4

Hierarchical clustering

4.1 Introduction

In a hierarchical classification the data are not partitioned into a particular number of classes or clusters at a single step. Instead the classification consists of a series of partitions, which may run from a single cluster containing all individuals, to n clusters each containing a single individual. Hierarchical clustering techniques may be subdivided into *agglomerative* methods, which proceed by a series of successive fusions of the n individuals into groups, and *divisive* methods, which separate the n individuals successively into finer groupings. Both types of hierarchical clustering can be viewed as attempting to find the optimal step, in some defined sense (see later), at each stage in the progressive subdivision or synthesis of the data, and each operates on a proximity matrix of some kind (see Chapter 3). A useful review of the standard methods has been given by Gordon (1987).

With hierarchical methods, divisions or fusions, once made, are irrevocable so that when an agglomerative algorithm has joined two individuals they cannot subsequently be separated, and when a divisive algorithm has made a split it cannot be undone. As Kaufman and Rousseeuw (1990) colourfully comment: 'A hierarchical method suffers from the defect that it can never repair what was done in previous steps'. Hawkins *et al.* (1982) illustrate the problem in the following way. Suppose a single variable is measured on eight objects, giving the results (−2.2, −2, −1.8, −0.1, 0.1, 1.8, 2, 2.2). The data contain three obvious 'clusters'. If the first split was into two clusters on the basis of the size of the usual t-statistic, the middle cluster (−0.1, 0.1) would be divided to produce the two clusters (−2.2, −2, −1.8, −0.1) and (0.1, 1.8, 2, 2.2). To recover them would

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

entail the nonhierarchical approach of continuing to a four-cluster solution and then merging these two.

Since all agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, and the divisive techniques will finally split the entire set of data into n groups each containing a single individual, the investigator wishing to have a solution with an 'optimal' number of clusters will need to decide when to stop. The tricky problem of deciding on the correct number of clusters is discussed in Section 4.4.4.

Hierarchical classifications produced by either the agglomerative or divisive route may be represented by a two-dimensional diagram known as a *dendrogram,* which illustrates the fusions or divisions made at each stage of the analysis. An example of such a diagram is given in Figure 4.1. Some further properties of dendrograms and how they may be used in interpreting the results of hierarchical clustering techniques are discussed in Section 4.4.

The structure in Figure 4.1 resembles an evolutionary tree, and it is in biological applications that hierarchical classifications are perhaps most relevant. According to Rohlf (1970), a biologist, 'all other things being equal', aims for a system of nested clusters. Other areas where hierarchical classifications might be particularly appropriate are studies of social systems, and in museology and

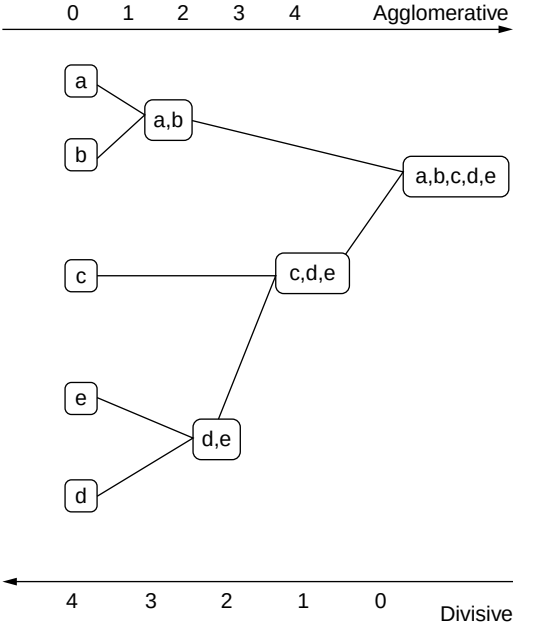


Figure 4.1 *Example of a hierarchical tree structure. (* Taken from *Finding Groups in Data,* 1990, Kaufman and Rousseeuw, with permission of the publisher, John Wiley & Sons, Inc.).

librarianship, where hierarchies are implicit in the subject matter. As will be seen later, hierarchical clustering methods have been applied in many other areas where there is not necessarily an underlying hierarchical structure. Although they may still often be usefully applied in these areas, if only to provide a starting point for a more complex clustering procedure, the following caveat of Hawkins *et al.* (1982) should be borne in mind: 'users should be very wary of using hierarchic methods if they are not clearly necessary'.

The following sections describe commonly used agglomerative techniques, and their properties. These properties are potentially applicable to divisive techniques also, but they are discussed here in relation to agglomerative techniques, since this is where most technical research has been concentrated. A description of some divisive techniques in Section 4.3 will be followed by a discussion of issues common to both agglomerative and divisive techniques. Several applications of hierarchical clustering techniques will be described in Section 4.5.

4.2 Agglomerative methods

Agglomerative procedures are probably the most widely used of the hierarchical methods. They produce a series of partitions of the data: the first consists of n single-member 'clusters'; the last consists of a single group containing all n individuals. The basic operation of all such methods is similar, and will be illustrated for two specific examples, *single linkage* and *centroid linkage.* At each stage the methods fuse individuals or groups of individuals which are closest (or most similar). Differences between the methods arise because of the different ways of defining distance (or similarity) between an individual and a group containing several individuals, or between two groups of individuals (see Chapter 3 for further details). Before giving a summary of the most widely used methods, we illustrate the general approach using two examples.

4.2.1 Illustrative examples of agglomerative methods

In this section, two hierarchical techniques are illustrated, the first requiring solely a proximity matrix, the second requiring access to a data matrix. The first illustration is of one of the simplest hierarchical clustering methods, *single linkage,* also known as the nearest-neighbour technique. It was first described by Florek *et al.* (1951) and later by Sneath (1957) and Johnson (1967). The defining feature of the method is that the distance between groups is defined as that of the closest pair of individuals, where only pairs consisting of one individual from each group are considered (nearest-neighbour distance; see Section 3.6). Single linkage serves to illustrate the general procedure of a hierarchical method, and in the example below it is applied as an agglomerative method. However, it could equally well be applied as a divisive method, by starting with a cluster containing all objects and then splitting into two clusters whose nearest-neighbour distance is a maximum.

Consider the following distance matrix:

$$\mathbf{D}_1 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}.$$

The smallest nonzero entry in the matrix is that for individuals 1 and 2, so these are joined to form a two-member cluster. Distances between this cluster and the other three individuals are obtained as

$$\begin{aligned} d_{(12)3} &= \min(d_{13}, d_{23}) = d_{23} = 5.0 \\ d_{(12)4} &= \min(d_{14}, d_{24}) = d_{24} = 9.0 \\ d_{(12)5} &= \min(d_{15}, d_{25}) = d_{25} = 8.0. \end{aligned}$$

A new matrix may now be constructed whose entries are inter-individual and cluster-individual distance values:

$$\mathbf{D}_2 = \begin{matrix} (12) \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0.0 & & & \\ 5.0 & 0.0 & & \\ 9.0 & 4.0 & 0.0 & \\ 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}.$$

The smallest entry in $\mathbf{D}_2$ is that for individuals 4 and 5, so these now form a second two-member cluster and a new set of distances are found:

$$\begin{aligned} d_{(12)3} &= 5.0 \text{ as before} \\ d_{(12)(45)} &= \min(d_{14}, d_{15}, d_{24}, d_{25}) = d_{25} = 8.0 \\ d_{(45)3} &= \min(d_{34}, d_{35}) = d_{34} = 4.0. \end{aligned}$$

These may be arranged in a matrix $\mathbf{D}_3$:

$$\mathbf{D}_3 = \begin{matrix} (12) \\ 3 \\ (45) \end{matrix} \begin{pmatrix} 0.0 & & \\ 5.0 & 0.0 & \\ 8.0 & 4.0 & 0.0 \end{pmatrix}.$$

The smallest entry is now $d_{(45)3}$, and individual 3 is added to the cluster containing individuals 4 and 5. Finally the groups containing individuals 1, 2 and 3, 4, 5 are combined into a single cluster.

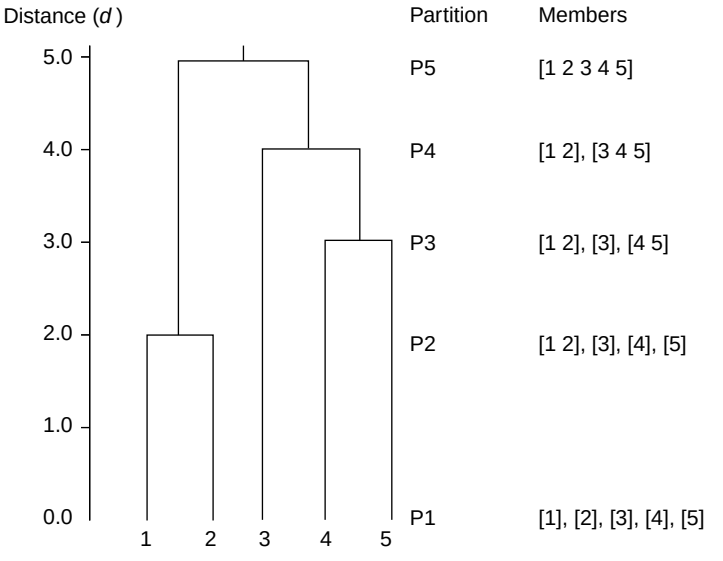


Figure 4.2 *Dendrogram for worked example of single linkage, showing partitions at each step.*

The dendrogram illustrating the process, and the partitions produced at each stage are shown in Figure 4.2; the *height* in this diagram represents the distance at which each fusion is made. Dendrograms and their features are described in more detail in Section 4.4.1.

Single linkage operates directly on a proximity matrix. Another type of clustering, *centroid* clustering, however, requires access to the original data. To illustrate this type of method, it will be applied to the following set of bivariate data:

Object	Variable 1	Variable 2
1	1.0	1.0
2	1.0	2.0
3	6.0	3.0
4	8.0	2.0
5	8.0	0.0

Suppose Euclidean distance is chosen as the inter-object distance measure, giving the following distance matrix:

$$\mathbf{C}_1 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0.00 & & & & \\ 1.00 & 0.00 & & & \\ 5.39 & 5.10 & 0.00 & & \\ 7.07 & 7.00 & 2.24 & 0.00 & \\ 7.07 & 7.28 & 3.61 & 2.00 & 0.00 \end{pmatrix}.$$

Examination of $\mathbf{C}_1$ shows that c_{12} is the smallest entry, and objects 1 and 2 are fused to form a group. The mean vector (centroid) of the group is calculated (1, 1.5) and a new Euclidean distance matrix is calculated:

$$\mathbf{C}_2 = \begin{matrix} (12) \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0.00 & & & \\ 5.22 & 0.00 & & \\ 7.02 & 2.24 & 0.00 & \\ 7.16 & 3.61 & 2.00 & 0.00 \end{pmatrix}.$$

The smallest entry in $\mathbf{C}_2$ is c_{45}, and objects 4 and 5 are therefore fused to form a second group, the mean vector of which is (8.0, 1.0). A further distance matrix $\mathbf{C}_3$ is now calculated:

$$\mathbf{C}_3 = \begin{matrix} (12) \\ 3 \\ (45) \end{matrix} \begin{pmatrix} 0.00 & & \\ 5.22 & 0.00 & \\ 7.02 & 2.83 & 0.00 \end{pmatrix}.$$

In $\mathbf{C}_3$ the smallest entry is $c_{(45)3}$, and so objects 3, 4 and 5 are merged into a three-member cluster. The final stage consists of the fusion of the two remaining groups into one.

An important point to note about the two methods mentioned above (and all the methods discussed in this chapter) is that the clusterings proceed hierarchically, each being obtained by the merger of clusters from the previous level. So, for example, in neither of the examples above could clusters (1, 2, 4) and (3, 5) have been formed, since neither is obtainable by merging existing clusters.

4.2.2 The standard agglomerative methods

In addition to those introduced in the previous section, there are several other possible inter-group proximity measures (see Section 3.6), each giving rise to a different agglomerative method. For example, *complete linkage* (or furthest neighbour) is opposite to single linkage, in the sense that distance between groups is now defined as that of the most distant pair of individuals. In *group average linkage* – also known as the unweighted pair-group method using the average approach (UPGMA) – the distance between two clusters is the average of the distance between all pairs of individuals that are made up of one individual from each group. All these three methods (single, complete and average) use a proximity matrix as input, and the inter-cluster distances they use are each illustrated graphically in Figure 4.3.

