Classification Society of North America

Short Course

An Introduction to Classification and Clustering

Session II

Cluster analysis

Pierre Legendre
Université de Montréal

École des Hautes Études Commerciales
Université de Montréal
Montréal, Québec, Canada

June 8, 2000

Outline:

1) Measures of resemblance (proximity)

The *Q* and *R* modes of analysis
Similarities (*S*) and distances (*D*)
Coefficients including (*symmetrical*) or excluding (*asymmetrical*) double-zeros
Coefficients for binary and for quantitative variables

Examples of binary coefficients: Simple matching, Jaccard
Quantitative coefficients: Euclidean distance, Steinhaus and Gower similarities
Examples of *R*-mode coefficients: Pearson's *r*, Spearman's *r*, chi-square

Choosing a coefficient

2) Hierarchical cluster analysis

Single linkage clustering (nearest neighbor)
Complete linkage (furthest neighbor)
Proportional-link linkage clustering
Unweighted arithmetic average clustering (average linkage; UPGMA)
Weighted arithmetic average clustering (WPGMA)
Unweighted centroid clustering (UPGMC)
Weighted centroid clustering (WPGMC)
Ward's minimum variance clustering — From raw data or from a *D* matrix
Flexible clustering

3) Ultrametric property and reversals

4) Partitioning data

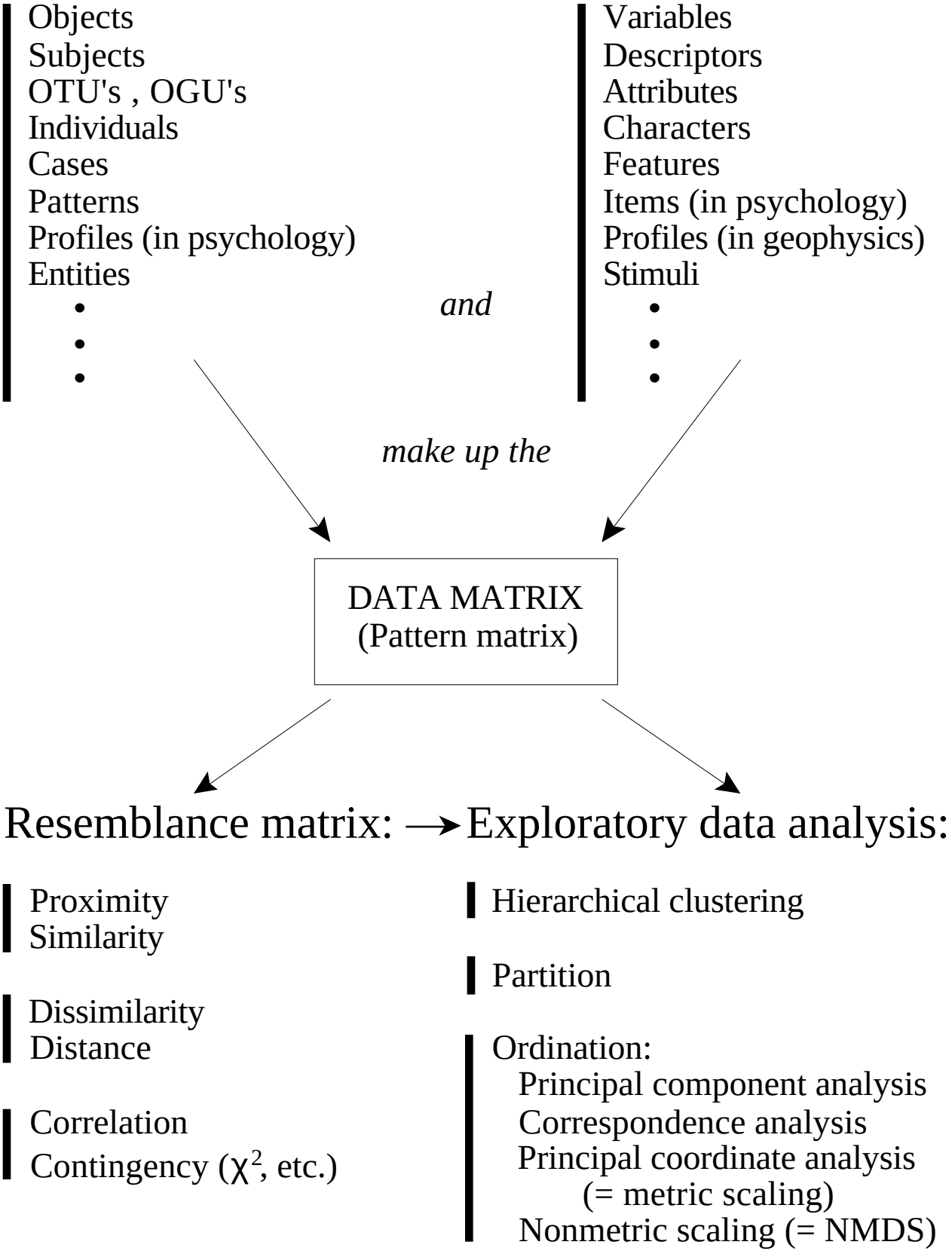
K-means clustering — From raw metric data or from a *D* matrix

