depicted by connected subgraphs or a dendrogram (Fig. 2a, b). Any dendrogram can be uniquely represented by a matrix in which the similarity (or distance) for a pair of objects is their cophenetic similarity (or distance). Consider the single linkage clustering dendrogram of Fig. 8.2. The clustering levels, read directly on the dendrogram, lead to the following matrices of similarity (**S**) and distance (**D**, where $D = 1 - S$):

S	212	214	233	431	432
212	—			(upper triangle symmetric to lower)	
214	0.600	—			
233	0.214	0.214	—		
431	0.214	0.214	0.300	—	
432	0.214	0.214	0.300	0.500	—

D	212	214	233	431	432
212	—			(upper triangle symmetric to lower)	
214	0.400	—			
233	0.786	0.786	—		
431	0.786	0.786	0.700	—	
432	0.786	0.786	0.700	0.500	—

Cophenetic matrix

Such a matrix is often called a *cophenetic matrix* (Sokal & Rohlf, 1962; Jain & Dubes, 1988). The ordering of objects in the cophenetic matrix is irrelevant; any order that suits the researcher is acceptable. The same applies to dendrograms; the order of the objects may be changed at will, provided that the dendrogram is redrawn to accommodate the new ordering.

For a *partition* of the data set (as in the *K*-means method, below), the resulting groups of objects are not related through a dendrogram. A cophenetic matrix may nevertheless be obtained. Consider the groups (212, 214) and (233, 431, 432) obtained by cutting the dendrogram of Fig. 8.2 at similarity level $S = 0.25$, ignoring the hierarchical structure of the two clusters. The cophenetic matrices would be:

S	212	214	233	431	432
212	—			(upper triangle symmetric to lower)	
214	1	—			
233	0	0	—		
431	0	0	1	—	
432	0	0	1	1	—

D	212	214	233	431	432
212	—			(upper triangle symmetric to lower)	
214	0	—			
233	1	1	—		
431	1	1	0	—	
432	1	1	0	0	—

2 — Ultrametric property

If there are no *reversals* in the clustering (Fig. 8.16), a classification has the following *ultrametric property* and the cophenetic matrix is called ultrametric:

$$D(\mathbf{x}_1, \mathbf{x}_2) \leq \max\,[D(\mathbf{x}_1, \mathbf{x}_3),\ D(\mathbf{x}_2, \mathbf{x}_3)] \qquad \textbf{(8.1)}$$

for every triplet of objects $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ in the study. Cophenetic distances also possess the four *metric properties* of Section 7.4. The ultrametric property may be expressed in terms of similarities:

$$S(\mathbf{x}_1, \mathbf{x}_2) \geq \min\,[S(\mathbf{x}_1, \mathbf{x}_3),\ S(\mathbf{x}_2, \mathbf{x}_3)] \qquad \textbf{(8.2)}$$

As an exercise, readers can verify that the five properties apply to all triplets of similarities and distances in the above matrices.

8.4 The panoply of methods

Clustering algorithms have been developed using a wide range of conceptual models and for studying a variety of problems. Sneath & Sokal (1973) propose a classification of clustering procedures. Its main dichotomies are now briefly described.

1 — Sequential versus simultaneous algorithms

Most clustering algorithms are sequential in the sense that they proceed by applying a recurrent sequence of operations to the objects. The agglomerative single linkage clustering of Section 8.2 is an example of a sequential method: the search for the equivalence relation R_c is repeated at all levels of similarity in the association matrix, up to the point where all objects are in the same cluster. In *simultaneous* algorithms, which are less frequent, the solution is obtained in a single step. Ordination techniques (Chapter 9), which may be used for delineating clusters, are of the latter type. This is also the case of the direct complete linkage clustering method presented in Section 8.9. The *K*-means (Section 8.8) and other non-hierarchical partitioning methods may be computed using sequential algorithms, although these methods are neither agglomerative nor divisive (next paragraph).

2 — Agglomeration versus division

Among the sequential algorithms, *agglomerative* procedures begin with the discontinuous partition of all objects, i.e. the objects are considered as being separate from one another. They are successively grouped into larger and larger clusters until a single, all-encompassing cluster is obtained. If the continuous partition of all objects is used instead as the starting point of the procedure (i.e. a single group containing all objects), *divisive* algorithms subdivide the group into sub-clusters, and so on until the discontinuous partition is reached. In either case, it is left to users to decide which of the intermediate partitions is to be retained, given the problem under study. Agglomerative algorithms are the most developed for two reasons. First, they are easier to program. Second, in clustering by division, the erroneous allocation of an object to a cluster at the beginning of the procedure cannot be corrected afterwards (Gower, 1967) unless a special procedure is embedded in the algorithm to do so.

3 — Monothetic versus polythetic methods

Divisive clustering methods may be monothetic or polythetic. *Monothetic* models use a single descriptor as basis for partitioning, whereas *polythetic* models use several descriptors which, in most cases, are combined into an association matrix (Chapter 7) prior to clustering. Divisive monothetic methods proceed by choosing, for each partitioning level, the descriptor considered to be the best for that level; objects are then partitioned following the state to which they belong with respect to that descriptor. For example, the most appropriate descriptor at each partitioning level could be the one that best represents the information contained in all other descriptors, after measuring the reciprocal information between descriptors (Subsection 8.6.1). When a single partition of the objects is sought, monothetic methods produce the clustering in a single step.

4 — Hierarchical versus non-hierarchical methods

In *hierarchical* methods, the members of inferior-ranking clusters become members of larger, higher-ranking clusters. Most of the time, hierarchical methods produce non-overlapping clusters, but this is not a necessity according to the definition of “hierarchy” in the dictionary or the usage recognized by Sneath & Sokal (1973). Single linkage clustering of Section 8.2 and the methods of Sections 8.5 and 8.6 are hierarchical. *Non-hierarchical methods* are very useful in ecology. They produce a single partition which optimizes within-group homogeneity, instead of a hierarchical series of partitions optimizing the hierarchical attribution of objects to clusters. Lance & Williams (1967d) restrict the term “clustering” to the non-hierarchical methods and call the hierarchical methods “classification”. Non-hierarchical methods include K-means partitioning, the ordination techniques (Chapter 9) used as clustering methods, the primary connection diagrams (dendrites) between objects with or without a reference space, the methods of similarity matrix seriation of Section 8.10, and one of the algorithms of Section 8.9 for the clustering of species into biological associations. These methods should be used in cases where the aim is to obtain a direct representation of the relationships among objects instead of a summary of their hierarchy. Hierarchical methods are easier to compute and more often available in statistical data analysis packages than non-hierarchical procedures.

Most hierarchical methods use a resemblance matrix as their starting point. This prevents their use with very large data sets because the resemblance matrix, with its $n(n-1)/2$ values, becomes extremely large and may exceed the handling capacity of computers. Jambu & Lebeaux (1983) have described a fast algorithm for the hierarchical agglomeration of very large numbers of objects (e.g. $n = 5000$). This algorithm computes a fraction only of the $n(n-1)/2$ distance values. Rohlf (1978, 1982a) has also developed a rather complex algorithm allowing one to obtain single linkage clustering after computing only a small fraction of the distances.

5 — Probabilistic versus non-probabilistic methods

Probabilistic methods include the clustering model of Clifford & Goodall (1967) and the parametric and nonparametric methods for estimating density functions in multivariate space.

In the method of Clifford & Goodall (1967), clusters are formed in such a way that the within-group association matrices have a given probability of being homogeneous. This clustering method is discussed at length in Subsection 8.9.2, where it is recommended, in conjunction with Goodall's similarity coefficient (S_{23}, Chapter 7), for the clustering of species into biological associations.

Sneath & Sokal (1973) describe other dichotomies for clustering methods, which are of lesser interest to ecologists. These are: global or local criteria, direct or iterative solutions, equal or unequal weights, and adaptive or non-adaptive clustering.

8.5 Hierarchical agglomerative clustering

Most methods of hierarchical agglomeration can be computed as special cases of a general model which is discussed in Subsection 8.5.9.

1 — Single linkage agglomerative clustering

In single linkage agglomeration (Section 8.2), two clusters fuse when the two objects closest to each other (one in each cluster) reach the similarity of the considered partition. (See also the method of simultaneous single linkage clustering described in Subsection 8.9.1). As a consequence of chaining, results of single linkage clustering are sensitive to noise in the data (Milligan, 1996), because noise changes the similarity values and may thus easily modify the order in which objects cluster. The origin of single linkage clustering is found in a collective work by Florek, Lukaszewicz, Perkal, Steinhaus, and Zubrzycki, published by Lukaszewicz in 1951.

2 — Complete linkage agglomerative clustering

Complete linkage rule

Opposite to the single linkage approach is *complete linkage agglomeration*, also called *furthest neighbour sorting*. In this method, first proposed by Sørensen (1948), the fusion of two clusters depends on the most distant pair of objects instead of the closest. Thus, an object joins a cluster only when it is linked (relationship G_c, Section 8.2) to all the objects already members of that cluster. Two clusters can fuse only when all objects of the first are linked to all objects of the second, and vice versa.

Coming back to the ponds of Ecological application 8.2, complete linkage clustering (Fig. 8.3) is performed on the table of ordered similarities of Section 8.2. The pair (212, 214) is formed at S = 0.6 and the pair (431, 432) at S = 0.5. The next clustering step must wait until S = 0.2, since it is only at S = 0.2 that pond 233 is finally linked (relationship G_c) to both ponds 431 and 432. The two clusters hence formed cannot fuse, because it is only at similarity zero that ponds 212 and 214 become linked to all the ponds of cluster (233, 431, 432). S = 0 indicating, by definition, distinct entities, the two groups are not represented as joining at that level.

In the compete linkage strategy, as a cluster grows, it becomes more and more difficult for new objects to join to it because the new objects should bear links with all the objects already in the cluster before being incorporated. For this reason, the growth of a cluster seems to move it away from the other objects or clusters in the analysis. According to Lance & Williams (1967c), this is equivalent to dilating the reference

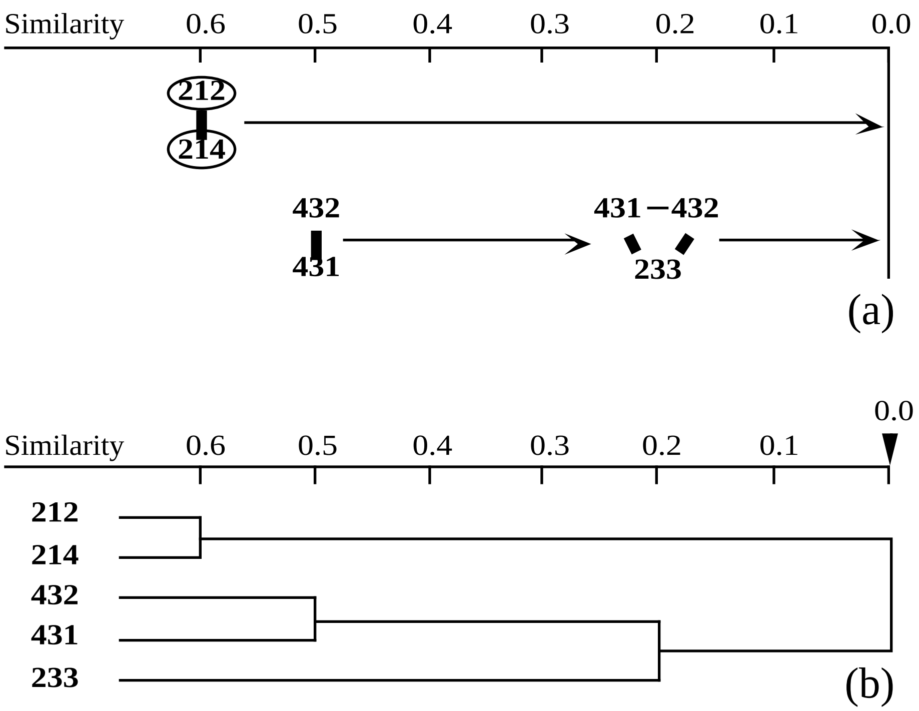


Figure 8.3 Complete linkage clustering of the ponds of Ecological application 8.2. Symbols as in Fig. 8.2.

space in the neighbourhood of that cluster; see also Fig. 8.23c and related text. This effect is opposite to what was found in single linkage clustering, which contracted the reference space. In reference space A (Fig. 7.2), complete linkage produces maximally linked and rather spherical clusters, whereas single linkage may produce elongated clusters with loose chaining. Complete linkage clustering is often desirable in ecology, when one wishes to delineate clusters with clear discontinuities.