Another agglomerative hierarchical method is *centroid clustering* – also known as the unweighted pair-group method using the centroid approach (UPGMC) – which uses a data matrix rather than a proximity matrix and involves merging clusters with the most similar mean vectors. *Median linkage* – the weighted

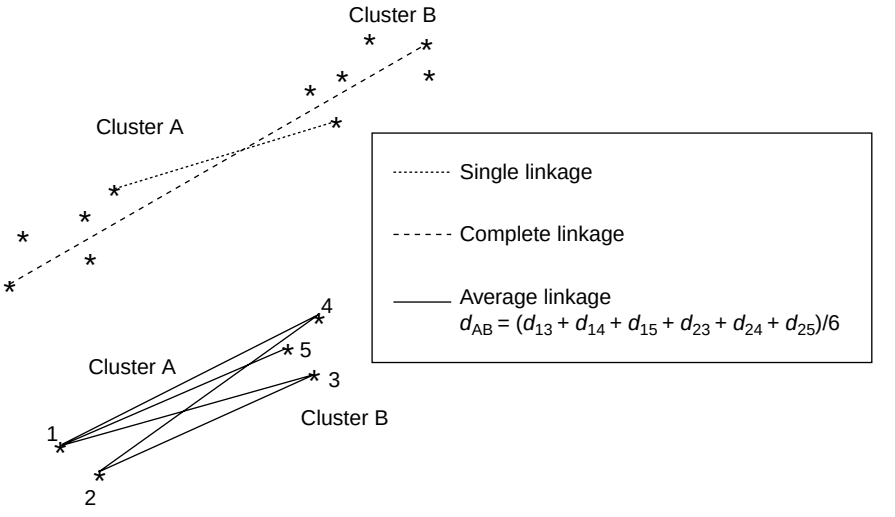


Figure 4.3 *Examples of three inter-cluster distance measures: single, complete and average.*

pair-group method using the centroid approach (WPGMC) – is similar, except that the centroids of the constituent clusters are weighted equally to produce the new centroid of the merged cluster. This is to avoid the objects in the more numerous of the pair of clusters to be merged dominating those in the smaller cluster. The new centroid is thus intermediate between the two constituent clusters.

In the numerical illustration of centroid linkage shown in Section 4.2.1, Euclidean distance was used, as is most common. While other proximity measures are possible with centroid or median linkage, they would lack interpretation in terms of the raw data (see Anderberg, 1973).

Ward (1963) introduced a third type of method, in which the fusion of two clusters is based on the size of an error sum-of-squares criterion. The objective at each stage is to minimize the increase in the total within-cluster error sum of squares, E, given by

$$E = \sum_{m=1}^{g} E_m,$$

where

$$E_m = \sum_{l=1}^{n_m} \sum_{k=1}^{p_k} \left(x_{ml,k} - \bar{x}_{m,k} \right)^2, \tag{4.1}$$

in which $\bar{x}_{m,k} = (1/n_m) \sum_{l=1}^{n_m} x_{ml,k}$ (the mean of the mth cluster for the kth variable), $x_{ml,k}$ being the score on the kth variable ($k = 1, \ldots, p$) for the lth object

($l = 1, \ldots, n_m$) in the mth cluster ($m = 1, \ldots, g$). This increase is proportional to the squared Euclidean distance between the centroids of the merged clusters, but the method differs from centroid clustering in that centroids are weighted by $n_m n_q/(n_m + n_q)$ when computing distances between centroids, where n_m and n_q are the numbers of objects in the two clusters m and q.

Weighted average linkage (McQuitty, 1966), also known as WPGMA, is similar to (group) average linkage but weights inter-cluster distances according to the inverse of the number of objects in each class, as in the case of median compared to centroid linkage.

The seven methods introduced so far are summarized in Table 4.1, along with some remarks about some of their typical characteristics, which will be amplified below.

Some other hierarchical methods, related to the techniques described above, should also be mentioned. The *sum-of-squares* method (Jambu, 1978; Podani, 1989) is similar to Ward's method but is based on the sum of squares within each cluster rather than the increase in sum of squares in the merged cluster.

Lance and Williams (1967) also introduced a new *flexible* method defined by values of the parameters of a general recurrence formula, outlined in the next subsection. Many of the mathematical properties of the standard hierarchical methods can be defined in terms of the parameters of the Lance and Williams formulation, and in Section 4.4.3 some of these are introduced.

4.2.3 Recurrence formula for agglomerative methods

The Lance and Williams recurrence formula gives the distance between a group k and a group (ij) formed by the fusion of two groups (i and j) as

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}|, \tag{4.2}$$

where d_{ij} is the distance between groups i and j. Lance and Williams used the formula to define a new 'flexible' scheme, with parameter values $\alpha_i + \alpha_j + \beta = 1$, $\alpha_i = \alpha_j$, $\beta < 1$, $\gamma = 0$. By allowing β to vary, clustering schemes with various characteristics can be obtained. They suggest small negative values for β, such as -0.25, although Scheibler and Schneider (1985) suggest -0.50.

The inter-group distance measures used by many standard hierarchical clustering techniques can, by suitable choice of the parameters α_i, α_j, β and γ, be contained within this formula, as shown in Table 4.2, which also shows additional properties of these methods, to be discussed in Section 4.4.3. Single linkage, for example, corresponds to the parameter values $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = 0$ and $\gamma = -\frac{1}{2}$, and (4.2) is

$$d_{k(ij)} = \frac{1}{2} d_{ki} + \frac{1}{2} d_{ki} d_{kj} - \frac{1}{2} \left| d_{ki} - d_{kj} \right| \tag{4.3}$$

If $d_{ki} > d_{kj}$, then $|d_{ki} - d_{kj}| = d_{ki} - d_{kj}$ and $d_{k(ij)} = d_{kj}$. Similarly, if $d_{ki} < d_{kj}$, $|d_{ki} - d_{kj}| = d_{kj} - d_{ki}$, and consequently the recurrence formula gives the

Table 4.1 Standard agglomerative hierarchical clustering methods.

Method	Alternative name[a]	Usually used with:	Distance between clusters defined as:	Remarks
Single linkage Sneath (1957)	Nearest neighbour	Similarity or distance	Minimum distance between pair of objects, one in one cluster, one in the other	Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure.
Complete linkage Sorensen (1948)	Furthest neighbour	Similarity or distance	Maximum distance between pair of objects, one in one cluster, one in the other	Tends to find compact clusters with equal diameters (maximum distance between objects). Does not take account of cluster structure.
(Group) Average linkage Sokal and Michener (1958)	UPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	Tends to join clusters with small variances. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust.
Centroid linkage Sokal and Michener (1958)	UPGMC	Distance (requires raw data)	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged cluster. Subject to reversals.
Weighted average linkage McQuitty (1966)	WPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	As for UPGMA, but points in small clusters weighted more highly than points in large clusters (useful if cluster sizes are likely to be uneven).
Median linkage Gower (1967)	WPGMC	Distance (requires raw data)	Squared Euclidean distance between weighted centroids	Assumes points can be represented in Euclidean space for geometrical interpretation. New group is intermediate in position between merged groups. Subject to reversals.
Ward's method Ward (1963)	Minimum sum of squares	Distance (requires raw data)	Increase in sum of squares within clusters, after fusion, summed over all variables	Assumes points can be represented in Euclidean space for geometrical interpretation. Tends to find same-size, spherical clusters. Sensitive to outliers.

[a]U = unweighted; W = weighted; PG = pair group; A = average; C = centroid.

Table 4.2 Hierarchical agglomerative clustering methods: admissibility conditions and Lance–Williams parameters.

	Admissibility conditions[a]				Lance–Williams parameters[b]		
Method	U	C	P	M	α_i	β	γ
Single linkage	N	N	Y	Y	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete linkage	N	N	Y	Y	$\frac{1}{2}$	0	$\frac{1}{2}$
Average linkage	N	N	N	N	$n_i/(n_i + n_j)$	0	0
Centroid linkage	Y	N	N	N	$n_i/(n_i + n_j)$	$-n_i n_j/(n_i + n_j)^2$	0
Median linkage	Y	N	Y	N	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward's method	N	Y	N	N	$(n_k + n_i)/(n_k + n_i + n_j)$	$-n_k/(n_k + n_i + n_j)$	0

[a] U = no reversals; C = convex; P = point proportional; M = monotone.
[b] n_k, n_i and n_j are the respective cluster sizes when cluster k is joined to the fusion of clusters i and j (see Equation (4.2)).

required

$$d_{k(ij)} = \min(d_{ki}, d_{kj}) \tag{4.4}$$

The Lance and Williams recursive formula can be used to program many hierarchical agglomerative methods, and such algorithms use computer time of the order of $n^2\log(n)$. Improvements can be made on this, as discussed by Hansen and Jaumard (1997). For example, algorithms based on merging edges in the minimum spanning tree representation of single linkage are proportional to n^2. (Divisive algorithms are intrinsically more difficult to program efficiently, and this is partly why they are less widely used.)

4.2.4 Problems of agglomerative hierarchical methods

To illustrate some of the potential problems of these agglomerative methods, a set of simulated data will be clustered using single, complete and average linkage. The data consist of 50 points simulated from two bivariate normal distributions with mean vectors (0, 0) and (4, 4), and common covariance matrix

$$\Sigma = \begin{pmatrix} 16.0 & 1.5 \\ 1.5 & 0.25 \end{pmatrix}.$$

Two intermediate points have been added for the first analysis, in order to illustrate a problem known as *chaining* often found when using single linkage. Figure 4.4 gives the single linkage dendrogram and Figure 4.5 shows some of the results of the cluster analyses.

Figure 4.4 shows a typical single linkage dendrogram. There is little clear structure, with the two intermediate points (51 and 52) linking the two main clusters, which are gradually pulled together into one large cluster, isolating two singletons until the final step. Note that although the outlying points 8 and 29 are

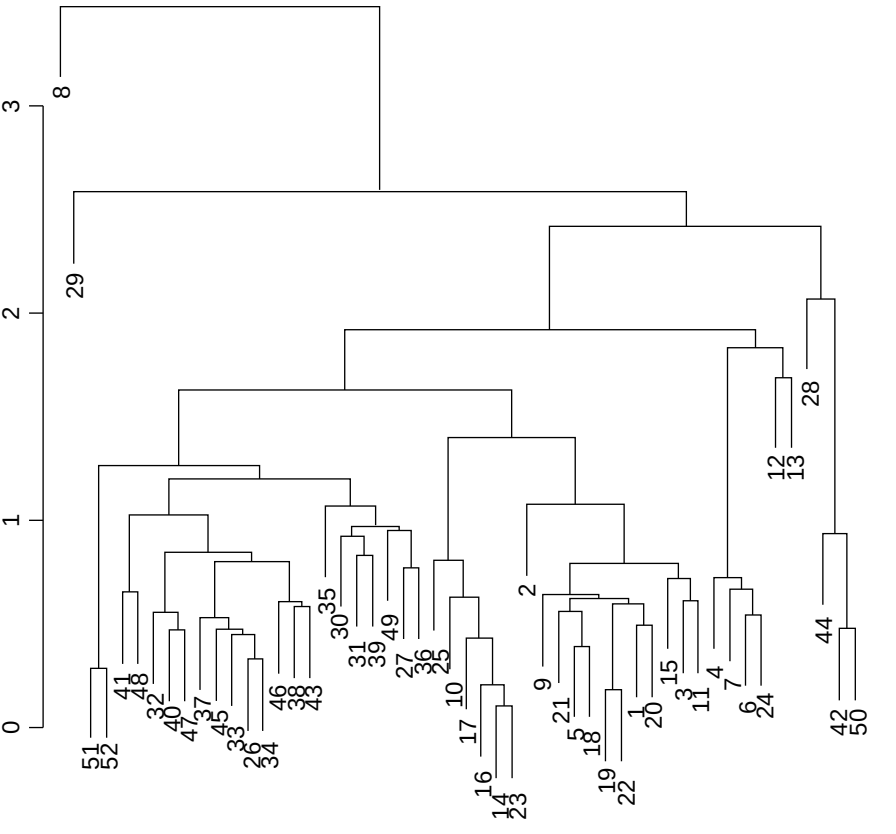


Figure 4.4 *Dendrogram showing single linkage clustering of simulated data set (see also Figure 4.5(a)).*

close together on the dendrogram, they are those at the extreme opposite ends of the main clusters. Note also that this form of dendrogram places the labels for the points just underneath the place where they first join a cluster. This makes the order of joining evident. Some other software would place all labels along the zero line at the bottom.

Figure 4.5(a) shows the chaining of the two main groups together in single linkage, and the isolation of one outlier, if two groups are specified. (If three groups are specified, the other outlier is hived off, still leaving one large cluster.) Despite the obvious lack of success in recovering the two groups, this example does illustrate a potential benefit of applying single linkage, namely that it can be used to identify outliers, since these are left as singletons if they are sufficiently far from their nearest neighbour.