5) Cophenetic correlations and other measures of adjustment

6) Representing results of clustering

7) A quick glance at different types of clustering algorithms

The vocabulary varies with the field of application



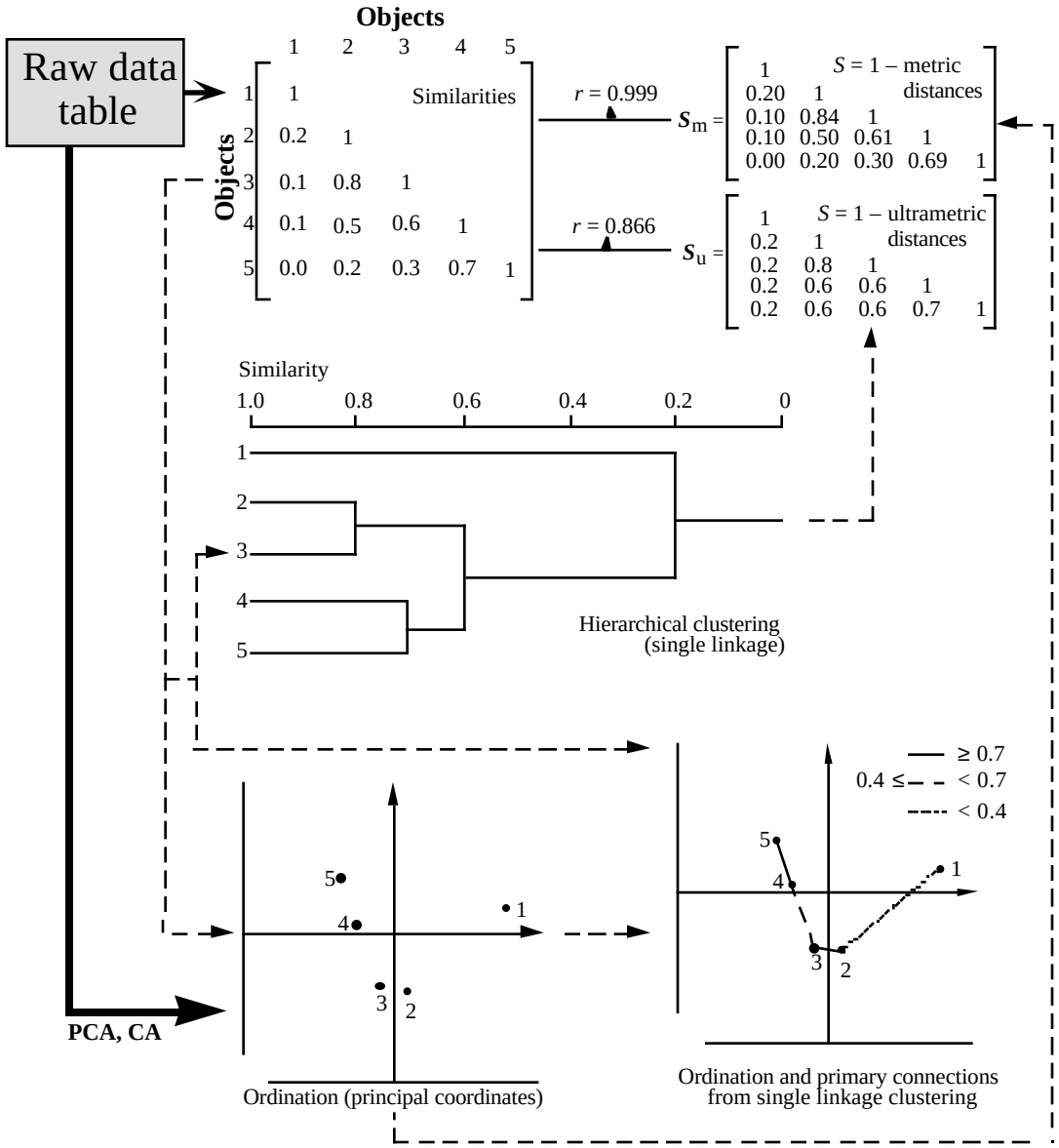


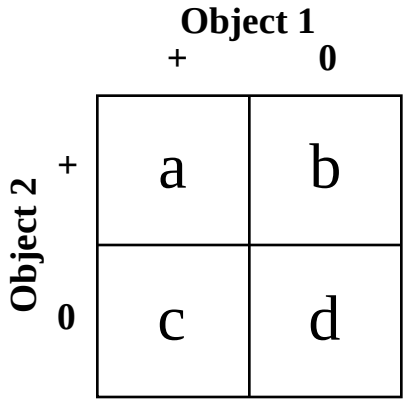
Figure 10.3 Identification of the structure of five objects, using clustering and ordination. Bottom right: the chain of primary connections is superimposed on a 2-dimensional ordination, as in Figs. 10.1 and 10.2. Top: the reduced-space ordination and the clustering results are compared to the resemblance matrix from which they originate. Upper right (top): a matrix of metric distances (or its complement $\mathbf{S_m} = [1 - D_m]$) is computed from the reduced-space ordination, and compared to the original similarities using matrix correlation ($r = 0.999$ is a rather high score). Upper right (below): a cophenetic matrix (Section 8.3) is computed from the dendrogram, and compared to the original similarities using matrix correlation ($r = 0.866$).

Some measures of resemblance

References: Sneath & Sokal (1973); Legendre & Legendre (1983, 1998)

1) Examples of binary similarity measures

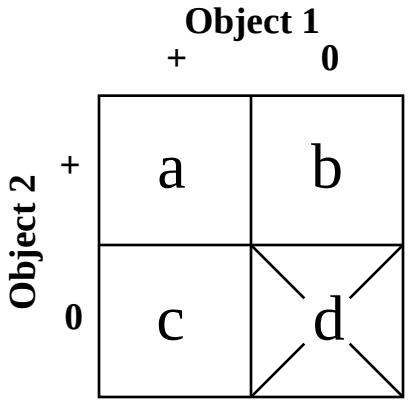
Coefficients including double-zeros (symmetrical coefficients)