The intermediate (next Subsection) and complete linkage clustering models have one drawback when compared to single linkage. In all cases where two incompatible candidates present themselves at the same time to be included in a cluster, algorithms use a preestablished and often arbitrary rule, called a “right-hand rule”, to choose one and exclude the other. This problem does not exist in single linkage. An example is when two objects or two clusters could be included in a third cluster, while these two objects or clusters have not completed the linkage with each other. For this problem, Sørensen (1948) recommends the following: (1) choose the fusion leading to the largest cluster; (2) if equality persists, choose the fusion that most reduces the number of clusters; (3) as a last criterion, choose the fusion that maximizes the average similarity within the cluster.

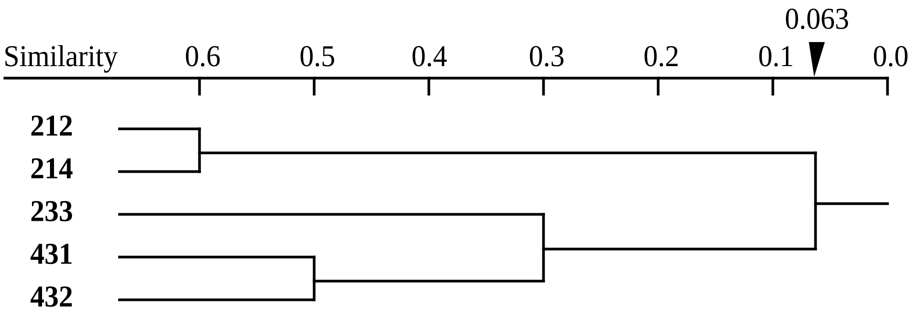


Figure 8.4 Intermediate linkage clustering, using the proportional link linkage criterion (*Co* = 50%), for the ponds of Ecological application 8.2 (dendrogram only).

3 — Intermediate linkage clustering

Between the chaining of single linkage and the extreme space dilation of complete linkage, the most interesting solution in ecology may be a type of linkage clustering that approximately conserves the metric properties of reference space A; see also Fig. 8.23b. If the interest only lies in the clusters shown in the dendrogram, and not in the actual similarity links between clusters shown by the subgraphs, the average clustering methods of Subsections 4 to 7 below could be useful since they also conserve the metric properties of the reference space.

Connectedness

In intermediate linkage clustering, the fusion criterion of an object or a cluster with another cluster is considered satisfied when a given proportion of the total possible number of similarity links is reached. For example, if the criterion of connectedness (*Co*) is 0.5, two clusters are only required to share 50% of the possible links to fuse; in other words, the fusion is authorized when $£/n_1n_2 \geq Co$ where £ is the actual number of *between-group* links at sorting level *L*, while n_1 and n_2 are the numbers of objects in the two clusters, respectively. This criterion has been called *proportional link linkage* by Sneath (1966). Fig. 8.4 gives the results of proportional link linkage clustering with *Co* = 50% for the pond example.

Proportional link linkage

Sneath (1966) has described three other ways of defining intermediate linkage clustering criteria: (1) by *integer link linkage*, which specifies the number of links required for the fusion of two groups (fusion when £ is larger than or equal to a fixed integer, *or else* when $£ = n_1n_2$); (2) by their *absolute resemblance*, based on the sum of similarities between the members of two clusters (the sum of between-group similarities, $\sum S_{12}$, must reach a given threshold before the fusion occurs); or (3) by their *relative resemblance*, where the sum of similarities between the two clusters, $\sum S_{12}$, is divided by the number of between-group similarities, n_1n_2 (fusion occurs at level *L* when the ratio $\sum S_{12}/n_1n_2$ is greater than *cL*, where *c* is an arbitrary constant.)

Table 8.2 Average clustering methods discussed in Subsections 8.5.4 to 8.5.7.

	Arithmetic average	Centroid clustering
Equal weights	4. Unweighted arithmetic average clustering (UPGMA)	6. Unweighted centroid clustering (UPGMC)
Unequal weights	5. Weighted arithmetic average clustering (WPGMA)	7. Weighted centroid clustering (WPGMC)

When c equals 1, the method is called *average linkage clustering*. These strategies are not combinatorial in the sense of Subsection 8.5.9.

4 — Unweighted arithmetic average clustering (UPGMA)

Average clustering

There are four methods of *average clustering* that conserve the metric properties of reference space A. These four methods were called “average linkage clustering” by Sneath & Sokal (1973), although they do not tally the links between clusters. As a consequence they are not object-linkage methods in the sense of the previous three subsections. They rely instead on average similarities among objects or on centroids of clusters. The four methods have nothing to do with Sneath’s (1966) “average linkage clustering” described in the previous paragraph, so that we prefer calling them “average clustering”. These methods (Table 8.2) result from the combinations of two dichotomies: (1) arithmetic average versus centroid clustering and (2) weighting versus non-weighting.

The first method in Table 8.2 is the *unweighted arithmetic average clustering* (Rohlf, 1963), also called “UPGMA” (“Unweighted Pair-Group Method using Arithmetic averages”) by Sneath & Sokal (1973) or “group-average sorting” by Lance & Williams (1966a and 1967c). It is also called “average linkage” by SAS, SYSTAT and some other statistical packages, thus adding to the confusion pointed out in the previous paragraph. The highest similarity (or smallest distance) identifies the next cluster to be formed. Following this event, the method computes the arithmetic average of the similarities or distances between a candidate object and each of the cluster members or, in the case of a previously formed cluster, between all members of the two clusters. All objects receive equal weights in the computation. The similarity or distance matrix is updated and reduced in size at each clustering step. Clustering proceeds by agglomeration as the similarity criterion is relaxed, just as it does in single linkage clustering.

For the ponds of Section 8.2, UPGMA clustering proceeds as shown in Table 8.3 and Fig. 8.5. At step 1, the highest similarity value in the matrix is

Table 8.3 Unweighted arithmetic average clustering (UPGMA) of the pond data. At each step, the highest similarity value is identified (italicized boldface value) and the two corresponding objects or groups are fused by averaging their similarities as described in the text (boxes).

Objects	212	214	233	431	432
212	—				**Step 1**
214	***0.600***	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—
212-214		—			**Step 2**
233		0.0355	—		
431		0.0315	0.300	—	
432		0.1070	0.200	***0.500***	—
212-214		—			**Step 3**
233		0.0355	—		
431-432		0.06925	***0.250***	—	
212-214		—			**Step 4**
233-431-432		***0.058***	—		

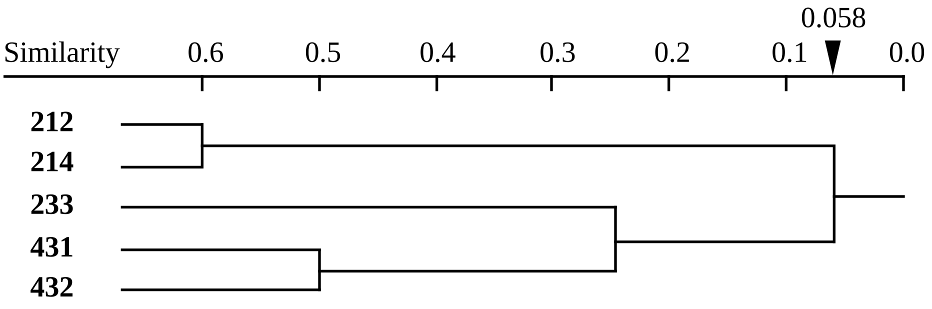


Figure 8.5 Unweighted arithmetic average clustering (UPGMA) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram. It cannot be represented by connected subgraphs since it is not a linkage clustering as in Figs. 8.2 and 8.3.

$S(212, 214) = 0.600$; hence the two objects fuse at level 0.600. As a consequence of this fusion, the similarity values of these two objects with each of the remaining objects in the study must be averaged (values in the inner boxes in the Table, step 1); this results in a reduction of the size of the similarity matrix. Considering the reduced matrix (step 2), the highest similarity value is $S = 0.500$; it indicates that objects 431 and 432 fuse at level 0.500. Again, this similarity value is obtained by averaging the boxed values; this produces a new reduced similarity matrix for the next step. In step 3, the largest similarity is 0.250; it leads to the fusion of the already-formed group (431, 432) with object 233 at level 0.250. In the example, this last fusion is the difficult point to understand. Before averaging the values, each one is multiplied by the number of objects in the corresponding group. There is one object in group (233) and two in group (431, 432), so that the fused similarity value is calculated as $[(0.0355 \times 1) + (0.06925 \times 2)]/3 = 0.058$. This is equivalent to averaging the six boxed similarities in the top matrix (larger box) with equal weights; the result would also be 0.058. So, this method is “unweighted” in the sense that it gives equal weights to the original similarities. To achieve this at step 3, one has to use weights that are equal to the number of objects in the groups. At step 4, there is a single remaining similarity value; it is used to perform the last fusion at level 0.058. In the dendrogram, fusions are drawn at the identified levels.

Because it gives equal weights to the original similarities, the UPGMA method assumes that the objects in each group form a representative sample of the corresponding larger groups of objects in the reference population under study. For that reason, UPGMA clustering should only be used in connection with simple random or systematic sampling designs if the results are to be extrapolated to some larger reference population.

Unlike linkage clustering methods, information about the relationships between pairs of objects is lost in methods based on progressive reduction of the similarity matrix, since only the relationships among groups are considered. This information may be extracted from the original similarity matrix, by making a list of the strongest similarity link found, at each fusion level, between the objects of the two groups. For the pond example, the *chain of primary connections* corresponding to the dendrogram would be made of the following links: (212, 214) for the first fusion level, (431, 432) for the second level, (233, 431) for the third level, and (214, 432) for the last level (Table 8.3, step 1). The topology obtained from UPGMA clustering may differ from that of single linkage; if this had been the case here, the chain of primary connections would have been different from that of single linkage clustering.

5 — Weighted arithmetic average clustering (WPGMA)

It often occurs in ecology that groups of objects, representing different regions of a territory, are of unequal sizes. Eliminating objects to equalize the clusters would mean discarding valuable information. However, the presence of a large group of objects, which are more similar *a priori* because of their common origin, may distort the UPGMA results when a fusion occurs with a smaller group of objects. Sokal &

Michener (1958) proposed a solution to this problem, called *weighted arithmetic average clustering* ("WPGMA" in Sneath & Sokal, 1973: "Weighted Pair-Group Method using Arithmetic averages"). This solution consists in giving equal weights, when computing fusion similarities, to the two *branches* of the dendrogram that are about to fuse. This is equivalent, when computing a fusion similarity, to giving different weights to the original similarities, i.e. down-weighting the largest group. Hence the name of the method.

Table 8.4 and Fig. 8.6 describe the WPGMA clustering sequence for the pond data. In this example, the only difference with UPGMA is the last fusion value. It is computed here by averaging the two similarities from the previous step: (0.0355 + 0.06925)/2 = 0.052375. Weighted arithmetic average clustering increases the separation of the two main clusters, compared to UPGMA. This gives sharper contrast to the classification.

6 — Unweighted centroid clustering (UPGMC)

Centroid

The *centroid* of a cluster of objects may be imagined as the type-object of the cluster, whether that object actually exists or is only a mathematical construct. In A-space (Fig. 7.2), the coordinates of the centroid of a cluster are computed by averaging the coordinates of the objects in the group.

Unweighted centroid clustering (Lance & Williams, 1967c; "UPGMC" in Sneath & Sokal, 1973: "Unweighted Pair-Group Centroid Method") is based on a simple geometric approach. Along a decreasing scale of similarities, UPGMC proceeds to the fusion of objects or clusters presenting the highest similarity, as in the previous methods. At each step, the members of a cluster are replaced by their common centroid (i.e. "mean point"). The centroid is considered to represent a new object for the remainder of the clustering procedure; in the next step, one looks again for the pair of objects with the highest similarity, on which the procedure of fusion is repeated.