Complete (Figure 4.5(c)) and average linkage (Figure 4.5(d)) techniques were equally unsuccessful in cluster recovery, with or without intermediate points, and whatever number of clusters was specified (from two to five). They tended to

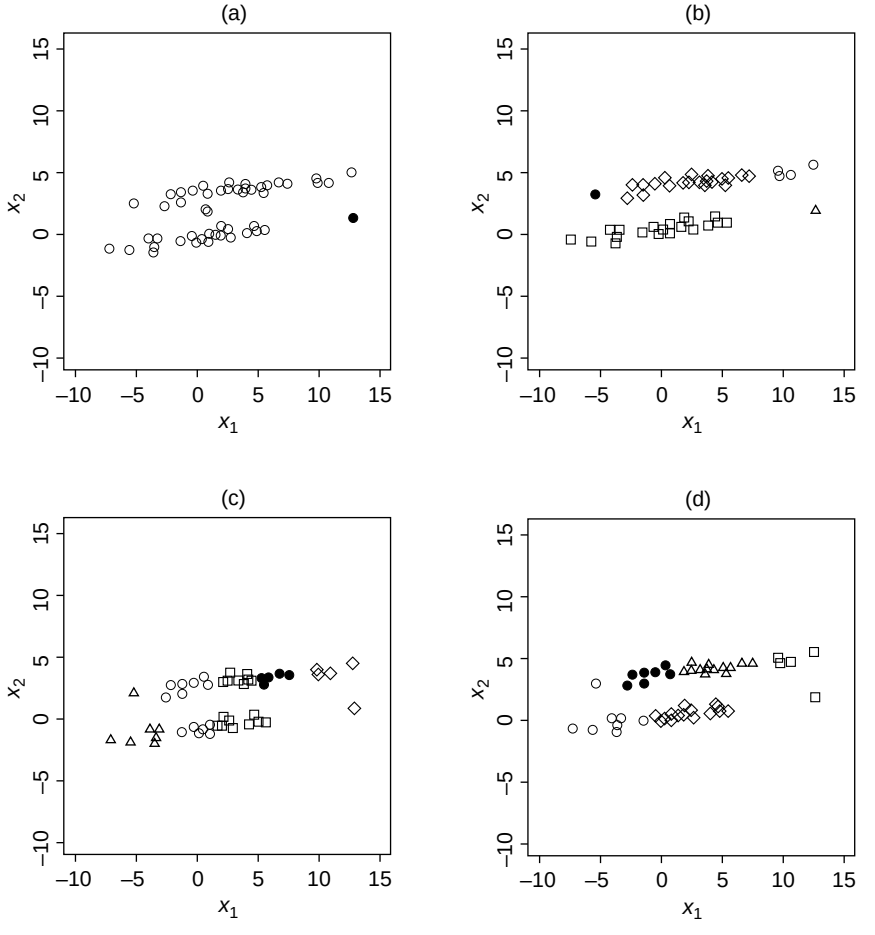


Figure 4.5 *Clusters obtained by four different methods from simulated data: (a) single linkage, with intermediate points, two-cluster solution; (b) single linkage, no intermediate points, five-cluster solution; (c) complete linkage, no intermediate points, five-cluster solution; (d) average linkage, no intermediate points, five-cluster solution.*

impose spherical clusters, forming a cluster in the middle, part from group 1 and part from group 2. The five-cluster solution for single linkage (Figure 4.5(b)) was relatively more successful, since the five clusters could be amalgamated into two, to form the correct groups. (Of course, such an amalgamation would destroy the hierarchy, just as in the Hawkins *et al.* example given in Section 4.1.)

These small examples show some of the problems of agglomerative methods and their relative failure to recover nonspherical clusters. (Similar problems were found in the more general empirical studies to be discussed in the next subsection.) It also shows how crucial it is to make the correct choice as to the number of clusters present (see Section 4.4.4), and the advisability of plotting the

raw data where feasible. A practical computational consideration is non-uniqueness due to ties (for single linkage and some other methods). In the case of non-uniqueness, a decision as to which clusters to fuse needs to be made, and this is usually a default choice determined by software. It is generally recommended to run analyses with different choices to check for robustness. The more sophisticated model-based clustering techniques to be discussed in Chapters 6 and 7 have the potential to overcome some of these problems.

4.2.5 Empirical studies of hierarchical agglomerative methods

Empirical studies of hierarchical methods are of two main types. One type simulates clusters in data of a particular type and then assesses the characteristics and recovery of clusters. The other is based on real data from a particular subject matter, the criterion in the latter usually being the interpretability of clusters. Examples of the former include a review by Milligan (1981) and a study reported by Hands and Everitt (1987). The latter concluded that Ward's method performed very well when the data contained clusters with approximately the same numbers of points, but poorly when the clusters were of different sizes. In that situation, centroid clustering appeared to give the most satisfactory results. Cunningham and Ogilvie (1972) and Blashfield (1976) also concluded that for clusters with equal numbers of points Ward's method is successful, otherwise centroid group average and complete linkage are preferable.

Studies that focus on the stability of clustering in the presence of outliers or noise include that by Hubert (1974), who found that complete linkage is less sensitive to observational errors than single linkage. (A related point is the observation of Hartigan (1975), that single linkage is dependent on the smallest distances, and they need to be measured with low error for single linkage to be successful.)

An empirical study based on the subject-matter approach is that of Duflou and Maenhaut (1990). These authors compared seven standard methods (those in Table 4.1 and one other) on data involving chemical concentrations in the brain. They rejected centroid and median linkage because of reversals (a type of inconsistency in the hierarchy; see Section 4.4.3), and concluded that, of the remainder, Ward's method and complete linkage gave interpretable results and correctly distinguished grey and white matter areas in the brain. A further example is provided by Baxter (1994), who summarizes the position in archaeology, where empirical studies generally favour Ward's method and average linkage.

It has to be recognized that hierarchical clustering methods may give very different results on the same data, and empirical studies are rarely conclusive. What is most clear is that no one method can be recommended above all others and, as Gordon (1998) points out, hierarchical methods are in any case only stepwise optimal. A few general observations can, however, be made. Single linkage, which has satisfactory mathematical properties and is also easy to program and apply to large data sets, tends to be less satisfactory than other methods because of

'chaining'; this is the phenomenon in which separated clusters with 'noise' points in between them tend to be joined together. Ward's method often appears to work well but may impose a spherical structure where none exists.

4.3 Divisive methods

Divisive methods operate in the opposite direction to agglomerative methods, starting with one large cluster and successively splitting clusters. They are computationally demanding if all $2^{k-1} - 1$ possible divisions into two subclusters of a cluster of k objects are considered at each stage. However, for data consisting of p binary variables, relatively simple and computationally efficient methods, known as *monothetic divisive methods*, are available. These generally divide clusters according to the presence or absence of each of the p variables, so that at each stage clusters contain members with certain attributes either all present or all absent. The data for these methods thus need to be in the form of a two-mode (binary) matrix. The term 'monothetic' refers to the use of a single variable on which to base the split at a given stage; *polythetic* methods, to be described in Section 4.3.2, use all the variables at each stage. While less commonly used than agglomerative methods, divisive methods have the advantage, pointed out by Kaufman and Rousseeuw (1990), that most users are interested in the main structure in their data, and this is revealed from the outset of a divisive method.

4.3.1 Monothetic divisive methods

The choice of the variable in monothetic divisive methods on which a split is made depends on optimizing a criterion reflecting either cluster homogeneity or association with other variables. This tends to minimize the number of splits that have to be made. An example of the homogeneity criterion is the *information content, C* (which in this case signifies disorder or chaos), defined by p variables and n objects (Lance and Williams, 1968):

$$C = pn\log n - \sum_{k=1}^{p} [f_k \log f_k - (n - f_k)\log(n - f_k)], \tag{4.5}$$

where f_k is the number of individuals having the kth attribute. If a group X is to be split into two groups A and B, the reduction in C is $C_X - C_A - C_B$. The ideal set of clusters would have members with identical attributes and C equal to zero; hence clusters are split at each stage according to possession of the attribute which leads to the greatest reduction in C.

Instead of cluster homogeneity, the attribute used at each step can be chosen according to its overall association with all attributes remaining at this step: this is sometimes termed *association analysis* (Williams and Lambert, 1959), especially in ecology. For example, for one pair of variables V_i and V_j with values 0 and 1, the frequencies observed might be:

V_j	V_i	
	1	0
1	a	b
0	c	d

Common measures of association (summed over all pairs of variables) are the following:

$$|ad-bc| \tag{4.6}$$

$$(ad-bc)^2 \tag{4.7}$$

$$(ad-bc)^2 n/[(a+b)(a+c)(b+d)(c+d)] \tag{4.8}$$

$$\sqrt{(ad-bc)^2 n/[(a+b)(a+c)(b+d)(c+d)]} \tag{4.9}$$

$$(ad-bc)^2/[(a+b)(a+c)(b+d)(c+d)] \tag{4.10}$$

The split at each stage is made according to the presence or absence of the attribute whose association with the others (i.e. the summed criterion above) is a maximum. The first two criteria, (4.6) and (4.7), have the advantage that there is no danger of computational problems if any of the marginal totals are zero (Kaufman and Rousseeuw, 1990). The last three, (4.8), (4.9) and (4.10), are all related to the usual chi-squared statistic, its square root, and the Pearson correlation coefficient, respectively. Hubálek (1982) gives a review of 43 such coefficients.

Appealing features of monothetic divisive methods are the easy classification of new members, and the inclusion of cases with missing values. The latter can be dealt with as follows. If there are missing values for a particular variable, V_1 say, the nonmissing variable with the largest absolute association with it is determined, V_2, say. The missing value for V_1 is replaced by the value of V_2 for the same observation (positive association between V_1 and V_2) or $1 - V_2$ (negative association).

A further advantage of monothetic divisive methods is that it is obvious which variables produce the split at any stage of the process. However, a general problem with these methods is that the possession of a particular attribute, which is either rare or rarely found in combination with others, may take an individual down the 'wrong' path. Typical uses of the method are in medicine (as diagnostic keys; see, for example, Payne and Preece, 1980) and in mortuary studies in archaeology, where it can be argued that social stratum in life might be reflected by the possession of a common set of grave goods (see O'Shea, 1985, for example).

A new method of divisive clustering has been proposed by Piccarreta and Billari (2007), which can be used for sequence data such as life-course histories. The method uses the logic of classification and regression tree (CART) analysis

(Breiman et al., 1984), thus enabling some of the most useful features of CART analysis to be employed, such as tree pruning by cross-validation to identify the appropriate number of clusters. Piccarreta and Billari define two new types of data derived from the original sequences: *auxiliary variables* and *state permanence sequences*. This means that, rather than having completely different dependent and independent variables (as in CART), the variables defining the splits, the criterion for assessing the homogeneity of the clusters, and the data characterizing the clusters are all derived from the sequence data.

The splits in this new method are made with the aim of producing 'pure' clusters. However, whereas, in CART, purity is defined in terms of a dependent or outcome variable, here 'impurity' is defined as the summed OMA distance between all pairs of units (see Chapter 3, Section 3.5 for a description of the OMA distance measure). The auxiliary variables, which are used to split the sample, can be defined in various ways according to the subject matter; for example, they might be the times at which a particular state is reached for the first time, second time, etc. The divisive procedure operates by choosing the auxiliary variables that lead to the greatest improvement in the within-cluster purity at each stage of the splitting process. This leads to a tree with nodes (clusters) which can then be simplified by pruning – cutting back branches. As in CART, this is at the expense of within-cluster purity, which has to be balanced against increased simplicity. Once a satisfactory solution has been found, the exemplar (see Section 4.4.1) – here the medoid of each cluster – can be used to summarize the clusters using a representation that retains some key features of the original sequence: the *state permanence sequence*, which indicates the length of time in each state. Section 4.5.4 describes the authors' application of this method. `SAS` routines are available from the authors of the paper.

4.3.2 Polythetic divisive methods

Polythetic divisive methods are more akin to the agglomerative methods discussed above, since they use all variables simultaneously, and can work with a proximity matrix. The procedure of MacNaughton-Smith *et al.* (1964) avoids considering all possible splits, a potential problem of polythetic divisive methods. It proceeds by finding the object that is furthest away from the others within a group, and using that as the seed for a splinter group. Each object is then considered for entry to the splinter group: any that are closer to the splinter group are moved into it. The step is repeated, the next cluster for splitting being chosen as the largest in diameter (defined by the largest dissimilarity between any two objects).

The process has been described as follows by Kaufman and Rousseeuw (1990), who have developed a program, `diana` (DIvisive ANAlysis clustering), which is implemented in `S-plus` and `R`, and this is often used as the method's appellation.

The mechanism somewhat resembles the way a political party might split up due to inner conflicts: firstly the most discontented member leaves the party and starts a new one, and then some others follow him until a kind of equilibrium is attained. So we first need to know which member disagrees most with the others.

To illustrate this, consider the following distance matrix for seven individuals:

$$\mathbf{D} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{pmatrix} 0 & & & & & & \\ 10 & 0 & & & & & \\ 7 & 7 & 0 & & & & \\ 30 & 23 & 21 & 0 & & & \\ 29 & 25 & 22 & 7 & 0 & & \\ 38 & 34 & 31 & 10 & 11 & 0 & \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{pmatrix}.$$

The individual used to initiate the splinter group is the one whose average distance from the other individuals is a maximum. This is found to be individual 1, giving the initial groups as (1) and (2, 3, 4, 5, 6, 7). Next the average distance of each individual in the main group to the individuals in the splinter group is found, followed by the average distances of each individual in the main group to the other individuals in this group. The difference between these two averages is then found. In this example this leads to:

Individual in main group	Average distance to splinter group (A)	Average distance to main group (B)	$B-A$
2	10.0	25.0	15.0
3	7.0	23.4	16.4
4	30.0	14.8	−15.2
5	29.0	16.4	−12.6
6	38.0	19.0	−19.0
7	42.0	22.2	−19.8

The maximum difference is 16.4, for individual 3, which is therefore added into the splinter group, giving the two groups (1, 3) and (2, 4, 5, 6, 7). Repeating the process gives the following:

Individual in main group	Average distance to splinter group (A)	Average distance to main group (B)	$B-A$
2	8.5	29.5	21.0
4	25.5	13.2	−12.3
5	25.5	15.0	−10.5
6	34.5	16.0	−18.5
7	39.0	18.7	−20.3

So now individual 2 joins the splinter group to give groups (1, 3, 2) and (4, 5, 6, 7), and the process is repeated to give:

Individual in main group	Average distance to splinter group (A)	Average distance to main group (B)	$B - A$
4	24.3	10.0	−14.3
5	25.3	11.7	−13.6
6	34.3	10.0	−24.3
7	38.0	13.0	−25.0

As all the differences are now negative, the process would continue (if desired) on each subgroup separately.