Simple matching coefficient:

$$S(\mathbf{x}_1,\mathbf{x}_2) = \frac{a + d}{a + b + c + d}$$

Coefficients excluding double-zeros (asymmetrical coefficients)



Jaccard's coefficient of community:

$$S(\mathbf{x}_1,\mathbf{x}_2) = \frac{a}{a + b + c}$$

Example:	**Binary variables** (n = 7)							
Object $\mathbf{x}_1$	1	0	0	1	1	0	1	
Object $\mathbf{x}_2$	0	0	1	1	1	0	0	
Agreement		1		1	1	1		Sum = 4
Positive agreement				1	1			Sum = 2
Presence	1		1	1	1		1	Sum = 5

$S(\mathbf{x}_1,\mathbf{x}_2) = 4/7 = 0.57$

$S(\mathbf{x}_1,\mathbf{x}_2) = 2/5 = 0.40$

2) Examples of quantitative measures

Distance not designed for shared species
No upper limit

Euclidean distance (preserved in principal component analysis):

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

Distance designed for shared species
Bounded in the interval [0, 1]

Bray-Curtis distance:

$$D(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{2W}{A + B}$$

Example:	**Species abundances**					A	B	W
Object $\mathbf{x}_1$	7	3	4	5	1	20		
Object $\mathbf{x}_2$	2	4	7	6	3		22	
Minima	2	3	4	5	1			15
Differences	5	1	3	1	2			

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{5^2 + 1^2 + 3^2 + 1^2 + 2^2} = 6.325 \qquad D(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{2 \times 15}{20 + 22} = 0.286$$