Gower (1967) proposed the following formula for centroid clustering, where the similarity of the centroid (**hi**) of the objects or clusters **h** and **i** with a third object or cluster **g** is computed from the similarities $S(\mathbf{h}, \mathbf{g})$, $S(\mathbf{i}, \mathbf{g})$, and $S(\mathbf{h}, \mathbf{i})$:

$$S(\mathbf{hi}, \mathbf{g}) = \frac{w_h}{w_h + w_i} S(\mathbf{h}, \mathbf{g}) + \frac{w_i}{w_h + w_i} S(\mathbf{i}, \mathbf{g}) + \frac{w_h w_i}{(w_h + w_i)^2} [1 - S(\mathbf{h}, \mathbf{i})] \qquad \mathbf{(8.3)}$$

were the w's are weights given to the clusters. To simplify the symbols, letters **g**, **h**, and **i** are used here to represent three objects considered in the course of clustering; **g**, **h**, and **i** may also represent centroids of clusters obtained during previous clustering steps. Gower's formula insures that the centroid **hi** of objects (or clusters) **h** and **i** is geometrically located on the line between **h** and **i**. In classical centroid clustering, the numbers of objects $n_\mathbf{h}$ and $n_\mathbf{i}$ in clusters **h** and **i** are taken as values for the weights w_h and w_i; these weights are 1 at the start of the clustering because there is then a single

Table 8.4 Weighted arithmetic average clustering (WPGMA) of the pond data. At each step, the highest similarity value is identified (italicized boldface value) and the two corresponding objects or groups are fused by averaging their similarities (boxes).

Objects	212	214	233	431	432
212	—				**Step 1**
214	***0.600***	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—
212-214		—			**Step 2**
233		0.0355	—		
431		0.0315	0.300	—	
432		0.1070	0.200	***0.500***	—
212-214		—			**Step 3**
233		0.0355	—		
431-432		0.06925	***0.250***	—	
212-214		—			**Step 4**
233-431-432		***0.05238***	—		

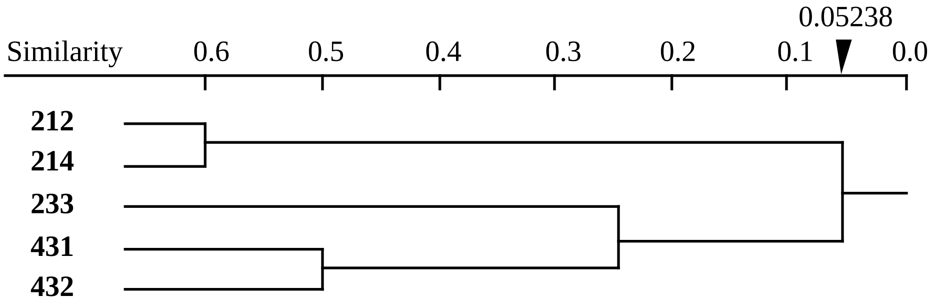


Figure 8.6 Weighted arithmetic average clustering (WPGMA) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram. It cannot be represented by connected subgraphs since it is not a linkage clustering as in Figs. 8.2 and 8.3.

object per cluster. If initial weights are attached to individual objects, they may be used instead of 1's in eq. 8.3.

Centroid clustering may lead to reversals (Section 8.6). Some authors feel uncomfortable about reversals since they violate the ultrametric property; such violations make dendrograms more difficult to draw. A reversal is found with the pond example (Table 8.5, Fig. 8.7): the fusion similarity found at step 4 is higher than that of step 3. The last fusion similarity (step 4), for instance, is calculated as follows:

$$S\,[\,(233, 431\text{-}432)\,,\,(212\text{-}214)\,] \;=\; \frac{1}{3}\times 0.1355 + \frac{2}{3}\times 0.29425 + \frac{2}{3^2}(1-0.375) \;=\; 0.38022$$

As indicated above, the geometric interpretation of UPGMC clustering is the fusion of objects into cluster centroids. Figure 8.8 presents the four clustering steps depicted by the dendrogram, drawn in an A-space (Fig. 7.2) reduced to two dimensions through principal coordinate analysis (Section 9.2) to facilitate representation. At the end of each step, a new cluster is formed and its centroid is represented at the *centre of mass* of the cluster members (examine especially steps 3 and 4).

Unweighted centroid clustering may be used with any measure of similarity, but Gower's formula above only retains its geometric properties for similarities corresponding to metric distances (Table 7.2). Note also that in this clustering procedure, the links between clusters do not depend upon identifiable pairs of objects; this was also the case with clustering methods 4 and 5 above. Thus, if the chain of primary connections is needed, it must be identified by some other method.

The assumptions of this model with respect to representativeness of the observations are the same as in UPGMA, since equal weights are given to all objects during clustering. So, UPGMC should only be used in connection with simple random or systematic sampling designs if the results are to be extrapolated to some larger reference population. When the branching pattern of the dendrogram displays asymmetry (many more objects in one branch than in the other), this can be attributed to the structure of the reference population if the sampling design was random.

In order to obtain clusters that were well separated even though the objects came from an ecological continuum, Flos (1976) provided his computer program with a strip (non-clustering zone) between centroids. The width of the zone was set by the user at the beginning of the calculations. Points found within that zone were not included in any cluster. At the end of the formation of clusters, the unclassified objects were allocated to the closest cluster centroid.

7 — Weighted centroid clustering (WPGMC)

Weighted centroid clustering was proposed by Gower (1967). It plays the same role with respect to UPGMC as WPGMA (method 5) plays with respect to UPGMA (method 4). When many observations of a given type have been included in the set to

Table 8.5 Unweighted centroid clustering (UPGMC) of the pond data. At each step, the highest similarity value is identified (italicized boldface value) and the two corresponding objects or groups are fused using eq. 8.3.

Objects	212	214	233	431	432
212	—				**Step 1**
214	***0.600***	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—
212-214		—			**Step 2**
233		0.1355	—		
431		0.1315	0.300	—	
432		0.2070	0.200	***0.500***	—
212-214		—			**Step 3**
233		0.1355	—		
431-432		0.29425	***0.375***	—	
212-214		—			**Step 4**
233-431-432		***0.3802***	—		

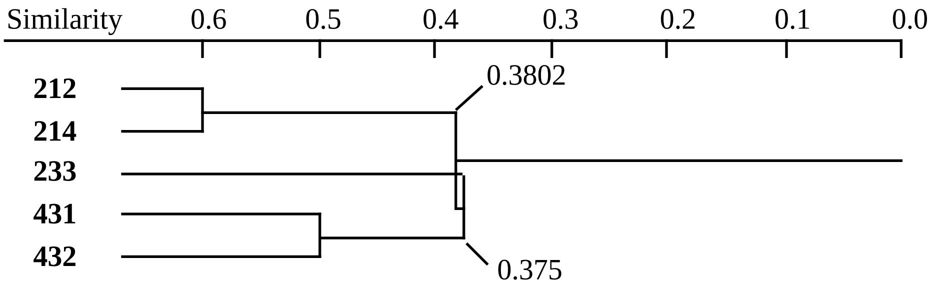


Figure 8.7 Unweighted centroid clustering (UPGMC) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram. The reversal in the structure of the dendrogram is explained in Section 8.6.

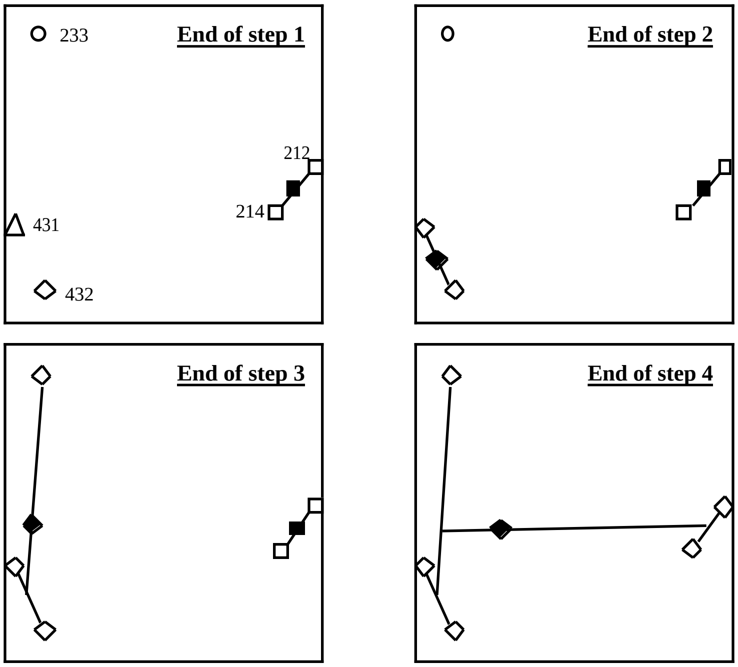


Figure 8.8 The four UPGMC clustering steps of Fig. 8.7 are drawn in A-space. Objects are represented by open symbols and centroids by dark symbols; object identifiers are shown in the first panel only. Distinct clusters are represented by different symbols. The first two principal coordinates, represented here, account for 87.4% of the variation of the full A-space.

be clustered, next to other types which were not as well-sampled (sampling design other than simple random or systematic), the positions of the centroids may be biased towards the over-represented types, which in turn could distort the clustering. In *weighted centroid clustering*, which Sneath & Sokal (1973) call “WPGMC” (“Weighted Pair-Group Centroid Method”), this problem is corrected by giving equal weights to two clusters on the verge of fusing, independently of the number of objects in each cluster. To achieve this, eq. 8.3 is replaced by the following formula (Gower, 1967):

$$S(\mathbf{hi}, \mathbf{g}) = \frac{1}{2}[S(\mathbf{h}, \mathbf{g}) + S(\mathbf{i}, \mathbf{g})] + \frac{1}{4}[1 - S(\mathbf{h}, \mathbf{i})] \qquad \textbf{(8.4)}$$

The five ponds of Ecological application 7.2 are clustered as described in Table 8.6 and Fig. 8.9. The last fusion similarity (step 4), for example, is calculated as follows:

$$S[(233, 431\text{-}432), (212\text{-}214)] = \frac{1}{2}[0.1355 + 0.29425] + \frac{1}{4}(1 - 0.375) = 0.371125$$

Table 8.6 Weighted centroid clustering (WPGMC) of the pond data. At each step, the highest similarity value is identified (italicized boldface value) and the two corresponding objects or groups are fused using eq. 8.4.

Objects	212	214	233	431	432
212	—				**Step 1**
214	***0.600***	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—
212-214		—			**Step 2**
233		0.1355	—		
431		0.1315	0.300	—	
432		0.2070	0.200	***0.500***	—
212-214		—			**Step 3**
233		0.1355	—		
431-432		0.29425	***0.375***	—	
212-214		—			**Step 4**
233-431-432		***0.37113***	—		

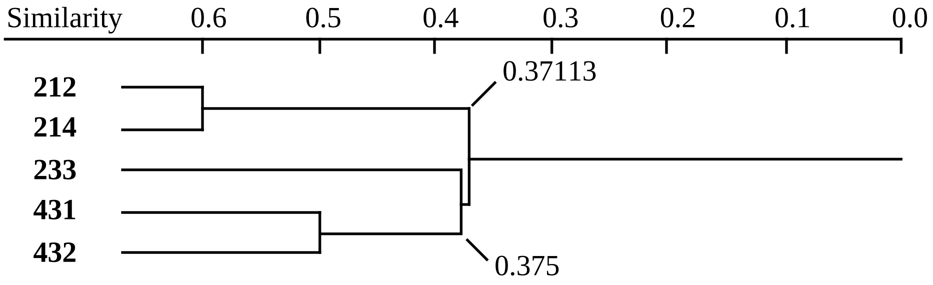


Figure 8.9 Weighted centroid clustering (WPGMC) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram.