4.4 Applying the hierarchical clustering process

To make best use of hierarchical techniques, both agglomerative and divisive, the user often needs to consider the following points (in addition to the choice of proximity measure):

- graphical display of the clustering process
- comparison of dendrograms
- mathematical properties of methods
- choice of partition
- hierarchical algorithms.

These will be discussed briefly in this section. (Further general information on some of these points will be given in Chapter 9.)

4.4.1 Dendrograms and other tree representations

The dendrogram, or tree diagram, is a mathematical and pictorial representation of the complete clustering procedure, as already illustrated. Here some terminology is given (see Figure 4.6). The *nodes* of the dendrogram represent clusters, and the lengths of the stems *(heights)* represent the distances at which clusters are joined, as already defined in Section 4.2.1. As noted in Section 4.2.4, the stems may be drawn so that they do not extend to the zero line of the diagram, in order to indicate the order in which objects first join clusters. Dendrograms which do not have numerical information attached to the stems are termed *unweighted* or *ranked.* Most dendrograms have two edges emanating from each node *(binary trees).* The arrangement of nodes and stems is the *topology* of the tree.

The names of objects attached to the terminal nodes are known as *labels.* Internal nodes are not usually labelled, although Carroll and Chang (1973) give an example of this where, for instance, the internal node ‘arm’ is above the terminal nodes ‘elbow’ and ‘hand’. Typical or representative members of the clusters can be associated with the internal nodes, called *exemplars* or *centrotypes,* and are defined as the objects having the maximum within-cluster average similarity

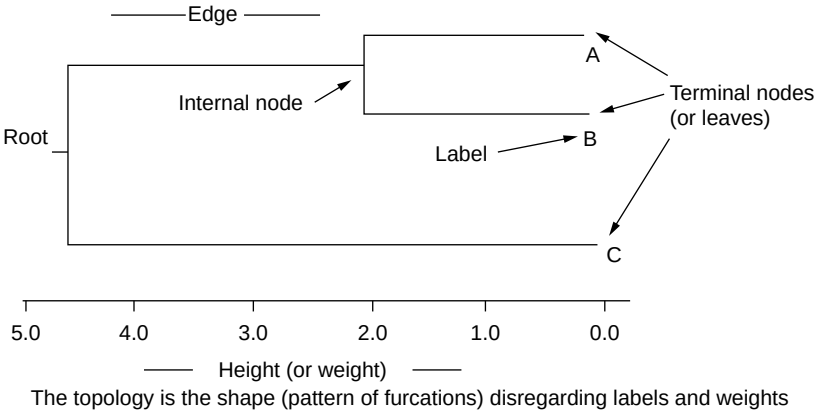


Figure 4.6 *Some terminology used in describing dendrograms.*

(or minimum dissimilarity). A particular type of centrotype is the *medoid* (the object with the minimum *absolute* distance to the other members of the cluster). The dendrogram itself describes the process by which the hierarchy has been obtained, whereas the exemplar and internal node labels describe particular partitions, once these have been chosen.

It is important to realize that the same data and clustering procedure can give rise to 2^{n-1} dendrograms with different appearances, depending on the order in which the nodes are displayed. This can be envisaged by imagining the dendrogram as a mobile in three-dimensional space: the stems from each node can swing around through 180 degrees without changing inter-cluster relationships. Most software packages choose the algorithm for drawing dendrograms automatically, but algorithms for optimizing the appearance of dendrograms have been developed, for example by using internal (Gale *et al.*, 1984) or external (Degerman, 1982) evidence. Wishart (1999) has proposed a robust method that optimizes the rank order of the proximities. The method involves considering each cluster fusion in turn, by reversing the order of the cases within the cluster but without affecting the topology of the tree so as to optimize an objective function.

A number of extensions to dendrograms have been developed. *Espaliers*, for example, are generalized dendrograms in which the length of the horizontal line conveys information about the relative homogeneity and separation of clusters. Hansen *et al.* (1996) discuss the details of these, and give a number of examples and an algorithm for converting a standard dendrogram into an espalier. The *pyramid* is a further specialized type of dendrogram for representing overlapping clusters (see Chapter 8). De Soete and Carroll (1996) give examples of these and other types of tree representation.

The additive tree (or *path length tree*) is a generalization of the dendrogram in which the lengths of the paths between the nodes represent proximities between objects, and in which the *additive inequality* (or *four-point condition*) holds. This

generalization of the ultrametric inequality (see Section 4.4.3) is a necessary and sufficient condition for a set of proximities to be represented in the form of an additive tree. The additive inequality is as follows:

$$d_{xy} + d_{uv} \leq \max|d_{xu} + d_{yv} + d_{yu}| \text{ for all } x,\ y,\ u,\ v. \tag{4.11}$$

Further details are given in Everitt and Rabe-Hesketh (1997), and an example of an additive tree showing genetic associations between various ethnic groups has kindly been provided by Kenneth Kidd (see Figure 4.7). This is a representation of pairwise genetic distances among 30 human populations, generated by a searching routine (Kidd and Sgaramella-Zonta, 1971) that makes topological changes around small or negative branches starting from the neighbour-joining tree produced by the PHYLIP package (Felsenstein, 1989). These branch lengths are the least-squares solution to the complete set of linear equations that relate each pairwise distance to the sum of the branch lengths connecting those populations. For these 30 populations there are $n(n-1)/2 = 435$ pairwise distances to be explained by addition of different combinations of $2n - 3 = 57$ branch lengths. Each tree topology is represented by a different set of equations. Of the 8.69×10^{36} possible trees (sets of linear equations), only about 100 were actually evaluated, and the tree in Figure 4.7 had the smallest $\sum e^2$ (the quantity minimized by least squares

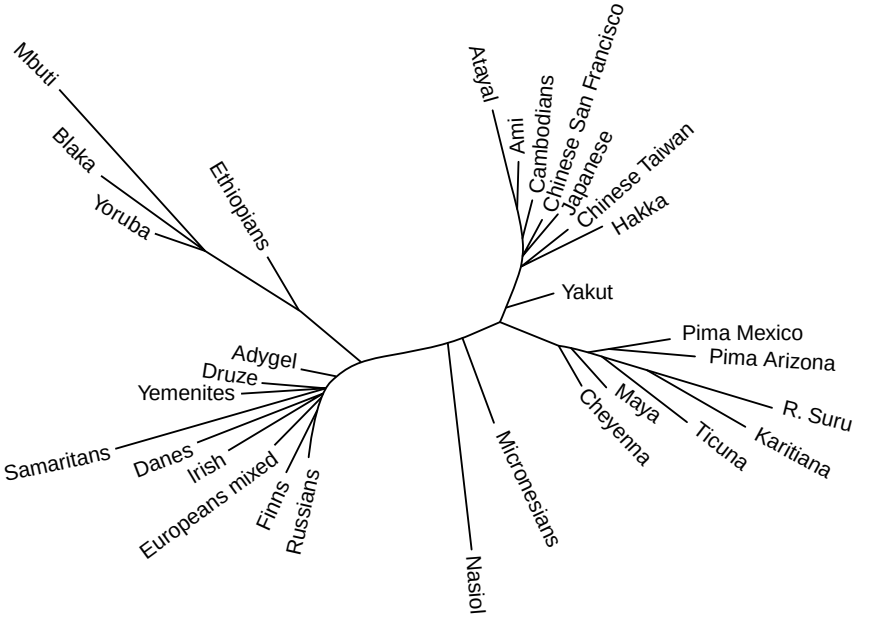


Figure 4.7 *An additive tree representation of pairwise genetic distances among 30 human populations; descriptions of these populations can be found in the ALFRED database at http://info.med.yale.edu/genetics/kkidd.* (Reproduced with permission from Kenneth K. Kidd.)

for each set of linear equations); several others were almost as good as this tree, with only small differences around the very small branches.

Unlike the dendrogram, where each terminal node is equidistant from a single node at the top of the hierarchy, this type of tree is not so rooted. As a concrete example of this, consider the case in which the distances represent genetic differences between species, based on the total number of mutations from a common origin. The additive tree would allow the estimation of the evolutionary time between the appearance of two species from the total path length between them, but it would not be possible to say for certain which species was earlier, since their common origin could be placed at any internal node in the tree. This example is typical of applications of additive trees, which are commonly employed in evolutionary studies to reconstruct phylogenies.

4.4.2 Comparing dendrograms and measuring their distortion

It may be required to compare two dendrograms without making a particular choice as to the particular partition corresponding to a specific number of clusters. Furthermore, hierarchical clustering techniques impose a hierarchical structure on data and it is usually necessary to consider whether this type of structure is acceptable or whether it introduces unacceptable distortion of the original relationships amongst the objects as implied by their observed proximities. Two measures commonly used for comparing a dendrogram with a proximity matrix or with a second dendrogram are the *cophenetic correlation* and *Goodman and Kruskal's* γ.

The starting point for either of these is the so-called *cophenetic matrix*. The elements of this matrix are the heights, h_{ij}, where two objects become members of the same cluster in the dendrogram. It is unaffected by the indeterminacy of the appearance of the dendrogram. The cophenetic matrix $\mathbf{H}$ for the single linkage example in Section 4.2.1 is

$$\mathbf{H} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 5.0 & 5.0 & 0.0 & & \\ 5.0 & 5.0 & 4.0 & 0.0 & \\ 5.0 & 5.0 & 4.0 & 3.0 & 0.0 \end{pmatrix}.$$

The *cophenetic correlation* is the product moment correlation between the $n(n-1)/2$ entries (h_{ij}) in the appropriate cophenetic matrices (excluding those on the diagonals). The matrix is most conveniently arranged in vector form. For example, the off-diagonal elements of $\mathbf{H}$ and the original proximity matrix $\mathbf{D}_1$ in Section 4.2.1 are as follows:

$$\begin{array}{lcccccccccc} \mathbf{H}: & 2, & 5, & 5, & 5, & 5, & 5, & 5, & 4, & 4, & 3 \\ \mathbf{D}_1: & 2, & 6, & 10, & 9, & 5, & 9, & 8, & 4, & 5, & 3. \end{array}$$

The cophenetic correlation between the distance matrix $\mathbf{D}_1$ and $\mathbf{H}$ is 0.82.

Another, nonparametric measure of association is Goodman and Kruskal's γ, defined as $(S_+ - S_-)/(S_+ + S_-)$, where S_+ and S_- are the number of concordances and discordances, respectively. A concordance or discordance in the context of matrix comparison is defined by comparing each pair of pairs. For example, the pairs h_{12} and h_{14} in $\mathbf{H}$ and d_{12} and d_{14} in $\mathbf{D}_1$ are discordant because $2 < 5$ in $\mathbf{H}$ and $2 < 10$ in D_1. For these data, γ is 1.0.

Further information on dendrogram comparison and some applications are given in Chapter 9.

4.4.3 Mathematical properties of hierarchical methods

A number of mathematical properties can be defined for clustering methods. One of these, the *ultrametric property,* was first introduced by Hartigan (1967), Jardine *et al.* (1967) and Johnson (1967), and has since been shown to be related to various features of clustering techniques, in particular the ability to represent the hierarchy by a dendrogram. The *ultrametric property* states that

$$h_{ij} \leq \max\left(h_{ij}, h_{jk}\right) \text{ for all } i, j \text{ and } k, \tag{4.12}$$

where h_{ij} is the distance between clusters i and j. An alternative way of describing this property is that for any three objects, the two largest distances between objects are equal. The property does not necessarily (or even usually) hold for the elements of proximity matrices. However, it does hold for the heights h_{ij} at which two objects become members of the same cluster in many hierarchical clustering techniques.

A consequence of failing to obey the ultrametric property is that *inversions* or *reversals* can occur in the dendrogram. This happens when the fusion levels do not form a monotonic sequence, so that a later fusion takes place at a lower level of dissimilarity than an earlier one. Morgan and Ray (1995) describe some empirical studies of inversions for a number of methods. Inversions are not necessarily a problem if the interest is in one particular partition rather than the complete hierarchical structure. They may also be useful in indicating areas where there is no clear structure (Gower, 1990). However, as Murtagh (1985) points out, reversals can make interpretation of the hierarchy very difficult, both in theoretical studies of cluster properties and also in applications where a hierarchical structure is an intrinsic part of the model. This is because the nested structure is not maintained, as shown in Figure 4.8. Both centroid and median clustering can produce reversals.

A related feature of clustering methods is their tendency to 'distort' space. The 'chaining' effect of single linkage, in which dissimilar objects are drawn into the same cluster, is an example of such distortion, *space contraction* in this case. The opposite type of distortion, where the process of fusing clusters tends to draw clusters together, is *space dilation,* as found in complete linkage. *Space-conserving* methods, such as group average linkage, obey the following inequality:

$$d_{iuv} \leq d_{i(uv)} \leq D_{iuv}, \tag{4.13}$$

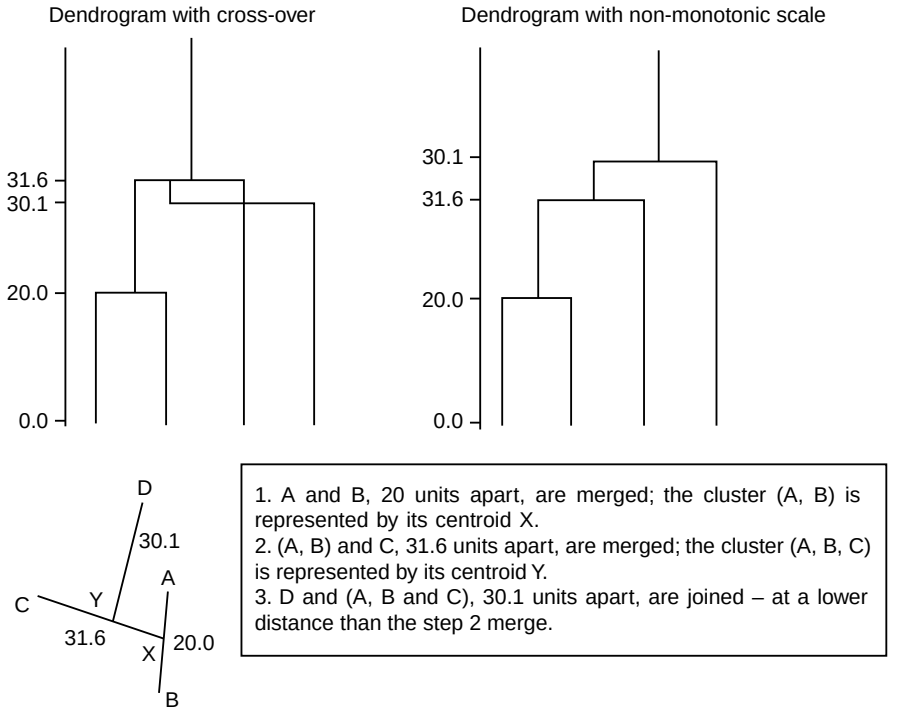


Figure 4.8 *Example of an inversion or reversal in a dendrogram.* (Adapted with permission of the publisher, Blackwell, from Morgan and Ray, 1995.)

where d_{iuv} and D_{iuv} are the minimum and maximum distances between object i and clusters u and v, respectively, and $d_{i(uv)}$ is the distance between object i and the fusion of clusters u and v. In other words, distances to merged clusters are intermediate between distances to the constituent clusters. Space-conserving methods can be thought of as 'averaging' the distances to clusters merged, while space-dilating (-contracting) methods move the merged clusters further from (closer to) each other.

A number of admissibility properties were introduced by Fisher and Van Ness (1971). Such properties would be desirable qualities, other things being equal, and as such they can aid in the choice of an appropriate clustering method. One of these, (*k-group*) *well-structured admissibility,* has been related to the Lance and Williams parameters by Mirkin (1996), who terms it *clump admissibility.* (There are a number of other subtypes of well-structured admissibility, but this one relates directly to space conservation and the ultrametric condition.) Mirkin defines this property as follows:

- *Clump admissibility*: there exists a clustering such that all within-cluster distances are smaller than all between-cluster distances.

Mirkin shows that clump admissibility and space conservation is equivalent to the following conditions, for any x and y such that $0 < x < 1$ and $y > 0$:

$$\begin{aligned} &\alpha(x,y) + \alpha(1-x,y) = 1; \\ &\beta(x,1-x,y) = 0; \\ &|\gamma(y)| \leq \alpha(x,y), \end{aligned} \tag{4.14}$$

where α, β and γ are the parameters in the Lance–Williams recurrence formula, expressed as functions of the cluster sizes, with $x = n_k/n_+$, $y = n_i/n_+$ and $z = n_j/n_+$, where $n_+ = n_i + n_j + n_k$ (see Equation (4.2) and Table 4.2).

Ohsumi and Nakamara (1989) and Chen and Van Ness (1996) also relate the Lance and Williams parameters to the space-conserving property. In addition to the well-structured admissibility property, a few of the more specialized, but nonetheless useful, properties are now summarized:

- *Convex admissibility*: if the objects can be represented in Euclidean space, the convex hulls of partitions never intersect.
- *Point proportional admissibility*: replication of points does not alter the boundaries of partitions.
- *Monotone admissibility*: monotonic transformation of the elements of the proximity matrix does not alter the clustering.

Table 4.2 summarizes the mathematical properties of the well-established hierarchical methods already introduced in Table 4.1, and also gives the Lance and Williams recurrence formula (see Section 4.2.3). The ultrametric property is denoted U, and the convex, point proportion and monotone admissibility properties are denoted C, P and M, respectively.

The convex admissibility property avoids one cluster 'cutting through' another and is only appropriate when the clusters are defined in Euclidean space. In the absence of further information about cluster structure, one would prefer to avoid such cuts, although many standard procedures (for example single and complete linkage) do not in fact have this property.

Point proportionality would be relevant in situations where samples might contain replicated observations (including cases in which sets of different objects had identical characteristics). Indeed, some software packages for clustering allow *case weights* to reflect replication. An example of replicated observations would be in systems for automatic monitoring of keywords contained in web pages; these are often derived from a number of different sources (see Kirriemuir and Willett (1995), for example). Point proportionality would also be helpful in achieving robustness if there were likely to be sets of observations differing only by a small amount.

The monotone property would be appropriate where only the rank-order information was reliable; for example, where the proximity matrix contained elements which had been obtained directly from subjective ratings, such as preference or brand switching matrices in market research (see the cola example in Chapter 1).

4.4.4 Choice of partition – the problem of the number of groups

It is often the case that an investigator is not interested in the complete hierarchy but only in one or two partitions obtained from it, and this involves deciding on the number of groups present. There are a variety of formal methods that apply equally well to hierarchical clustering and optimization methods. Some of these will be discussed in Section 5.5. In this chapter we concentrate on the methods that are specific to hierarchical techniques.

In standard agglomerative or polythetic divisive clustering, partitions are achieved by selecting one of the solutions in the nested sequence of clusterings that comprise the hierarchy, equivalent to cutting a dendrogram at a particular height (sometimes termed the *best cut*). This defines a partition such that clusters below that height are distant from each other by at least that amount, and the appearance of the dendrogram can thus informally suggest the number of clusters. Large changes in fusion levels are taken to indicate the best cut. A more flexible development of this idea is 'dynamic tree cutting' (Langfelder *et al.*, 2008). This allows for different branches of the tree to be cut at different levels. The process iterates until the number of clusters is stable by combining and decomposing clusters, making successive cuts of the sub-dendrograms within clusters based on their shape. This has been implemented as a package in `R` (`dynamicTreeCut`), and is particularly appropriate where there are sets of nested clusters. As with the more inflexible fixed-height cut methods, parameters for the cut heights and the minimum cluster sizes must be chosen, so there is the possibility of influence from *a priori* expectations.

More formal approaches to the problem of determining the number of clusters have been reviewed by Milligan and Cooper (1985). They identified five best-performing rules which were further investigated by Gordon (1998), who found that the two developed by Duda and Hart (1973) and Beale (1969a) are appropriate for the nested structure inherent in hierarchical methods, since they test whether a cluster should be divided. Both are based on the ratio of between-cluster to within-cluster sums of squares, when the cluster is optimally divided into two (see Section 5.5).

Other 'number of groups' procedures which are particularly suitable for hierarchical methods have been suggested by Mojena (1977). The first is based on the relative sizes of the different fusion levels in the dendrogram and is sometimes known as the *upper tail rule.* In detail, the proposal is to select the number of groups corresponding to the first stage in the dendrogram satisfying

$$\alpha_{j+1} > \bar{\alpha} + ks_{\alpha}, \tag{4.15}$$

where $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ are the fusion levels corresponding to stages with n, $n-1, \ldots, 1$ clusters. The terms $\bar{\alpha}$ and s_{α} are respectively the mean and unbiased standard deviation of the j previous fusion levels, and k is a constant. Mojena suggests that values of k in the range 2.75–3.50 give the best results, although

Milligan and Cooper suggest 1.25. Alternatively, one can use a t-distribution (although this assumes an underlying normal distribution which is clearly not applicable to fusion levels). A visual approach is to identify breaks in the plot of the values $(\alpha_{j+1}-\bar{\alpha})/s_\alpha$ against the number of clusters j.

The second method proposed by Mojena is based on a moving average approach, familiar from ideas in quality control. Here the rule is to use the partition corresponding to the first stage j, in a partial cluster sequence from $j=r$ to $j=n-2$ clusters, satisfying

$$\alpha_{j+1} > \bar{\alpha} + L_j + b_j + ks_j, \tag{4.16}$$

where $\bar{\alpha}$ and s_j are the mean and standard deviation of the fusion values, now based on the previous t-values; L_j and b_j are corrections to the mean for the upward trend in fusion values (L_j is the 'trend lag', in quality control jargon, equal under certain simplifying assumptions to $(r-1)b_j/2$, where b_j is the moving least-squares slope of the fusion levels).

According to Wishart (1987), this second rule has the advantage that the fusion level being considered does not enter into the sample statistics, but the disadvantage is that the value of r has to be chosen by the investigator. In both cases it is usual to order the criterion values and then to choose the lowest number of clusters where the rule is satisfied. Results using the upper tail rule will be illustrated in Section 4.5.5.

Given the lack of consensus about which rule to apply (and the varying results on the same data), the following comment is appropriate (Baxter, 1994): 'informal and subjective criteria, based on subject expertise, are likely to remain the most common approach. In published studies practice could be improved by making such criteria more explicit than is sometimes the case'. A discussion of whether a data set shows *any* evidence of clustering is left to Chapter 9.

4.4.5 Hierarchical algorithms

It is worth distinguishing between hierarchical *methods* and hierarchical *algorithms* for computing the clustering. For any given hierarchical method, several different computational algorithms may be used to achieve the same result. Day (1996) discusses efficient algorithms for a wide variety of clustering methods, including hierarchical techniques, and gives a comprehensive bibliography. A more recent review is by Gascuel and McKenzie (2004). Many algorithms produce a nested structure by optimizing some criterion in a stepwise manner, whereas others operate globally, for example by minimizing the distortion, as discussed in Section 4.4.2, that results from representing the proximity matrix by a hierarchical structure. These global algorithms are known as *direct optimizing algorithms.* An early example was given by De Soete (1984a); Gordon (1998) and De Soete and Carroll (1996) give more examples and further references. A method for finding *parsimonious trees* (those with a minimum number of levels

in the hierarchy) was proposed by Sriram and Lewis (1993). Direct optimizing algorithms can be useful when elements of the proximity matrix are missing (De Soete, 1984b).

Zahn (1971) gives a number of graph-theoretical clustering algorithms based on the *minimum spanning tree*. A *graph* is a set of nodes and of relations between pairs of nodes indicated by joining the nodes by *edges*. A set of observations and their dissimilarities can be represented in a graph as nodes and edges, respectively. A spanning tree of a graph is a set of edges which provides a unique path between every pair of *nodes*, and a minimum spanning tree is the shortest of all such spanning trees. Minimum spanning trees are related to single linkage algorithms (see Gower and Ross, 1969).