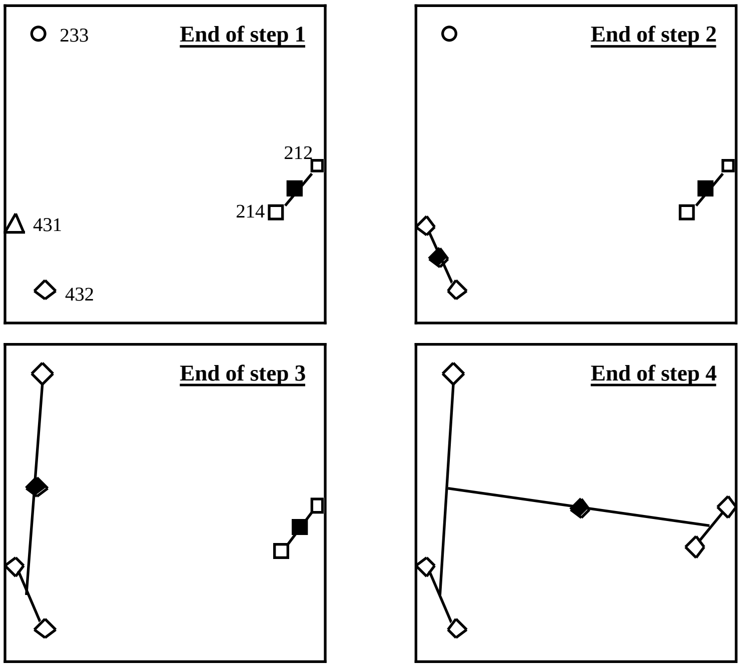


Figure 8.10 The four WPGMC clustering steps of Fig. 8.9 are drawn in A-space. Objects are represented by open symbols and centroids by dark symbols; object identifiers are shown in the first panel only. Distinct clusters are represented by different symbols. The first two principal coordinates, represented here, account for 87.4% of the variation of the full A-space.

This value is the level at which the next fusion takes place. Note that no reversal appears in this result, although WPGMC may occasionally produce reversals, like UPGMC clustering.

As indicated above, the geometric interpretation of WPGMC clustering is the fusion of objects into cluster centroids. Fig. 8.10 presents the four clustering steps depicted by the dendrogram, in A-space (Fig. 7.2) reduced to two dimensions through principal coordinate analysis (Section 9.2) to facilitate representation. At the end of each step, a new cluster is formed and its centroid is represented at the *geometric centre* of the last line drawn (examine especially steps 3 and 4 and compare to Fig. 8.8).

In the R mode, weighted centroid clustering does not make sense if the measure of association is Pearson's *r*. Correlations are cosine transformations of the angles between descriptors; these cannot be combined using eq. 8.4.

8 — Ward's minimum variance method

Ward's (1963) minimum variance method is related to the centroid methods (Subsections 6 and 7 above) in that it also leads to a geometric representation in which cluster centroids play an important role. To form clusters, the method minimizes an *objective function* which is, in this case, the same "squared error" criterion as that used in multivariate analysis of variance.

Objective function

At the beginning of the procedure, each objects is in a cluster of its own, so that the distance of an object to its cluster's centroid is 0; hence, the sum of all these distances is also 0. As clusters form, the centroids move away from actual object coordinates and the sums of the squared distances from the objects to the centroids increase. At each clustering step, Ward's method finds the pair of objects or clusters whose fusion increases as little as possible the sum, over all objects, of the *squared distances* between objects and cluster centroids. The distance of object $\mathbf{x}_i$ to the centroid $\mathbf{m}$ of its cluster is computed using the Euclidean distance formula (eq. 7.33) over the various descriptors $\mathbf{y}_j$ ($j = 1 \dots p$):

$$\sum_{j=1}^{p} [y_{ij} - m_j]^2$$

The centroid $\mathbf{m}$ of a cluster was defined at the beginning of Subsection 8.5.6. The sum of squared distances of all objects in cluster k to their common centroid is called "error" in ANOVA, hence the symbol e_k^2:

Squared error

$$e_k^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{p} [y_{ij}^{(k)} - m_j^{(k)}]^2 \qquad \textbf{(8.5)}$$

where $y_{ij}^{(k)}$ is the value of descriptor $\mathbf{y}_j$ for an object i member of group (k) and $m_j^{(k)}$ is the mean value of descriptor j over all members of group k. Alternatively, the within-cluster sums of squared errors e_k^2 can be computed as the mean of the squared distances among cluster members:

$$e_k^2 = \left[\sum_{h,i=1}^{n_k} D_{hi}^2 \right] \Big/ n_k \qquad \textbf{(8.6)}$$

where the D_{hi}^2 are the squared distances among objects in cluster k (Table 8.7) and n_k is the number of objects in that cluster. Equations 8.5 and 8.6 both allow the calculation of the squared error criterion. The equivalence of these two equations is stated in a theorem whose demonstration, for the univariate case, is found in Kendall & Stuart (1963, parag. 2.22). Numerical examples illustrating the calculation of eqs. 8.5 and 8.6 are given at the end of Section 8.8 (K-means partitioning).

Table 8.7 Ward's minimum variance clustering of the pond data. Step 1 of the table contains *squared distances* computed as $D^2 = (1 - S)^2$ from the similarity values in Tables 8.3 to 8.6. At each step, the lowest squared distance is identified (italicized boldface value) and the two corresponding objects or groups are fused using eq. 8.10.

Objects	212	214	233	431	432
212	—				**Step 1**
214	***0.16000***	—			
233	1.00000	0.86304	—		
431	1.00000	0.87797	0.49000	—	
432	1.00000	0.61780	0.64000	0.25000	—
212-214		—			**Step 2**
233		1.18869	—		
431		1.19865	0.49000	—	
432		1.02520	0.64000	***0.25000***	—
212-214		—			**Step 3**
233		1.18869	—		
431-432		1.54288	***0.67000***	—	
212-214		—			**Step 4**
233-431-432		***1.67952***	—		

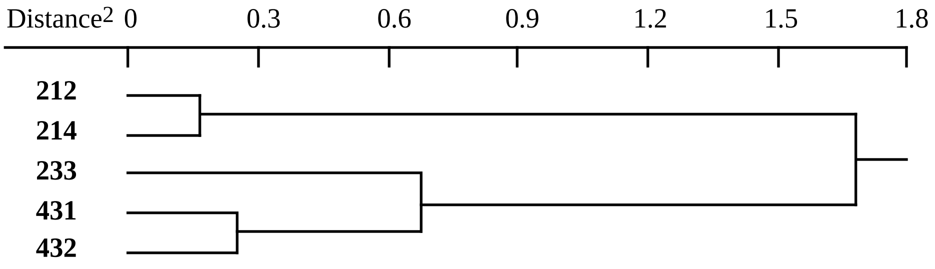


Figure 8.11 Ward's minimum variance clustering of the ponds from Ecological application 8.2. The scale of this dendrogram is the squared distances computed in Table 8.7.

The sum of squared errors E_K^2, for all K clusters corresponding to given partition, is the criterion to be minimized at each clustering step:

Sum of squared errors

$$E_K^2 = \sum_{k=1}^{K} e_k^2 \qquad \textbf{(8.7)}$$

At each clustering step, two objects or clusters **h** and **i** are merged into a new cluster **hi**, as in previous sections. Since changes occurred only in the groups **h**, **i**, and **hi**, the change in the overall sum of squared errors, $\Delta E_{\mathbf{hi}}^2$, may be computed from the changes that occurred in these groups only:

$$\Delta E_{\mathbf{hi}}^2 = e_{\mathbf{hi}}^2 - e_{\mathbf{h}}^2 - e_{\mathbf{i}}^2 \qquad \textbf{(8.8)}$$

It can be shown that this change depends only on the distance between the centroids of clusters **h** and **i** and on their numbers of objects $n_{\mathbf{h}}$ and $n_{\mathbf{i}}$ (Jain & Dubes, 1988):

$$\Delta E_{\mathbf{hi}}^2 = \frac{n_{\mathbf{h}} n_{\mathbf{i}}}{n_{\mathbf{h}} + n_{\mathbf{i}}} \sum_{j=1}^{p} [m_j^{(\mathbf{h})} - m_j^{(\mathbf{i})}]^2 \qquad \textbf{(8.9)}$$

So, one way of identifying the next fusion is to compute the $\Delta E_{\mathbf{hi}}^2$ statistic for all possible pairs and select the pair which generates the smallest value of this statistic for the next merge. Another way is to use the following updating formula to compute the fusion distances between the new cluster **hi** and all other objects or clusters **g**, in the agglomeration table (Table 8.7):

$$D^2(\mathbf{hi}, \mathbf{g}) = \frac{n_{\mathbf{h}} + n_{\mathbf{g}}}{n_{\mathbf{h}} + n_{\mathbf{i}} + n_{\mathbf{g}}} D^2(\mathbf{h}, \mathbf{g}) + \frac{n_{\mathbf{i}} + n_{\mathbf{g}}}{n_{\mathbf{h}} + n_{\mathbf{i}} + n_{\mathbf{g}}} D^2(\mathbf{i}, \mathbf{g}) - \frac{n_{\mathbf{g}}}{n_{\mathbf{h}} + n_{\mathbf{i}} + n_{\mathbf{g}}} D^2(\mathbf{h}, \mathbf{i}) \qquad \textbf{(8.10)}$$

Squared distances are used instead of similarities in eq. 8.10 and in Table 8.7.

Dendrograms for Ward's clustering may be represented along a variety of scales. They all lead to the same clustering topology.

- Fig. 8.11 uses the same scale of squared distances as Table 8.7. This is the solution advocated by Jain & Dubes (1988) and other authors.

- One can easily compute the *square roots* of the fusion distances of Table 8.7 and draw the dendrogram accordingly. This solution, illustrated in Fig. 8.12a, removes the distortions created by squaring the distances. It is especially suitable when one wants to compare the fusion distances of Ward's clustering to the original distances, either graphically (Shepard-like diagrams, Fig. 8.23) or numerically (cophenetic correlations, Subsection 8.11.2).

- The sum of squared errors E_K^2 (eq. 8.7) is used in some computer programs as the clustering scale. This statistic is also called the *total error sum of squares* (TESS) by Everitt (1980) and other authors. This solution is illustrated in Fig. 8.12b.

TESS

- The SAS package (1985) recommends two scales for Ward's clustering. The first one is the proportion of variance (R^2) accounted for by the clusters at any given partition level. It is computed as the total sum of squares (i.e. the sum of squared distances from the centroid of all objects) minus the within-cluster squared errors E_K^2 of eq. 8.7 for the given partition, divided by the total sum of squares. R^2 decreases as clusters grow. When all the objects are lumped in a single cluster, the resulting one-cluster partition does not explain any of the objects' variation so that $R^2 = 0$. The second scale recommended by SAS is called the *semipartial* R^2. It is computed as the between-cluster sum of squares divided by the (corrected) total sum of squares. This statistic increases as the clusters grow.