4.4.6 Methods for large data sets

For very large data sets, where standard methods may be unable to cope, specialized methods have been developed, for example by Zupan (1982). The use of parallel computer hardware has become a possibility (see, for example, Tsai *et al*., 1997; Rasmussen and Willett, 1989). Some methods for large data sets combine a hierarchical method with a preclustering or sampling phase. Three that have become relatively widely used are BIRCH, CURE and `SPSS TwoStep`. BIRCH (Zhang *et al*., 1996) employs a preclustering phase where dense regions are summarized, the summaries being then clustered using a hierarchical method based on centroids. CURE (Guha *et al*., 1998) starts with a random sample of points, and represents clusters by a smaller number of points that capture the shape of the cluster, which are then shrunk towards the centroid so as to dampen the effects of outliers; hierarchical clustering then operates on the representative points. CURE has been shown to be able to cope with arbitrary-shaped clusters, and in that respect may be superior to BIRCH, although it does require a judgement as to the number of clusters and also a parameter which favours more or less compact clusters. `Two-Step`'s first step is similar to BIRCH in that it forms 'preclusters' by detecting dense regions; at this step, outliers (clusters with few cases, e.g. $<5\%$,) can be rejected before the next stage. The second step has some overlap with the model-based methods described in Chapter 6, in that one of the possible distance measures used is a combination of the likelihoods calculated for the continuous (assuming multivariate mixtures of normal distributions) and for the categorical variables (assuming multinomial distributions). The original paper is that by Chiu *et al*. (2001); some further details were obtained by Bacher *et al*. (2004) by personal communication from these authors (and may be subject to change as the method is developed). Bacher *et al*. described an evaluation by simulation, and found that while the method worked well for continuous variables, it was outperformed by latent class models (see Chapter 6). They also point out an important aspect of the method, to do with the analysis of data with different measurement types, which is one of its features. This is the problem of commensurability – the need to standardize variables to a common scale of measurement – as discussed in general in Chapter 3. In the `TwoStep` method

using the log-likelihood distance, a difference of 1 in a categorical variable is equal to a difference of 2.45 scale units in a z-scored variable (and other standardizations give different relative weights to continuous and categorical variables). An interesting application of the method from meteorology is given by Michailidou *et al.* (2009).

4.5 Applications of hierarchical methods

In this section we describe applications of a number of the clustering techniques discussed earlier, with particular reference to some of the points raised in Section 4.4.

4.5.1 Dolphin whistles – agglomerative clustering

In a study of dolphin behaviour, Janik (1999) used two hierarchical methods of clustering to categorize dolphin whistles. The results were compared to human classification and to a previously published *k*-means approach (see Chapter 5 for a description of *k*-means). Bottle-nosed dolphins produce a variety of whistles, but each animal usually has a stereotypical 'signature whistle', a characteristic sound which is used exclusively and frequently by that animal. The data for this study were a sample of 104 line spectrograms of the whistles of four dolphins, both signature and nonsignature. A selection of signature spectrograms is shown in Figure 4.9, which also shows the human classification (D was subdivided into D_1 and D_2).

Proximity matrices were calculated using two methods: cross-correlation between the spectrograms, after first aligning them to have maximum correlation, thus producing a similarity coefficient; and average absolute difference between the spectrograms, thus producing a dissimilarity coefficient (the city block distance; see Chapter 3). Two procedures, average linkage and complete linkage, were applied to the data. The first was chosen as a representative of a commonly used procedure in biological science, and the second selected as a procedure that would be expected to identify very stereotypical whistle types.

The results for the average linkage procedure applied to the city block (average distance) proximity matrix are shown in Figure 4.10. The dendrogram has been cut at a level where the human-identified signature whistles cluster into reasonably distinct (although not perfectly separated) groups. The other procedure gave very similar results.

The categorization by a group of five humans was taken to be 'correct' for the signature whistles, since they showed high internal consistency. Neither human nor automatic methods showed consistency in grouping the 'nonsignature' whistles, and the authors concluded that the difficulty in using automated methods was not in the methods used but in the definition of similarity. They discuss the effect of duration of the whistle in the comparison of the spectrograms, the weight to be applied to different parts of the spectrogram, and

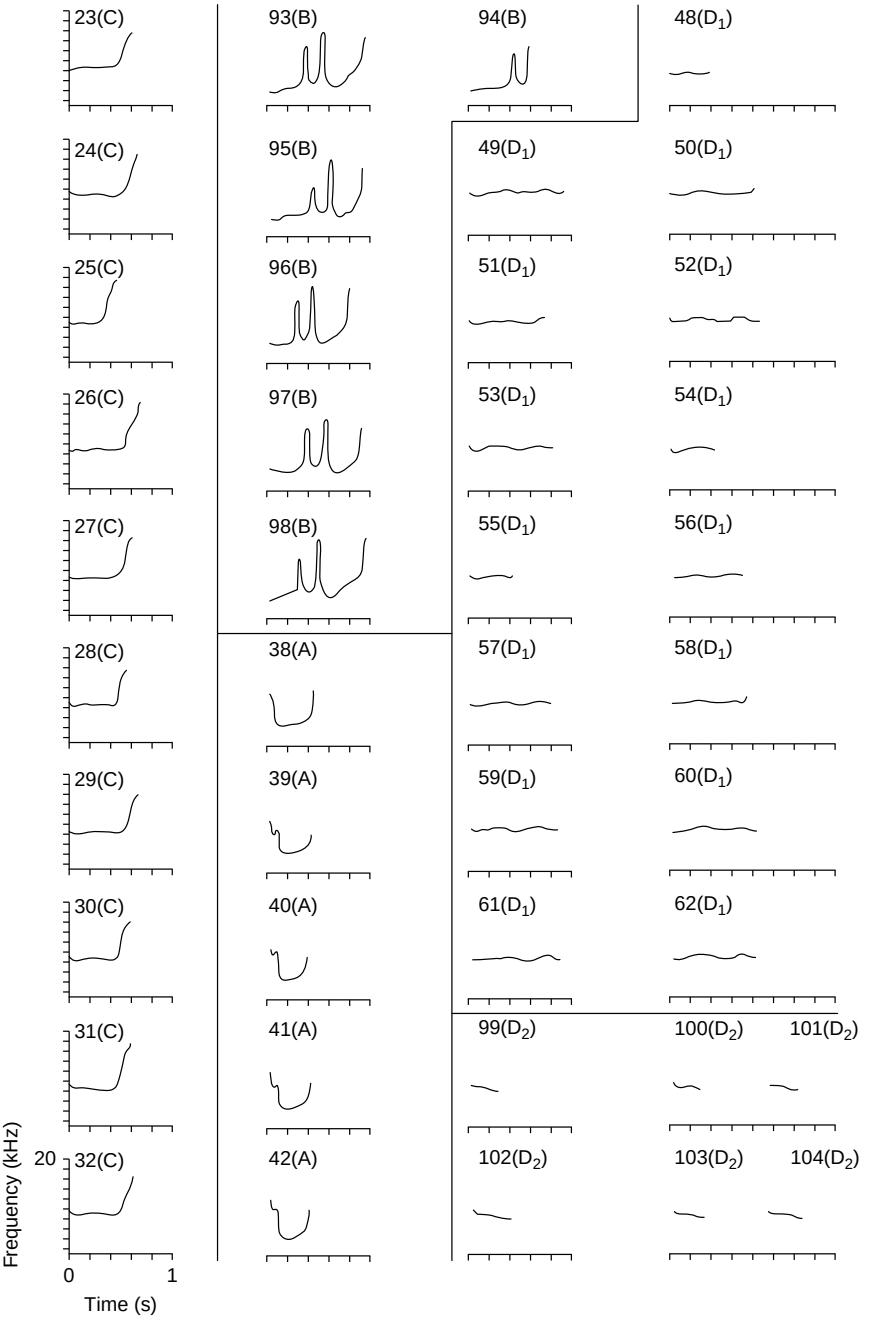


Figure 4.9 *Line spectrograms representing signature dolphin whistles. Letters indicate an expert's classification.* (Source: Janik, 1999.)

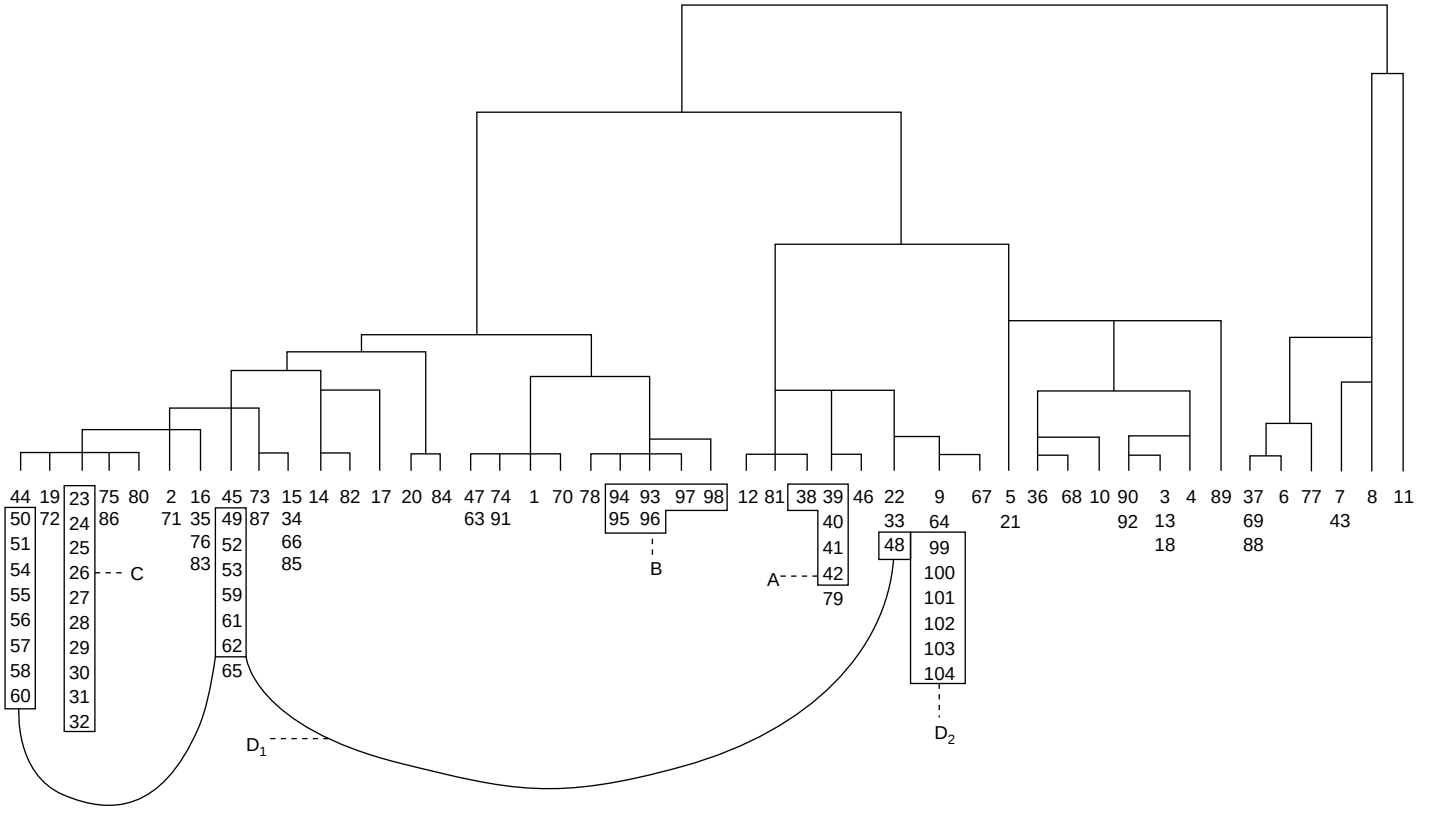


Figure 4.10 *Dendrogram of average linkage applied to dolphin whistle data, also showing human classification; see also Figure 4.9. Signature whistles are boxed.* (Source: Janik, 1999.)

whether dolphins are sensitive to the actual frequency or to the general shape of the signal. They conclude that proximity measures need to take account of the patterns that are relevant to the animal, and these need to be elicited by experimental methods.

4.5.2 Needs of psychiatric patients – monothetic divisive clustering

To illustrate the use of a monothetic divisive method, a data set from a European study of the needs, symptoms and mental health service use of schizophrenics in Amsterdam, Copenhagen, London, Verona and Santander (Becker *et al*., 1999) will be analysed. A subset of the data (male patients in London) is used. The binary variables are the presence or absence of a need in the following 'needs domains': accommodation, daytime activities, physical health, information about treatment, company. These are a subset of the 22 domains measured by the Camberwell Assessment of Need – European Version (McCrone *et al*., 2000). The data are shown in Table 4.3.

The criterion for splitting at each stage is the difference of cross-products (Equation (4.9)). At each stage the variable with the highest criterion value (summed over all the other variables) is chosen to make a split: by splitting on this variable one automatically carries other associated variables with it, so that the total number of splits is minimized. The results are illustrated in the banner plot shown in Figure 4.11.

In this case the 'best' splitting variable (that with the highest association with others) is 'information', which divides the data into 11 (with a need for information) and 23 (without an information need). Those with a need for information can then be divided into those with and without a need for daytime activities, and those without a need for information can be divided into those with and without a need for accommodation. At this separation stage there are four clusters. Looking at the right-hand side of the diagram, it can be seen that the final splits are based on the physical health needs, which are relatively unassociated with the other needs. There is one substantial group with completely homogeneous needs: five people have no need apart from physical health, a medical domain, rather than a social domain like the others. The identification of such subgroups could be used (in a larger study) to devise appropriate mental health services by grouping together people with similar needs.

4.5.3 Globalization of cities – polythetic divisive method

The following example of a polythetic divisive method involves the possible globalization of cities; that is, the tendency of cities across the globe to become (commercially) similar through the influence of multinational companies. The data describe the presence or absence of six multinational advertising agencies, and the status of the office of five multinational banks – head office, branch or agency, coded as 1, 2 or 3. The data are from Data Set 4 of the

Table 4.3 Needs of schizophrenic men in London according to the Camberwell Assessment of Need.

Patient	Accommodation	Daytime activities	Physical health	Information	Company
1	0	0	1	1	0
2	1	0	1	0	0
3	0	0	0	0	0
4	0	0	0	0	1
5	0	1	1	1	1
6	0	0	1	0	0
7	1	1	1	0	0
8	0	0	0	0	1
9	0	0	1	1	0
10	0	0	0	0	0
11	0	0	0	0	0
12	0	1	1	1	1
13	0	1	0	0	0
14	0	0	1	1	1
15	0	0	1	0	0
16	0	1	0	1	1
17	0	0	0	0	0
18	0	1	0	1	1
19	0	0	0	0	0
20	0	0	0	1	0
21	0	0	1	0	0
22	0	0	1	1	1
23	0	0	0	0	0
24	1	1	0	0	0
25	0	1	0	0	1
26	1	0	0	0	1
27	1	1	1	0	0
28	1	1	1	0	1
29	0	0	0	0	0
30	0	1	1	1	1
31	1	0	1	1	1
32	0	0	1	0	0
33	0	0	0	0	0
34	0	0	1	0	0

Globalization and World Cities Study Group and Network (GaWC) Research Group (Beaverstock *et al.*, 1999). Figure 4.12 shows the cities, which have been previously categorized as alpha, beta or gamma depending on the overall level of economic activity.

The proximity matrix was computed using Gower's mixed data coefficient (see Chapter 3), with the contributions from the binary variables (presence or absence of each advertising agency) entered into an asymmetric (Jaccard)

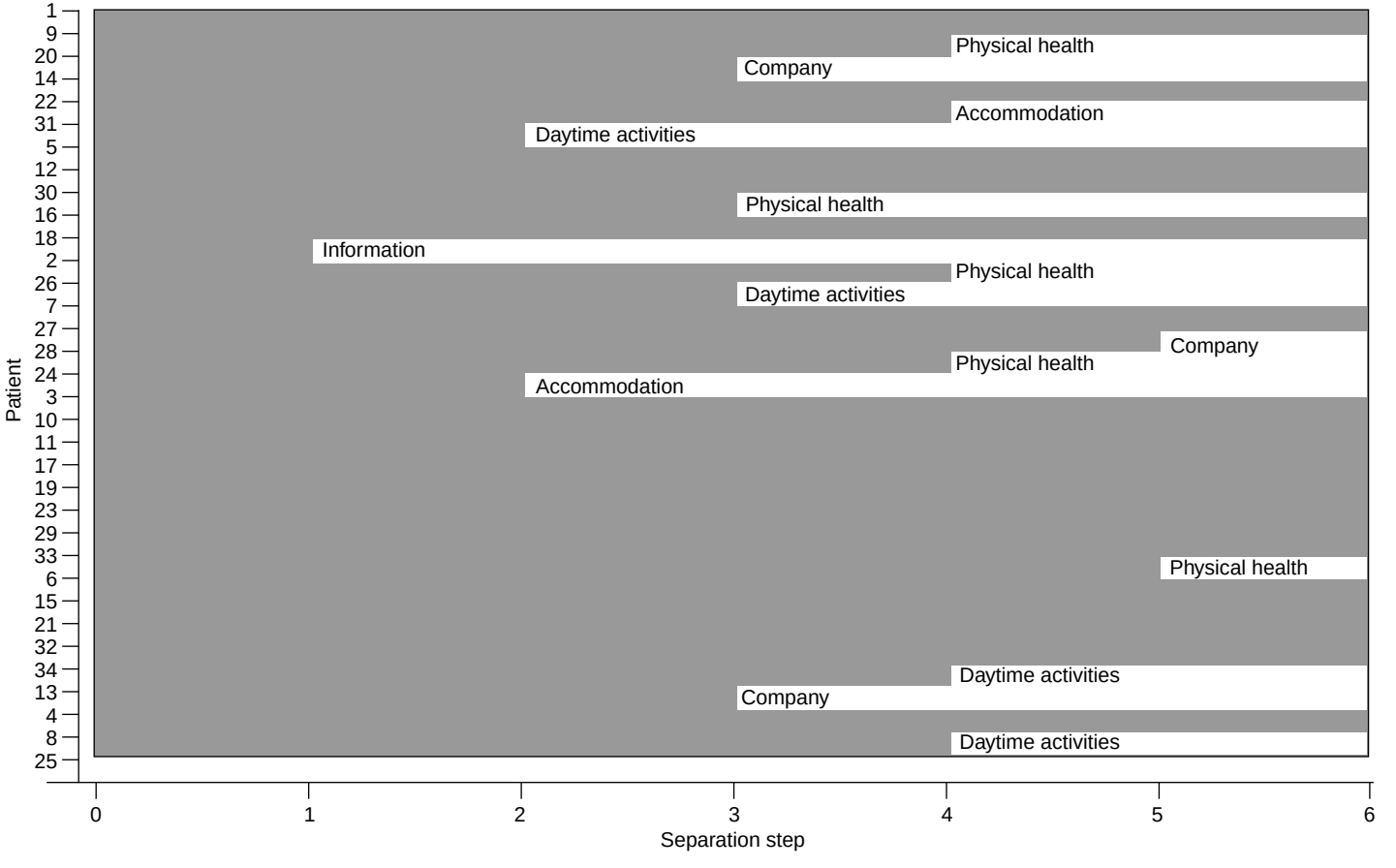
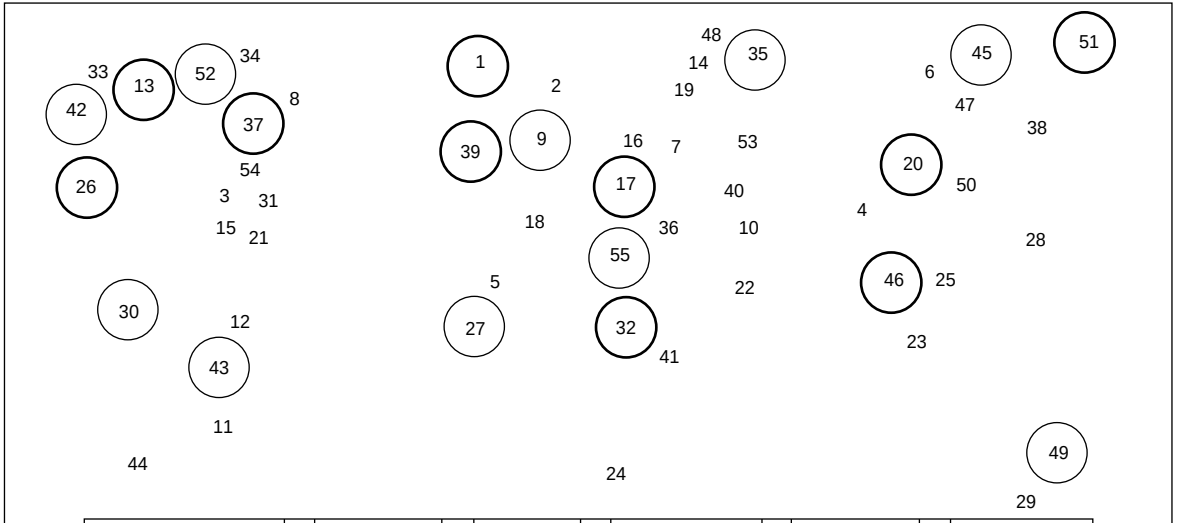


Figure 4.11 *Banner plot showing successive monothetic division of a set of male schizophrenics in London according to their needs* (see also Table 4.3)



Key Bold circles: alpha cities Other circles: beta cities Uncircled: gamma cities																				
		1	London	6	Beijing	11	Buenos Aires	16	Düsseldorf	21	Houston
		2	Amsterdam	7	Berlin	12	Caracas	17	Frankfurt	22	Istanbul
		3	Atlanta	8	Boston	13	Chicago	18	Geneva	23	Jakarta
		4	Bangkok	9	Brussels	14	Copenhagen	19	Hamburg	24	Johannesburg
		5	Barcelona	10	Budapest	15	Dallas	20	Hong Kong	25	Kuala Lumpur
26	Los Angeles	31	Miami	36	Munich	41	Rome	46	Singapore	51	Tokyo
27	Madrid	32	Milan	37	New York	42	San Francisco	47	Shanghai	52	Toronto
28	Manila	33	Minneapolis	38	Osaka	43	São Paulo	48	Stockholm	53	Warsaw
29	Melbourne	34	Montreal	39	Paris	44	Santiago	49	Sydney	54	Washington
30	Mexico City	35	Moscow	40	Prague	45	Seoul	50	Taipei	55	Zurich

Figure 4.12 *Diagrammatic map of global cities, classified according to overall economic activity as 'alpha', 'beta' or 'gamma'.* (Source: Beaverstock et al., 1999.)

coefficient matrix. The ordered variable (head office, subsidiary office or 'representative only', for banks) was included by treating the data as continuous and dividing by 3 to give a value between 0 and 1. An algorithm of Kaufman and Rousseeuw (1990) based on the method of MacNaughton-Smith *et al.* (1964) was applied, and this is illustrated with a dendrogram in Figure 4.13. Many of the so-called 'alpha' cities appear in the same cluster (far left); this is because the proximity measure is strongly influenced by overall activity, and there are many advertising agencies and head, rather than subsidiary, offices for banks in those areas.

Geographical clustering is also obvious and could be used to infer the common areas of influence of specific companies. The geographical clustering is not perfect: if this had been required, a *constrained* method could be used to allow clustering only of cities that were close either geographically as the crow flies or, more usefully, those with good transport links (see Chapter 8 for a more detailed discussion of constrained clustering).

4.5.4 Women's life histories – divisive clustering of sequence data

Piccarreta and Billari illustrate their CART-like monothetic divisive method described in Section 4.3.1 by analysing the employment and family trajectories of women using data from the British Household Panel Survey. The women are categorized in each month in terms of their work status (W), number of children (1, 2, $\geq$3) and whether in a co-resident partnership (U). The possible life states in each month from age 13–30 are then coded from these data, for example the state WU1 denotes work and in a partnership, with 1 child. As an illustration of the concepts described above, the sequence 0–0–W–W–WU–WU–W–W–W produces the auxiliary variable $\boldsymbol{W}_1$ (the month when a woman worked for the first time; in this example, 3), and $\boldsymbol{WU}_1$ (when she was in a partnership and work for the first time; in this example, 5). The state permanence sequence is 0^2–W^2–WU^2–W^3. Figure 4.14 shows the tree derived by the authors from the British Household Panel Survey using their new method.

A full discussion of the tree and its implications for life courses of women is given by the authors. As a starting point for orientation around the tree, we merely comment here that the first split is made on the basis of the time to reaching 'two children without work or further study'; the left-hand side of the tree thus contains women who wait to have two children later in life and who are interpreted as less family oriented. The characteristics of specific clusters are informative: for example, the medoid woman of cluster 3 (which contains 12.3% of the sample) has state permanence sequence 0^{105}–W^{31}–WU^{61}– $1WU^7$. She becomes employed just before 22 years, starts a union within 3 years after and becomes a mother about 5 years after that. This cluster is interpreted as a group of women who combine work and a (relatively delayed) family.