Like the K-means partitioning method (Section 8.8), Ward's agglomerative clustering can be computed from either a raw data table using eq. 8.8, or a matrix of squared distances through eq. 8.10. The latter is the most usual approach in computer programs. It is important to note that distances are computed as (squared) Euclidean

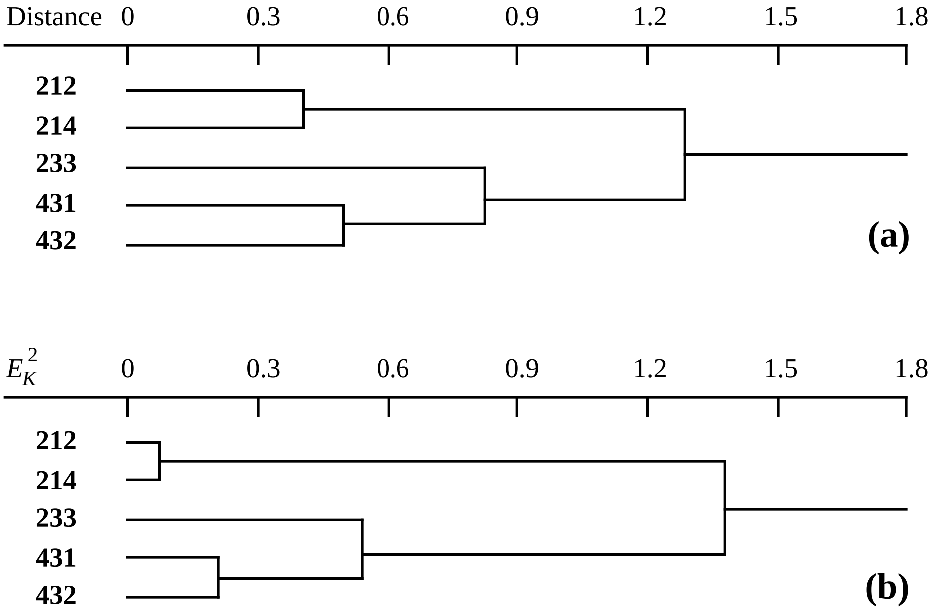


Figure 8.12 Ward's minimum variance clustering of the ponds from Ecological application 8.2. The scale of dendrogram (a) is the square root of the squared distances computed in Table 8.7; in dendrogram (b), it is the E_K^2 (or TESS) statistic.

distances in Ward's method. So, unless the descriptors are such that Euclidean distances (D_1 in Chapter 7) are an appropriate model for the relationships among objects, one should not use a Ward's algorithm based or raw data. It is preferable in such cases to compute a distance matrix using an appropriate coefficient (Tables 7.3 to 7.5), followed by clustering of the resemblance matrix in A-space by a distance-based Ward algorithm. Section 9.2 will show that resemblance matrices can also be used for plotting the positions of objects in A-space, "as if" the distances were Euclidean.

Because of the squared error criterion used as the objective function to minimize, clusters produced by the Ward minimum variance method tend to be hyperspherical, i.e. spherical in multidimensional A-space, and to contain roughly the same number of objects if the observations are evenly distributed through A-space. The same applies to the centroid methods of the previous subsections. This may be seen as either an advantage or a problem, depending on the researcher's conceptual model of a cluster.

9 — General agglomerative clustering model

Lance & Williams (1966a, 1967c) have proposed a general model that encompasses all the agglomerative clustering methods presented up to now, except intermediate linkage (Subsection 3). The general model offers the advantage of being translatable into a single, simple computer program, so that it is used in most statistical packages that offer agglomerative clustering. The general model allows one to select an agglomerative clustering model by choosing the values of four parameters, called α_h, α_i, β, and γ, which determine the clustering strategy. This model only outputs the branching pattern of the clustering tree (the dendrogram), as it was the case for the methods described in Subsections 8.5.4 to 8.5.8. For the linkage clustering strategies (Subsections 8.5.1 to 8.5.3), the list of links responsible for cluster formation may be obtained afterwards by comparing the dendrogram to the similarity matrix.

Combinatorial method

The model of Lance & Williams is limited to *combinatorial* clustering methods, i.e. those for which the similarity $S(\mathbf{hi}, \mathbf{g})$ between an external cluster **g** and a cluster **hi**, resulting from the prior fusion of clusters **h** and **i**, is a function of the three similarities $S(\mathbf{h}, \mathbf{g})$, $S(\mathbf{i}, \mathbf{g})$, and $S(\mathbf{h}, \mathbf{i})$ and also, eventually, the numbers $n_\mathbf{h}$, $n_\mathbf{i}$, and $n_\mathbf{g}$ of objects in clusters **h**, **i**, and **g**, respectively (Fig. 8.13). Individual objects are considered to be single-member clusters. Since the similarity of cluster **hi** with an external cluster **g** can be computed from the above six values, **h** and **i** can be condensed into a single row and a single column in the updated similarity matrix; following that, the clustering proceeds as in the Tables of the previous Subsections. Since the new similarities at each step can be computed by *combining* those from the previous step, it is not necessary for a computer program to retain the original similarity matrix or data set. Non-combinatorial methods do not have this property. For similarities, the general model for combinatorial methods is the following:

$$S(\mathbf{hi}, \mathbf{g}) = (1 - \alpha_h - \alpha_i - \beta) + \alpha_h S(\mathbf{h}, \mathbf{g}) + \alpha_i S(\mathbf{i}, \mathbf{g}) + \beta S(\mathbf{h}, \mathbf{i}) - \gamma \left| S(\mathbf{h}, \mathbf{g}) - S(\mathbf{i}, \mathbf{g}) \right| \tag{8.11}$$

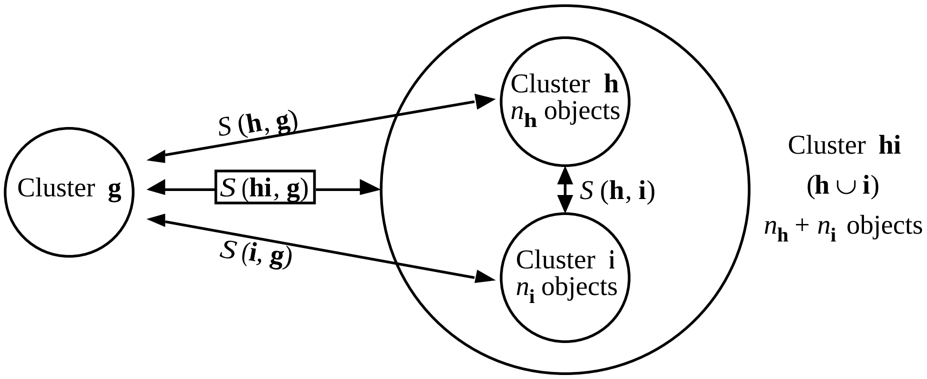


Figure 8.13 In combinatorial clustering methods, the similarity between a cluster **hi**, resulting from the fusion of two previously formed clusters **h** and **i**, and an external cluster **g** is a function of the three similarities between (**h** and **i**), (**h** and **g**), and (**i** and **g**), and of the number of objects in **h**, **i**, and **g**.

When using distances, the combinatorial equation becomes:

$$D(\mathbf{hi}, \mathbf{g}) = \alpha_h D(\mathbf{h}, \mathbf{g}) + \alpha_i D(\mathbf{i}, \mathbf{g}) + \beta D(\mathbf{h}, \mathbf{i}) + \gamma \left| D(\mathbf{h}, \mathbf{g}) - D(\mathbf{i}, \mathbf{g}) \right| \qquad \textbf{(8.12)}$$

Clustering proceeds in the same way for all combinatorial agglomerative methods. As the similarity decreases, a new cluster is obtained by the fusion of the two most similar objects or groups, after which the algorithm proceeds to the fusion of the two corresponding rows and columns in the similarity (or distance) matrix using eq. 8.11 or 8.12. The matrix is thus reduced by one row and one column at each step. Table 8.8 gives the values of the four parameters for the most commonly used combinatorial agglomerative clustering strategies. Values of the parameters for some other clustering strategies are given by Gordon (1996a).

In the case of equality between two mutually exclusive pairs, the decision may be made on an arbitrary basis (the so-called “right-hand rule” used in most computer programs) or based upon ecological criteria (as, for example, Sørensen's criteria reported at the end of Subsection 8.5.2, or those explained in Subsection 8.9.1).

In several strategies, $\alpha_h + \alpha_i + \beta = 1$, so that the term $(1 - \alpha_h - \alpha_i - \beta)$ becomes zero and disappears from eq. 8.11. One can show how the values chosen for the four parameters make the general equation correspond to each specific clustering method. For single linkage clustering, for instance, the general equation becomes:

$$S(\mathbf{hi}, \mathbf{g}) = \frac{1}{2}\left[S(\mathbf{h}, \mathbf{g}) + S(\mathbf{i}, \mathbf{g}) + \left| S(\mathbf{h}, \mathbf{g}) - S(\mathbf{i}, \mathbf{g}) \right|\right]$$

Table 8.8 Values of parameters α_h, α_i, β, and γ in Lance and Williams' general model for combinatorial agglomerative clustering. Modified from Sneath & Sokal (1973) and Jain & Dubes (1988).

Clustering method	α_h	α_i	β	γ	Effect on space A
Single linkage	1/2	1/2	0	–1/2	Contracting*
Complete linkage	1/2	1/2	0	1/2	Dilating*
UPGMA	$\frac{n_h}{n_h + n_i}$	$\frac{n_i}{n_h + n_i}$	0	0	Conserving*
WPGMA	1/2	1/2	0	0	Conserving
UPGMC	$\frac{n_h}{n_h + n_i}$	$\frac{n_i}{n_h + n_i}$	$\frac{-n_h n_i}{(n_h + n_i)^2}$	0	Conserving
WPGMC	1/2	1/2	–1/4	0	Conserving
Ward's	$\frac{n_h + n_g}{n_h + n_i + n_g}$	$\frac{n_i + n_g}{n_h + n_i + n_g}$	$\frac{-n_g}{n_h + n_i + n_g}$	0	Conserving
Flexible	$\frac{1 - \beta}{2}$	$\frac{1 - \beta}{2}$	$-1 \le \beta < 1$	0	Contracting if $\beta \approx 1$ Conserving if $\beta \approx -.25$ Dilating if $\beta \approx -1$

* Terms used by Sneath & Sokal (1973).

The last term (absolute value) completes the smallest of the two similarities S(**h**, **g**) and S(**i**, **g**), making it equal to the largest one. Hence, S(**hi**, **g**) = max[S(**h**, **g**), S(**i**, **g**)]. In other words, the similarity between a newly-formed cluster **hi** and some other cluster **g** becomes equal to the largest of the similarity values previously computed between the two original clusters (**h** and **i**) and **g**.

Intermediate linkage clustering is not a combinatorial strategy. All along the clustering procedure, it is necessary to refer to the original association matrix in order to calculate the connectedness of pairs of clusters. This is why it cannot be obtained using the Lance & Williams general agglomerative clustering model.

10 — Flexible clustering

Lance & Williams (l966a. 1967c) proposed to vary the parameter β (eq. 8.11 or 8.12) between –1 and +1 to obtain a series of intermediates solutions between single linkage chaining and the space dilation of complete linkage. The method is also called *beta-flexible clustering* by some authors. Lance & Williams (*ibid.*) have shown that, if the other parameters are constrained a follows:

$$\alpha_h = \alpha_i = (1 - \beta)/2 \qquad \text{and} \qquad \gamma = 0$$

the resulting clustering is ultrametric (no reversals; Section 8.6).

When β is close to 1, strong chaining is obtained. As β decreases and becomes negative, space dilation increases. The space properties are conserved for small negative values of β, near –0.25. Figure 8.14 shows the effect of varying β in the clustering of 20 objects. Like weighted centroid clustering, flexible clustering is compatible with all association measures except Pearson's *r*.

Ecological application 8.5a

Pinel-Alloul *et al.* (1990) studied phytoplankton in 54 lakes of Québec to determine the effects of acidification, physical and chemical characteristics, and lake morphology on species assemblages. Phytoplankton was enumerated into five main taxonomic categories (microflagellates, chlorophytes, cyanophytes, chrysophytes, and pyrrophytes). The data were normalized using the generalized form of the Box-Cox method that finds the best normalizing transformation for all species (Subsection 1.5.6). A Gower (S_{19}) similarity matrix, computed among lakes, was subjected to flexible clustering with parameter $\beta = -0.25$. Six clusters were found, which were roughly distributed along a NE-SW geographic axis and corresponded to increasing concentrations of total phytoplankton, chlorophytes, cyanophytes, and microflagellates. Explanation of the phytoplankton-based lake typology was sought by comparing it to the environmental variables (Section 10.2.1).

11 — Information analysis

The Q-mode clustering method called *information analysis* was developed for ecological purposes by Williams *et al.* (1966) and Lance & Williams (1966b). It does not go through the usual steps of similarity calculation followed by clustering. It is a direct method of clustering, based on information measures.

Entropy

Shannon's formula (eq. 6.1) can be used to measure the diversity or information in a frequency or probability distribution:

$$H = -\sum_{j=1}^{p} \mathrm{p}_j \log \mathrm{p}_j$$

Information analysis is a type of unweighted centroid clustering, adapted to species data. At each step, the two objects or clusters causing the smallest gain in within-group diversity (or information) are fused. As a consequence, the clusters are as homogeneous as possible in terms of species composition.

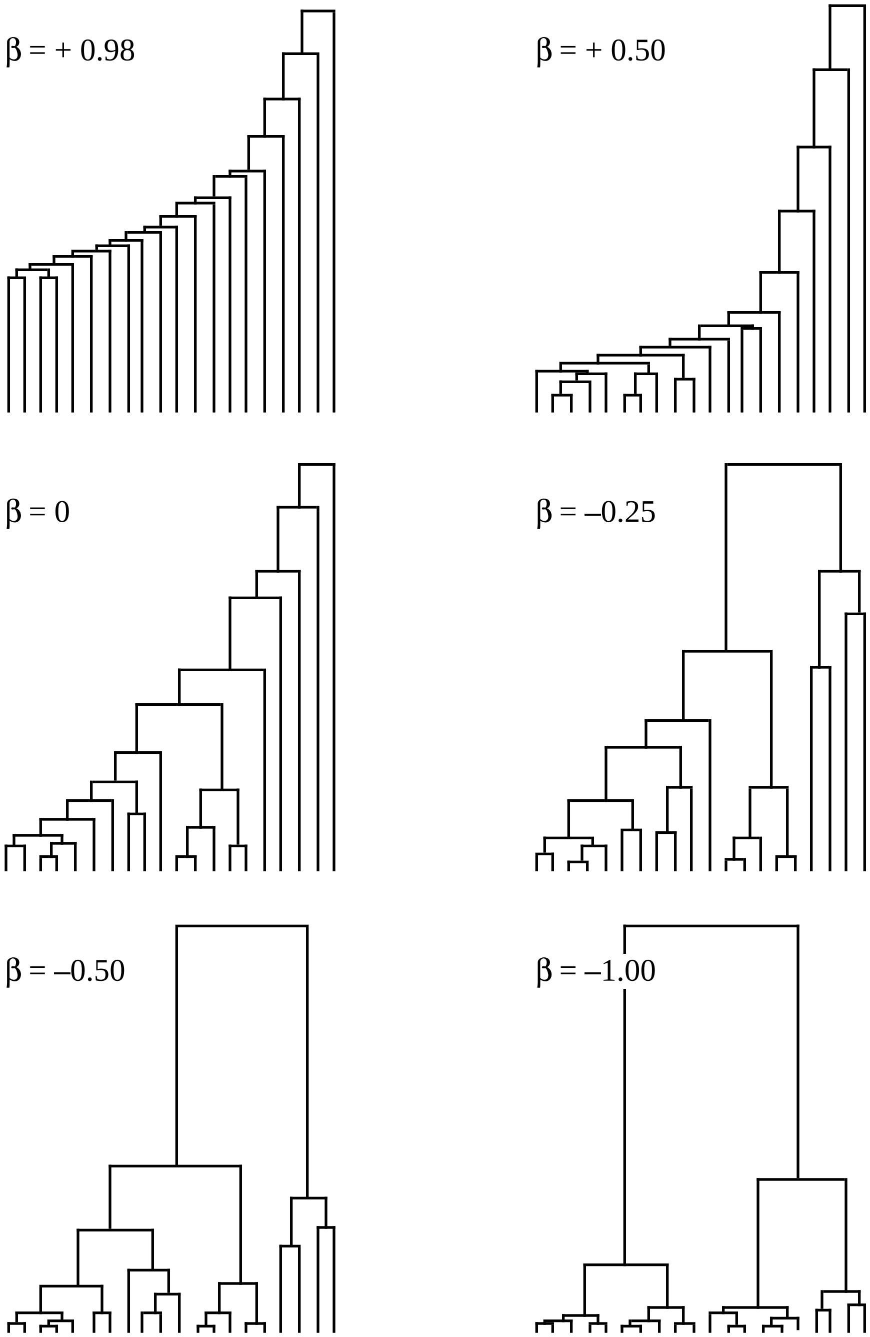


Figure 8.14 Flexible clustering of 20 objects for six values of β. The measure of association is the squared Euclidean distance D_1^2. Adapted from Lance & Williams (1967c: 376).

The method could be applied to species abundance data, divided into a small number of classes but, in practice, it is mostly used with presence-absence data. The information measure described below is not applicable to raw abundance data because the number of different states would then vary from one species to another, which would give them different weights in the overall measure.

To illustrate the method, the pond zooplankton counts used in Chapter 7 (coefficient S_{23}) are transformed here into presence-absence data (see also Ecological application 8.2):

Species j	Ponds					p_j	$(1-\mathrm{p}_j)$
	212	214	233	431	432		
1	1	1	0	0	0	0.4	0.6
2	0	0	1	1	0	0.4	0.6
3	0	1	1	0	1	0.6	0.4
4	0	0	1	1	1	0.6	0.4
5	1	1	0	0	0	0.4	0.6
6	0	1	0	1	1	0.6	0.4
7	0	0	0	1	1	0.4	0.6
8	1	1	0	0	0	0.4	0.6

Total information in this group of ponds is computed using an information measure derived from the following reasoning (Lance & Williams, 1966b). The entropy of each species presence-absence descriptor j is calculated on the basis of the probabilities of presence p_j and absence $(1-\mathrm{p}_j)$ of species j, which are written in the right-hand part of the table. The probability of presence is estimated as the number of ponds where species j is present, divided by the total number of ponds *in the cluster under consideration* (here, the group of five ponds). The probability of absence is estimated likewise, using the number of ponds where species j is absent. The entropy of species j is therefore:

$$H(j) = -[\mathrm{p}_j \log \mathrm{p}_j + (1-\mathrm{p}_j)\log(1-\mathrm{p}_j)] \qquad \text{for } 0 < \mathrm{p}_j < 1 \qquad \textbf{(8.13)}$$

The base of the logarithms is indifferent, as long as the same base is used throughout the calculations. Natural logarithms are used throughout the present example. For the first species, $H(1)$ would be:

$$H(1) = -[0.4\ \ln(0.4) + 0.6\ \ln(0.6)] = 0.673$$

The information of the conditional probability table can be calculated by summing the entropies per species, considering that all species have the same weight. Since the measure of *total information* in the group must also take into account the number of objects in the cluster, it is defined as follows:

$$I = -n \sum_{j=1}^{p} [\mathrm{p}_j \log \mathrm{p}_j + (1 - \mathrm{p}_j) \log (1 - \mathrm{p}_j)] \quad \text{for } 0 < \mathrm{p}_j < 1 \qquad \textbf{(8.14)}$$

where p is the number of species represented in the group of n objects (ponds). For the group of 5 ponds above,

$$I = –5 [8 (–0.673)] = 26.920$$

If I is to be expressed as a function of the number a_j of ponds with species j present, instead of a function of probabilities $\mathrm{p}_j = \mathrm{a}_j/n$, it can be shown that the following formula is equivalent to eq. 8.14:

$$I = np \log n - \sum_{j=1}^{p} [\mathrm{a}_i \log \mathrm{a}_j + (n - \mathrm{a}_j) \log (n - \mathrm{a}_j)] \qquad \textbf{(8.15)}$$

I is zero when all ponds in a group contain the same species. Like entropy H, I has no upper limit; its maximum value depends on the number of species present in the study.

At each clustering step, three series of values are considered: (a) the total information I in each group, which is 0 at the beginning of the process since each object (pond) then forms a distinct cluster; (b) the value of I for all possible combinations of groups taken two at a time; (c) the increase of information ΔI resulting from each possible fusion. As recommended by Sneath & Sokal (1973), all these values can be placed in a matrix, initially of dimension $n \times n$ which decreases as clustering proceeds. For the example data, values (a) of information in each group are placed on the diagonal, values (b) in the lower triangle, and values (c) of ΔI in the upper triangle, in italics.

Ponds	Ponds				
	212	214	233	431	432
212	0	*2.773*	*8.318*	*9.704*	*9.704*
214	2.773	0	*8.318*	*9.704*	*6.931*
233	8.318	8.318	0	*4.159*	*4.159*
431	9.704	9.704	4.159	0	*2.773*
432	9.704	6.931	4.159	2.773	0

The ΔI for two groups is found by subtracting the corresponding values I, on the diagonal, from the value I of their combination in the lower triangle. Values on the diagonal are 0 in this first calculation matrix, so that values in the upper triangle are the same as in the lower triangle, but this will not be the case in subsequent matrices.

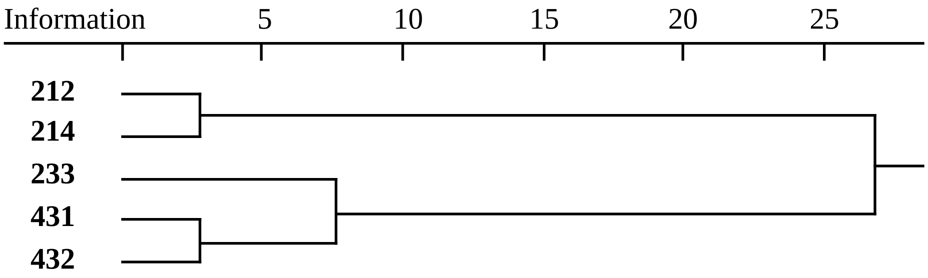


Figure 8.15 Clustering of the ponds from Ecological application 8.2, using information analysis.

The first fusion is identified by the lowest Δ*I* value found in the upper triangle. This value is 2.773 for pairs (212, 214) and (431, 432), which therefore fuse. A new matrix of *I* values is computed:

Groups	Groups		
	212 214	233	431 432
212-214	2.773	*10.594*	*15.588*
233	13.367	0	*4.865*
431-432	21.134	7.638	2.773

This time, the Δ*I* values in the upper triangle differ from the *I*'s in the lower triangle since there are *I* values on the diagonal. The Δ*I* corresponding to group (212, 214, 431, 432), for example, is computed as: 21.134 – 2.773 – 2.773 = 15.588. The lowest value of Δ*I* is for the group (233, 431, 432), which therefore fuses at this step.

For the last clustering step, the only *I* value to calculate in the lower triangle is for the cluster containing the five ponds. This value is 26.920, as computed above from eq. 8.14. Δ*I* is then 26.920 – 2.773 – 7.638 = 16.509.

Groups	Groups	
	212 214	233 431-432
212-214	2.773	*16.509*
233-431-432	26.920	7.638

The last fusion occurs at *I* = 26.920; computing Δ*I* would not have been necessary in this case. The values of *I* can be used as the scale for a dendrogram summarizing the clustering steps (Fig. 8.15).

According to Williams *et al.* (1966), information analysis minimizes chaining and quickly delineates the main clusters, at least with ecological data. Field (1969) pointed out, however, that information analysis bases the similarity between objects on double absences as well as double presences. This method may therefore not be appropriate when a gradient has been sampled and the data matrix contains many zeros; see Section 7.3 and Subsection 9.4.5 for a discussion of this problem.

Efficiency coefficient

The inverse of ΔI is known as the *efficiency coefficient* (Lance & Williams, 1966b). An analogue to the efficiency coefficient may be computed for dendrograms obtained using other agglomerative clustering procedures. In that case, the efficiency coefficient is still computed as $1/\Delta$, where Δ represents the amount by which the information in the classification is reduced due to the fusion of groups. The reduction is computed as the entropy in the classification before a fusion level minus the entropy after that fusion. In Fig. 8.2 for instance, the partition at $S = 0.40$ contains three groups of 2, 2, and 1 objects respectively; using natural logarithms, Shannon's formula (eq. 6.1) gives $H = 1.05492$. The next partition, at $S = 0.25$, contains two groups of 2 and 3 objects respectively; Shannon's formula gives $H = 0.67301$. The difference is $\Delta = 0.38191$, hence the efficiency coefficient $1/\Delta = 2.61843$ for the dendrogram fusion level $S = 0.3$.