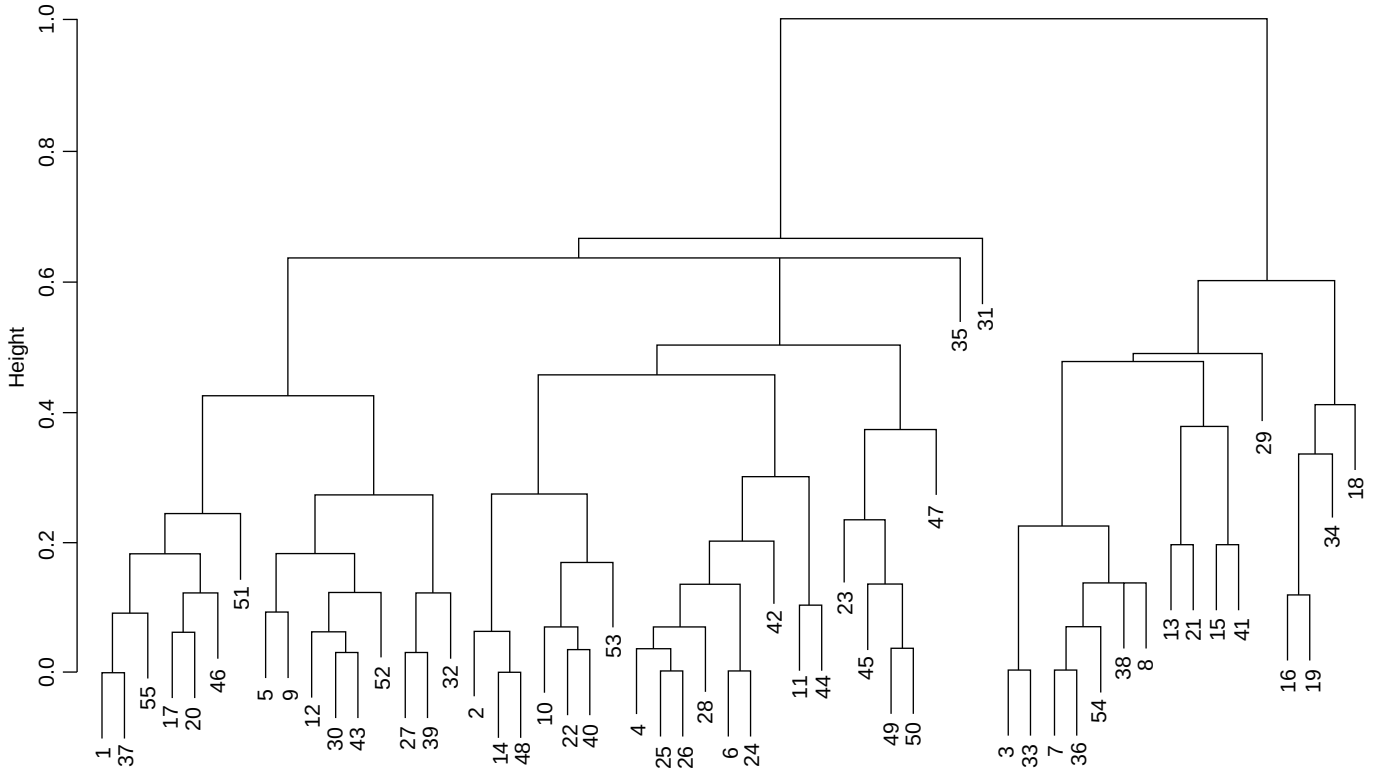


Figure 4.13 *Dendrogram showing polythetic divisive clustering of global cities, based on the strength of presence of advertising agencies and international banks (see also Figure 4.12).* Acknowledgement: the data used is from Data Set 4 from the GaWC Research Group and Network (www.lboro.ac.uk/gawc/datasets/da4.html). It was collected by J.V. Beaverstock, R.G. Smith and P.J. Taylor as part of their ESRC project 'The Geographical Scope of London as a World City' (R000222050)

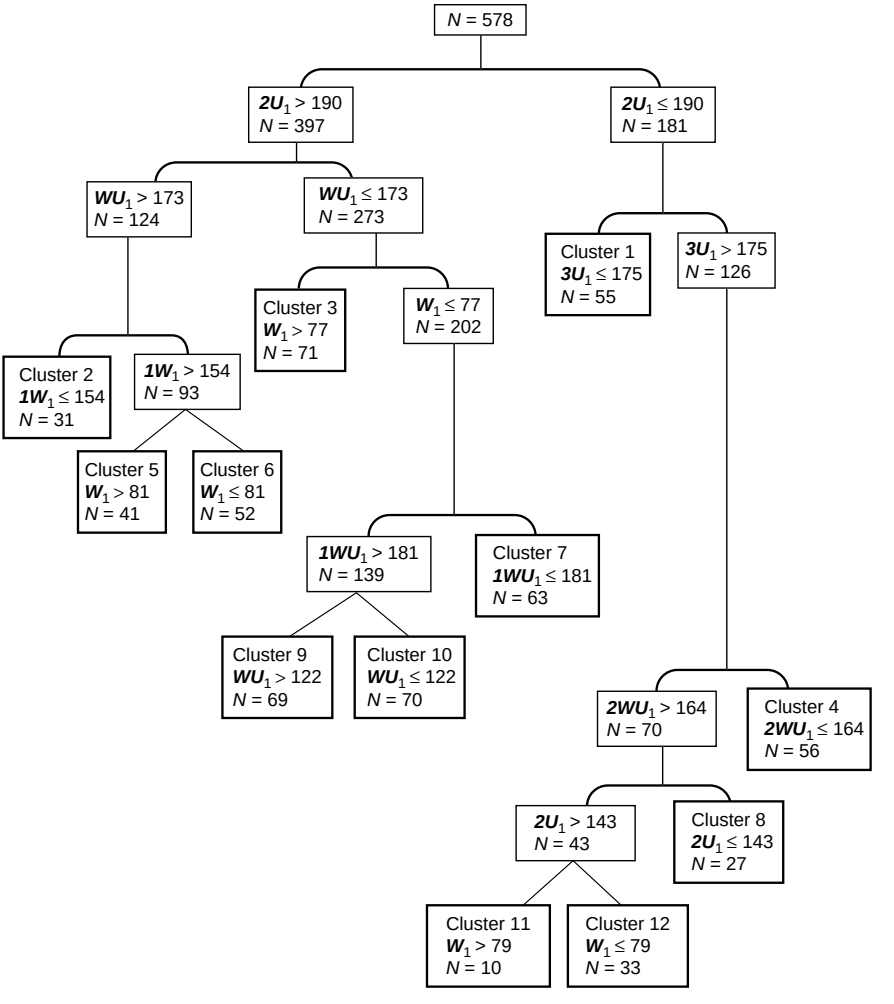


Figure 4.14 *Tree obtained from divisive cluster analysis after pruning. Auxiliary variables used are U_1 (time to being in a union); $1U_1$, $2U_1$, $3U_1$ (time to being in a union with 1, 2 or 3 or more children respectively); W_1 (time to first job); WU_1 (time to first job, being in a union without children); $1WU_1$, $2WU_1$ (time to first job, being in a union and with 1 or 2 children, respectively).*

4.5.5 Composition of mammals' milk – exemplars, dendrogram seriation and choice of partition

The compositions of various mammals' milk are given in Table 4.4 (Hartigan, 1975). Figure 4.15 shows an average linkage clustering based on Euclidean distances, in which the second dendrogram has been seriated so that the order of the labels is optimal. In this second dendrogram, unlike the first, all the equines are contiguous, as are the pairs of primates, camelids and bovines. The number of

Table 4.4 Composition of mammals' milk (percentage) standardized to have zero mean and unit standard deviation.

Mammal	Water	Protein	Fat	Lactose	Ash
Bison	0.681	−0.387	−0.818	0.856	0.073
Buffalo	0.307	−0.085	−0.229	0.310	−0.165
Camel	0.743	−0.742	−0.657	0.365	−0.303
Cat	0.268	1.064	−0.381	0.146	−0.224
Deer	−0.955	1.147	0.893	−0.836	1.063
Dog	−0.145	0.845	−0.077	−0.618	0.667
Dolphin	−2.592	1.201	2.338	−1.764	−0.660
Donkey	0.946	−1.235	−0.847	1.129	−0.918
Elephant	0.628	−0.715	0.693	0.801	−0.462
Fox	0.268	0.106	−0.419	0.419	0.132
Guinea pig	0.291	0.325	−0.295	−0.782	−0.026
Hippo	0.954	−1.536	−0.552	0.146	−1.512
Horse	0.930	−0.989	−0.885	1.511	−1.017
Llama	0.650	−0.633	−0.676	0.801	−0.125
Monkey	0.798	−1.098	−0.723	1.238	−1.353
Mule	0.923	−1.153	−0.809	0.747	−0.779
Orangutan	0.806	−1.317	−0.647	1.020	−1.234
Pig	0.362	0.243	−0.495	−0.236	0.469
Rabbit	−0.535	1.667	0.265	−1.218	2.846
Rat	−0.441	0.818	0.218	−0.454	1.063
Reindeer	−1.041	1.229	0.950	−0.891	1.063
Seal	−2.475	0.955	3.013	−2.256	−0.026
Sheep	0.299	−0.168	−0.372	0.310	0.093
Whale	−1.041	1.338	1.036	−1.382	1.658
Zebra	0.627	−0.879	−0.524	0.638	−0.323

(Taken with permission from Hartigan, 1975.)

clusters was assessed using the upper tail rule, as described in Equation (4.15); the criterion values are as follows (t-values in brackets, found by multiplying by the square root of $n-1$, where n is the number of objects):

- 2 clusters, 4.16 (20.4)
- 3 clusters, 1.59 (7.81)
- 4 clusters, 0.56 (2.73).

In this case, the four-cluster solution seems to be interpretable and is significant at $p = 0.05$, and exemplars for the four-cluster solution are underlined, with average compositions shown in Table 4.5. As might be expected, fat composition is the main identifying variable for cluster 4, with exemplar 'seal'. Cluster 3, with exemplar 'whale', is similar but with less fat and more ash. Cluster 1, with exemplar 'zebra', has the highest average lactose. Cluster 2, with exemplar 'dog', does not have any strong identifying features.

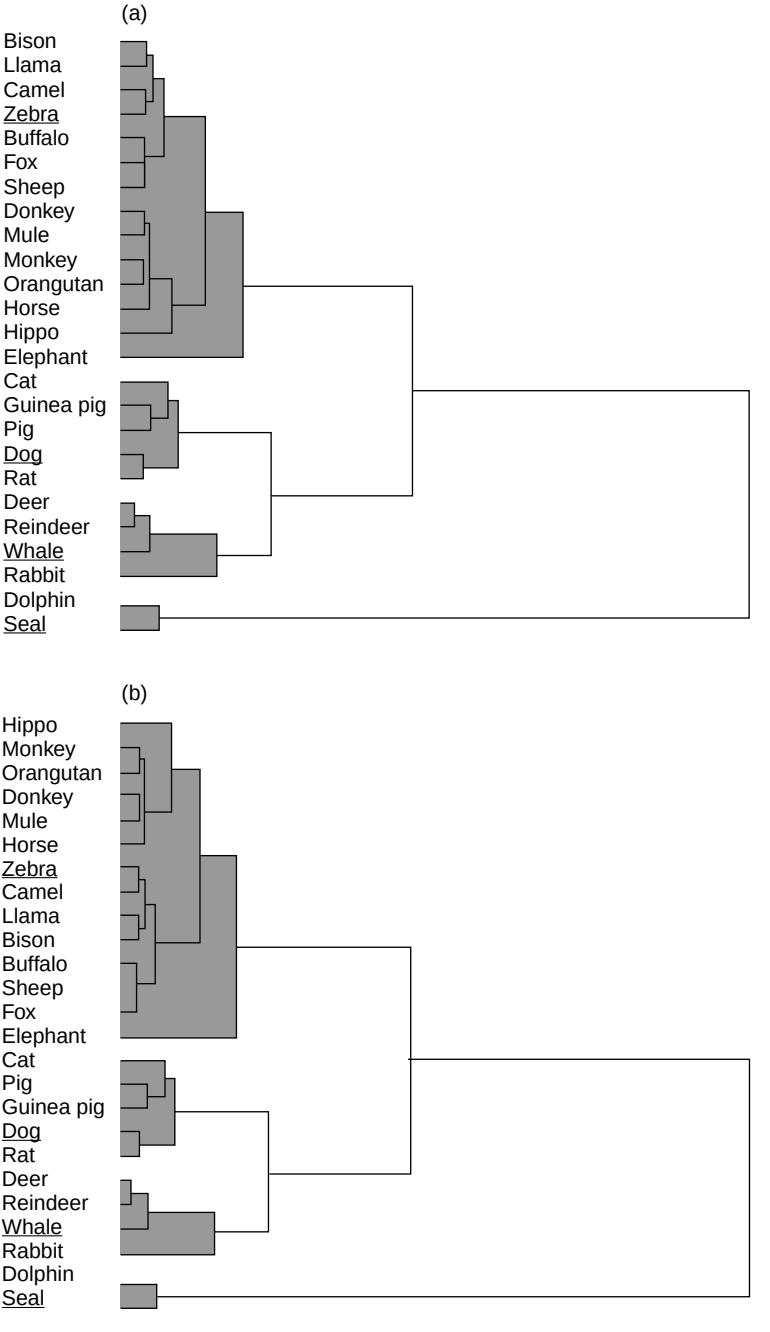


Figure 4.15 *Dendrograms for average linkage clustering of mammals' milk composition, showing 'best cut' (four-cluster solution) and the cluster exemplars; the lower dendrogram (b) was produced by seriating the objects (see also Tables 4.4 and 4.5)*

Table 4.5 Mean percentage composition of mammals' milk clusters.

Cluster (exemplar)	Water	Protein	Fat	Lactose	Ash
1 (Zebra)	85.76	3.39	4.70	5.48	0.58
2 (Dog)	79.02	8.62	8.14	3.42	1.06
3 (Whale)	66.71	11.12	18.58	2.15	1.70
4 (Seal)	45.68	10.15	38.47	0.45	0.69

4.6 Summary

Hierarchical methods form the backbone of cluster analysis in practice. They are widely available in software packages and they are easy to use, although clustering large data sets is time-consuming (methods to get round this can involve hybrid techniques which have preclustering or sampling phases and are usually available only in specialized packages). Choices that the investigator needs to make are the measure of proximity, the clustering method and, often, the number of clusters. The main problem in practice is that no particular clustering method can be recommended, since methods with favourable mathematical properties (such as single linkage) often do not seem to produce interpretable results empirically. Furthermore, to use the results involves choosing the partition, and the best way of doing this is unclear. When a particular partition is required and there is no underlying hierarchy, the methods of Chapter 5 may be more appropriate. Some of the problems of traditional hierarchical methods can be overcome by the use of model-based techniques, as will be discussed in Chapter 6.