When $1/\Delta I$ is high, the procedure clusters objects that are mostly alike. The efficiency coefficient does not monotonically decrease as the clustering proceeds. With real data, it may decrease, reach a minimum, and increase again. If $1/\Delta I$ is plotted as a function of the successive fusion levels, the minima in the graph indicate the most important partitions. If one wants to select a single cutting level through a dendrogram, this graph may help in deciding which partition should be selected. In Fig. 8.2 for example, one would choose the value $1/\Delta I = 1.48586$, which corresponds to the last fusion level ($S = 0.214$), as the most informative partition. The efficiency coefficient is not a rigorous decision criterion, however, since no test of statistical significance is performed.

8.6 Reversals

Reversals may occasionally occur in the clustering structure when using UPGMC or WPGMC (Subsections 8.5.6 and 8.5.7), or with some unusual combinations of parameters in the general agglomerative model of Lance & Williams (Subsection 8.5.9). As an example, a reversal was produced in Fig. 8.7. Two types of situations lead to reversals:

- When $\mathbf{x}_1$ and $\mathbf{x}_2$ cluster first, because they represent the closest pair, although the distance from $\mathbf{x}_3$ to the centroid $\mathbf{c}_{12}$ is smaller than the distance from $\mathbf{x}_1$ to $\mathbf{x}_2$ (Fig. 8.16a).

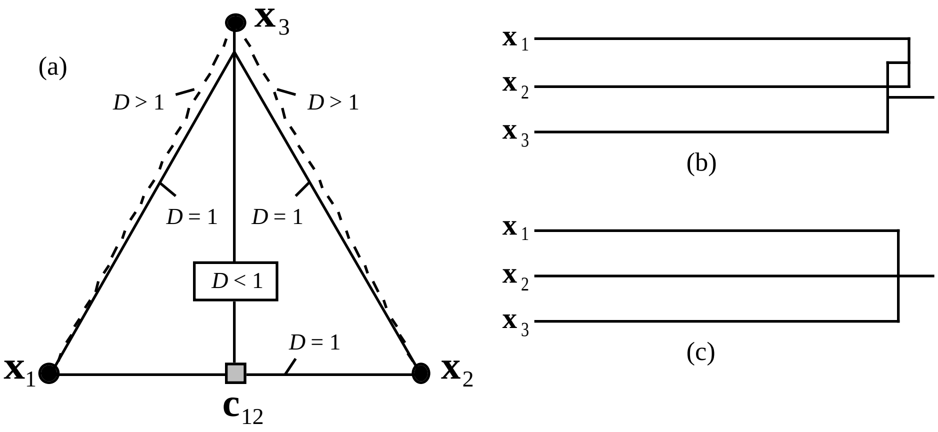


Figure 8.16 A reversal may occur in situations such as (a), where $\mathbf{x}_1$ and $\mathbf{x}_2$ cluster first because they represent the closest pair, although the distance from $\mathbf{x}_3$ to the centroid $\mathbf{c}_{12}$ is smaller than the distance from $\mathbf{x}_1$ to $\mathbf{x}_2$. (b) The result is usually depicted by a non-ultrametric dendrogram. (c) The reversal may also be interpreted as a trichotomy.

- When $D(\mathbf{x}_1, \mathbf{x}_2) = D(\mathbf{x}_1, \mathbf{x}_3) = D(\mathbf{x}_2, \mathbf{x}_3)$. In such a situation, most computer programs use an arbitrary rule (“right-hand rule”) and first cluster two of the three objects. A reversal appears when the third object is added to the cluster.

When this happens, the cophenetic matrix (Subsection 8.3.1) violates the ultrametric property (Subsection 8.3.2) and the dendrogram is more difficult to draw than in the no-reversal cases (Fig. 8.16b). However, departures from ultrametricity are never large in practice. For this reason, a reversal may be interpreted as nearly equivalent to a trichotomy in the hierarchical structure (Fig. 8.16c). They may also indicate true trichotomies, as discussed above; this can be checked by examination of the similarity or distance matrix.

A clustering method is said to be *monotonic* (i.e. without reversals) if:

$$S(\mathbf{x}_1 \cup \mathbf{x}_2, \mathbf{x}_3) \le S(\mathbf{x}_1, \mathbf{x}_2)$$

or

$$D(\mathbf{x}_1 \cup \mathbf{x}_2, \mathbf{x}_3) \ge D(\mathbf{x}_1, \mathbf{x}_2)$$

Assuming that $\alpha_h > 0$ and $\alpha_i > 0$ (Table 8.8), necessary and sufficient conditions for a clustering method to be monotonic in all situations are the following:

$$\alpha_h + \alpha_i + \beta \ge 1$$

and

$$\gamma \ge -\min(\alpha_h, \alpha_i)$$

(Milligan, 1979; Jain & Dubes, 1988). Some authors use the term *classification* only for hierarchies *without reversals* or for single non-overlapping partitions of the objects (Section 8.8).

8.7 Hierarchical divisive clustering

Contrary to the agglomerative methods of Section 8.5, hierarchical divisive techniques use the whole set of objects as the starting point. They divide it into two or several subgroups, after which they consider each subgroup and divide it again, until the criterion chosen to end the divisive procedure is met (Lance & Williams, 1967b).

In practice, hierarchical divisive clustering can only be achieved in the monothetic case, or when working in an ordination space. In monothetic divisive methods, the objects are divided, at each step of the procedure, according to the states of a single descriptor. This descriptor is chosen because it best represents the whole set of descriptors (next subsection). Polythetic algorithms have been developed, but it will be seen that they are not satisfactory.

An alternative is to use a partitioning method (Section 8.8) for all possible numbers of groups from $K = 2$ to $K = (n - 1)$ and assemble the results into a graph. There is no guarantee, however, that the groups will be nested and form a hierarchy, unless the biological or ecological processes that have generated the data are themselves hierarchical.

1 — Monothetic methods

Association analysis

The clustering methods that use only one descriptor at a time are less than ideal, even when the descriptor is chosen after considering all the others. The best-known monothetic method in ecology is Williams & Lambert's (1959) *association analysis*, originally described for species presence-absence data. Association analysis may actually be applied to any binary data table, not only species. The problem is to identify, at each step of the procedure, which descriptor is the most strongly associated with all the others. First, X^2 (chi-square) values are computed for 2 × 2 contingency tables comparing all pairs of descriptors in turn. X^2 is computed using the usual formula:

$$X^2 = n\,(ad - bc)^2/[(a + b)\,(c + d)\,(a + c)\,(b + d)]$$

The formula may include Yates' correction for small sample sizes, as in similarity coefficient S_{25}. The X^2 values relative to each descriptor k are summed up:

$$\sum_{j=1}^{p} X^2_{jk} \quad \text{for} \quad j \neq k \qquad \textbf{(8.16)}$$

of selecting large gaps in the data, sites that may be very close in species composition may become separated.

There are other problems with TWINSPAN when it comes to identifying clusters of species or computing indicator values.

- Firstly, when identifying clusters of species or computing indicator values, one cannot introduce some other classification of the sites in the program; only the classifications produced by TWINSPAN, which are based on correspondence analysis (CA) or detrended correspondence analysis (DCA, Subsection 9.4.5), may be used to delineate species groups.

- Secondly, the pseudospecies concept is based on species relative abundances. The relative abundance of a species depends on the absolute abundances of the other species present at the site. Such relative frequencies may be highly biased when sampling or harvesting mobile organisms; all species are not sampled with the same efficiency because of differences in behaviour. There is always a distortion between the estimated (i.e. observed) and real relative frequencies of species at any given site. A species abundance value observed at a site should only be compared to abundance values for the same species at other sites.

- Finally, whereas simple CA is well suited for studying species abundances observed at several sites (ter Braak 1985), DCA has recently been severely criticized (Subsection 9.4.5). Jackson and Somers (1991b) have shown that results obtained with DCA vary depending on the number of segments used to remove the arch effect. Therefore, several runs with different numbers of segments must be done to find stable factorial axes and interpretable results.

8.8 Partitioning by *K*-means

Partitioning consists in finding a single partition of a set of objects (Table 8.1). Jain & Dubes (1988) state the problem in the following terms: given *n* objects in a *p*–dimensional space, determine a partition of the objects into *K* groups, or clusters, such that the objects within each cluster are more similar to one another than to objects in the other clusters. The number of groups, *K*, is determined by the user. This problem has first been stated in statistical terms by Fisher (1958), who proposed solutions (with or without constraint of contiguity; see Sections 12.6 and 13.2) for a single variable.

The difficulty is to define what “more similar” means. Several criteria have been suggested; they can be divided into global and local criteria. A *global criterion* would be, for instance, to represent each cluster by a type-object (on *a priori* grounds, or using the centroids of Subsections 8.5.6 and 8.5.7) and assign each object to the nearest type-object. A *local criterion* uses the local structure of the data to delineate clusters; groups are formed by identifying high-density regions in the data. The

K–means method, described in the next paragraphs, is the most commonly used of the latter type.

Objective function

In *K*-means, the objective function that the partition to discover should minimize is the same as in Ward's agglomerative clustering method (Subsection 8.5.8): the total error sum of squares (E_K^2, or TESS). The major problem encountered by the algorithms is that the solution on which the computation eventually converges depends to some extent on the initial positions of the centroids. This problem does not exist in Ward's method, which proceeds iteratively by hierarchical agglomeration. However, even though Ward's algorithm guarantees that the *increase* in sum of squared errors ($\Delta E_{\mathbf{hi}}^2$, eq. 8.8) is minimized at each step of the agglomeration (so that any order of entry of the objects should lead to the same solution, except in cases of equal distances where a "right-hand" programming rule may prevail), there is no guarantee that any given Ward's partition is optimal in terms of the E_K^2 criterion — surprising at this may seem. This same problem occurs with all stepwise statistical methods.

Local minimum

The problem of the final solution depending on the initial positions of the centroids is known as the "local minimum" problem in algorithms. The concept is illustrated in Fig. 8.17, by reference to a *solution space*. It may be explained as follows. Solutions to the *K*-means problem are the different ways to partition *n* objects into, say, $K = 4$ groups. If a single object is moved from one group to another, the corresponding two solutions will have slightly different values for the criterion to be minimized (E_K^2). Imagine that all possible solutions form a "space of solutions". The different solutions can be plotted as a graph with the E_K^2 criterion as the ordinate. It is not necessary to accurately describe the abscissa to understand the concept; it would actually be a multidimensional space. A *K*-means algorithm starts at some position in this space, the initial position being assigned by the user (see below). It then tries to navigate the space to find the solution that minimizes the objective criterion (E_K^2). The space of solutions is not smooth, however. It may contain *local minima* from which the algorithm may be unable to escape. When this happens, the algorithm has not found the overall minimum and the partition is not optimal in terms of the objective criterion.

Overall minimum

Several solutions may be used to help a *K*-means algorithm converge towards the overall minimum of the objective criterion E_K^2. They involve either selecting specific objects as "group seeds" at the beginning of the run, or attributing the objects to the *K* groups in some special way. Here are some commonly-used approaches:

- Provide an initial configuration corresponding to an (ecological) hypothesis. The idea is to start the algorithm in a position in the solution space which is, hopefully, close to the final solution sought. This ideal situation is seldom encountered in real studies, however.

- Provide an initial configuration corresponding to the result of a hierarchical clustering, obtained from a space-conserving method (Table 8.8). One simply chooses the partition into *K* groups found on the dendrogram and lists the objects pertaining to

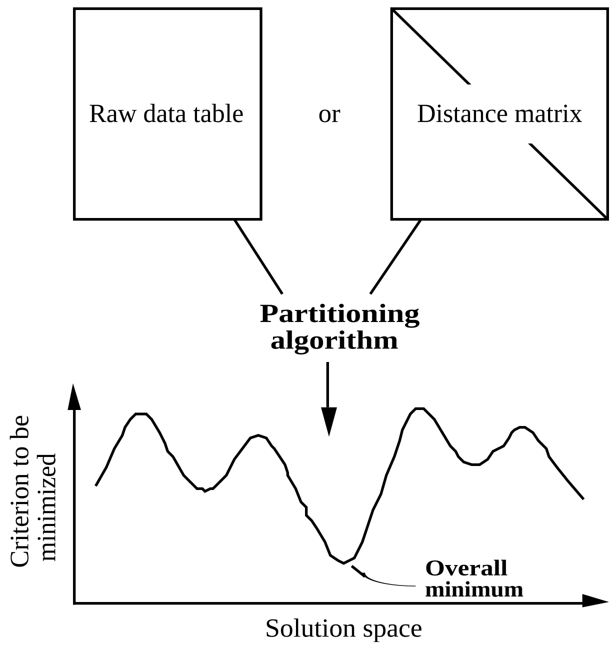


Figure 8.17 *K*-means algorithms search the space of solutions, trying to find the overall minimum (arrow) of the objective criterion to be minimized, while avoiding local minima (troughs).

each group. The *K*-means algorithm will then be asked to rearrange the group membership and look for a better overall solution (lower E_K^2 statistic).

- If the program allows it, select as "group seed", for each of the *K* groups to be delineated, some object located near the centroid of that group. For very large problems, Lance & Williams (1967d) suggested to use as starting point the result of a hierarchical clustering of a *random subset of the objects*, using as "group seeds" either the centroids of *K* clusters, or objects located near these centroids.

- Attribute the objects at random to the various groups. All *K*-means computer programs offer this option. Find a solution and note the E_K^2 value. It is possible, of course, that the solution found corresponds to a local minimum of E_K^2. So, repeat the whole procedure a number of times (for example, 100 times), starting every time from a different random configuration. Retain the solution that minimizes the E_K^2 statistic. One is more confident that this solution corresponds to the overall minimum when the corresponding value of E_K^2 is found several times across the runs.

Several algorithms have been proposed to solve the *K*–means problem, which is but one of a family of problems known in computer sciences as the *NP–complete* or

NP-hard *NP–hard problems*[*]. In all these problems, the only way to be sure that the optimal solution has been found is to try all possible solutions in turn. This is impossible, of course, for any real-size problems, even with modern-day computers, as explained in Subsection 8.7.2. Classical references to *K*-means algorithms are Anderberg (1973), Hartigan (1975), Späth (1975, 1980), Everitt (1980), and Jain & Dubes (1988). Milligan & Cooper (1987) have reviewed the most commonly used algorithms and compared them for structure recovery, using artificial data sets. One of the best algorithms available is the following; it frequently converges to the solution representing the overall minimum for the E_K^2 statistic. It is a very simple alternating least-squares algorithm, which iterates between two steps:

- Compute cluster centroids and use them as new cluster seeds.

- Assign each object to the nearest seed.

At the start of the program, *K* observations are selected as "group seeds". Each iteration reduces the sum of squared errors E_K^2, if possible. Since only a finite number of partitions are possible, the algorithm eventually reaches a partition from which no improvement is possible; iterations stop when E_K^2 can no longer be improved. The FASTCLUS procedure of the SAS package, mentioned here because it can handle very large numbers of objects, uses this algorithm. Options of the program can help deal with outliers if this is a concern. The SAS (1985) manual provides more information on the algorithm and available options.

This algorithm was originally proposed in a pioneering paper by MacQueen (1967) who gave the method its name: *K*-means. Lance & Williams made it popular by recommending it in their review paper (1967d). In the MacQueen paper, group centroids are recomputed after each addition of an object; this is also an option in SAS. MacQueen's algorithm contains procedures for the fusion of clusters, if centroids become very close, and for creating new clusters if an object is very distant from existing centroids.

K-means partitioning may be computed from either a table of raw data or a distance matrix, because the total error sum of squares E_K^2 is equal to the sum of squares of the distances from the points to their respective centroids (eq. 8.5; Fig. 8.18a) and to the sum (over groups) of the mean squared within-group distances (eq. 8.6; Fig. 8.18b). It is especially advantageous to compute it on raw data when the number of objects is large because, in such a situation, the distance matrix may become very cumbersome or even impossible to store and search. In contrast, when using a table of original data, one only needs to compute the distance of each object to each group centroid, rather than to all other objects.

[*] *NP* stands for *Non-deterministic Polynomial.* In theory, these problems can be solved in polynomial time (i.e. some polynomial function of the number of objects) on a (theoretical) non-deterministic computer. NP-hard problems are probably not solvable by efficient algorithms.

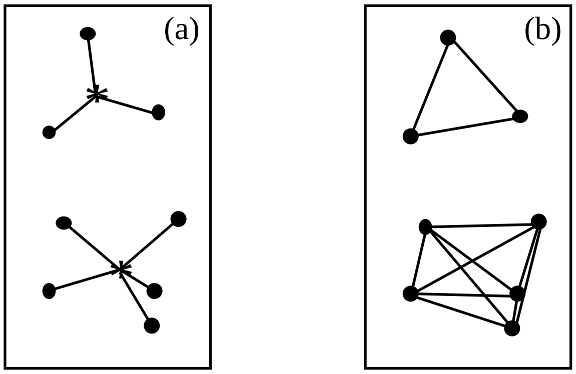
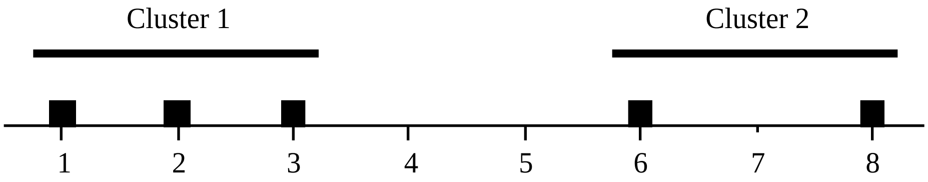


Figure 8.18 The total error sum of squares (E_K^2, TESS) is equal (a) to the sum of squares of the distances from the points to their respective centroids. (b) It is also equal to the sum (over groups) of the mean squared within-group distances.

The disadvantage of using a table of raw data is that the only distance function among points, available during *K*-means partitioning, is the Euclidean distance (D_1, Chapter 7) in A-space. This is not suitable with species counts and other ecological problems (Chapter 7). When the Euclidean distance is unsuitable, one may first compute a suitable similarity or distance measure among objects (Table 7.3 and 7.4); run the resemblance matrix through a metric or nonmetric scaling procedure (principal coordinate analysis, Section 9.2; nonmetric multidimensional scaling, Section 9.3); obtain a new table of coordinates for the objects in A-space; and run *K*-means partitioning using this table.

Following are two numerical examples that illustrate the behaviour of the E_K^2 criterion (eq. 8.5 and 8.6).

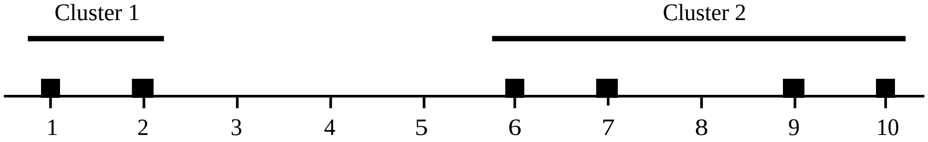
Numerical example 1. For simplicity, consider a single variable. The best partition of the following five objects (dark squares) in two clusters (boldface horizontal lines) is obviously to put objects with values 1, 2 and 3 in one group, and objects with values 6 and 8 in the other:



This example is meant to illustrate that the E_K^2 criterion can be computed from either raw data (eq. 8.5) or distances among objects (eq. 8.6). Using raw data (left-hand column, below), the group centroids are at positions 2 and 7 respectively; deviations from the centroids are calculated for each object, squared, and added within each cluster. Distances among objects (right-hand column, below) are easy to calculate from the object positions along the axis; the numbers of objects (n_k), used in the denominators, are 3 for cluster 1 and 2 for cluster 2.

$$\begin{array}{llllll}
e_1^2 = & (1^2 + 0^2 + (-1)^2) & = 2 & \qquad e_1^2 = & (2^2 + 1^2 + 1^2)/3 & = 2 \\
e_2^2 = & (1^2 + (-1)^2) & = 2 & \qquad e_2^2 = & 2^2/2 & = 2 \\
\cline{3-3}\cline{6-6}
 & E_K^2 & = 4 & & E_K^2 & = 4
\end{array}$$

Numerical example 2. Considering a single variable again, this example examines the effect on the E_K^2 statistic of changing the cluster membership. There are six objects and they are to be partitioned into $K = 2$ clusters. The optimal solution is that represented by the boldface horizontal lines:



Calculations are as above. Using raw data (left-hand column, below), the group centroids are at positions 0.5 and 8 respectively; deviations from the centroids are calculated for each object, squared, and added within each cluster. Distances among objects (right-hand column, below) are easy to calculate from the object positions along the axis; the numbers of objects (n_k), used in the denominators, are 2 for cluster 1 and 4 for cluster 2.

$$\begin{array}{llllll}
e_1^2 = & (0.5^2 + (-0.5)^2) & = \ 0.5 & \qquad e_1^2 = & 1^2/2 & = \ 0.5 \\
e_2^2 = & (2^2 + 1^2 + (-1)^2 + (-2)^2) & = 10.0 & \qquad e_2^2 = & (1^2 + 3^2 + 4^2 + 2^2 + 3^2 + 1^2)/4 & = 10.0 \\
\cline{3-3}\cline{6-6}
 & E_K^2 & = 10.5 & & E_K^2 & = 10.5
\end{array}$$

Consider now a sub-optimal solution where the clusters would contain the objects located at positions (1, 2, 6, 7) and (9, 10), respectively. The centroids are now at positions 4 and 9.5 respectively. Results are the following:

$$\begin{array}{llllll}
e_1^2 = & (3^2 + 2^2 + (-2)^2 + (-3)^2) & = 26.0 & \qquad e_1^2 = & (1^2 + 5^2 + 6^2 + 4^2 + 5^2 + 1^1)/4 & = 26.0 \\
e_2^2 = & (0.5^2 + (-0.5)^2) & = \ 0.5 & \qquad e_2^2 = & 1^2/2 & = \ 0.5 \\
\cline{3-3}\cline{6-6}
 & E_K^2 & = 26.5 & & E_K^2 & = 26.5
\end{array}$$

This example shows that the E_K^2 criterion quickly increases when the cluster membership departs from the optimum.

In some studies, the number of clusters K to be delineated is determined by the ecological problem, but this it is not often the case. The problem of determining the most appropriate number of clusters has been extensively discussed in the literature. Over 30 different methods, called “stopping rules”, have been proposed to do so. The efficiency coefficient, described in the last paragraph of Section 8.5, is one of them. Milligan & Cooper (1985; see also Milligan, 1996) compared them through an extensive series of simulations using artificial data sets with known numbers of clusters. Some of these rules recover the correct number of clusters in most instances, but others are appallingly inefficient. SAS, for instance, has implemented two among the best of these rules: a pseudo-*F* statistic and the cubic clustering criterion.

8.9 Species clustering: biological associations

Most of the methods discussed in the previous sections may be applied to clustering descriptors as well as objects. When searching for species associations, however, it is important to use clustering methods that model as precisely as possible a clearly formulated concept of association. The present section attempts (1) to define an operational concept of association and (2) to show how species associations can be identified in that framework.

Several concepts of species association have been developed since the nineteenth century; Whittaker (1962) wrote a remarkable review about them. These concepts are not always operational, however. In other words, they cannot always be translated into a series of well-defined analytical steps which would lead to the same result if they were applied by two independent researchers, using the same data. In general, the concept of association refers to a group of species that are “significantly” found together, without this implying necessarily any positive interaction among these species. In other words, an association is simply a group of species (or of taxa pertaining to some systematic category) recognized as a cluster following the application of a clearly stated set of rules.

Several procedures have been proposed for the identification of species associations. Quantitative algorithms have progressively replaced the empirical methods, as they have in other areas of science. All these methods, whether simple or elaborate, have two goals: first, identify the species that occur together and, second, minimize the likelihood that the co-occurrences so identified be fortuitous. The search for valid associations obviously implies that the sampling be random and planned in accordance with the pattern of variability under study (e.g. geographical, temporal, vertical, experimental). This pattern defines the framework within which the groups of species, found repeatedly along the sampling axes, are called associations; one then speaks of association of species over geographic space, or in time, etc. The criterion is the recurrence of a group of species along the sampling axes under study.

Ecologists are interested in associations of species as a conceptual framework to synthesize environmental characteristics. When associations have been found, one can concentrate on finding the ecological requirements common to most or all species of