Distance designed for shared species
No upper limit

Chi-square (χ^2) distance (preserved in correspondence analysis):

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j} / y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

Example:

$$\begin{bmatrix} y_{i+} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 7 & 3 & 4 & 5 & 1 \\ 2 & 4 & 7 & 6 & 3 \\ 12 & 8 & 5 & 14 & 6 \end{bmatrix} \begin{bmatrix} 20 \\ 22 \\ 45 \end{bmatrix} \rightarrow \begin{bmatrix} \frac{y_{ij}}{y_{i+}} \end{bmatrix} = \begin{bmatrix} 0.350 & 0.150 & 0.200 & 0.250 & 0.050 \\ 0.091 & 0.182 & 0.318 & 0.273 & 0.136 \\ 0.267 & 0.178 & 0.111 & 0.311 & 0.133 \end{bmatrix}$$

$$\begin{bmatrix} y_{+j} \end{bmatrix} = \begin{bmatrix} 21 & 15 & 16 & 25 & 10 \end{bmatrix} \quad 87$$

$D(\mathbf{x}_1, \mathbf{x}_2) = 0.653$

Fig. 1 – Species abundance paradox data, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 1 and 2, which have no species in common, is smaller than that between sites 1 and 3 which share species 2 and 3; this example shows that the Euclidean distance is not appropriate for species abundance data. With the other coefficients listed below, the distance between sites 1 and 2 is larger than that between sites 1 and 3; furthermore, the distance between sites 1 and 2 is the same as between sites 2 and 3.

Species abundance paradox data
(3 sites, 3 species)

	Species 1	Species 2	Species 3
Site 1	0	1	1
Site 2	1	0	0
Site 3	0	4	8

$$D_{\text{Euclidean}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad \mathbf{D} = \begin{bmatrix} 0.0000 & 1.7321 & 7.6158 \\ 1.7321 & 0.0000 & 9.0000 \\ 7.6158 & 9.0000 & 0.0000 \end{bmatrix}$$

$$D_{\text{Bray-Curtis}}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} = 1 - \frac{2W}{A+B} \qquad \mathbf{D} = \begin{bmatrix} 0.0000 & 1.0000 & 0.7143 \\ 1.0000 & 0.0000 & 1.0000 \\ 0.7143 & 1.0000 & 0.0000 \end{bmatrix}$$

$$D_{\text{chord}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}} \right)^2} \qquad \mathbf{D} = \begin{bmatrix} 0.0000 & 1.4142 & 0.3204 \\ 1.4142 & 0.0000 & 1.4142 \\ 0.3204 & 1.4142 & 0.0000 \end{bmatrix}$$

$$D_{\chi^2\text{metric}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad \mathbf{D} = \begin{bmatrix} 0.0000 & 1.0382 & 0.0930 \\ 1.0382 & 0.0000 & 1.0352 \\ 0.0930 & 1.0352 & 0.0000 \end{bmatrix}$$

$$D_{\chi^2\text{distance}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j} / y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad \mathbf{D} = \begin{bmatrix} 0.0000 & 4.0208 & 0.3600 \\ 4.0208 & 0.0000 & 4.0092 \\ 0.3600 & 4.0092 & 0.0000 \end{bmatrix}$$

$$D_{\text{Hellinger}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2} \qquad \mathbf{D} = \begin{bmatrix} 0.0000 & 1.4142 & 0.1697 \\ 1.4142 & 0.0000 & 1.4142 \\ 0.1697 & 1.4142 & 0.0000 \end{bmatrix}$$

Another useful coefficient is Gower's similarity coefficient (next page). It accepts mixtures of quantitative and binary variables.

3) Examples of coefficients for the *R* mode of analysis (comparison of variables):

Quantitative data: Pearson's *r* correlation coefficient.

Quantitative or semi-quantitative data: Spearman's *r* correlation.

Qualitative (nominal) data: chi-square and derived forms
such as the symmetric and asymmetric uncertainty coefficients, Rajski's metric, Pearson's and Tschuproff's contingency coefficients, etc.

Gower (1971a) proposed a general coefficient of similarity which can combine different types of descriptors and process each one according to its own mathematical type. Although the description of this coefficient may seem a bit complex, it can be easily translated into a short computer program. The coefficient initially takes the following form (see also the final form, eq. 7.20):

$$S_{15}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12j}$$

Partial similarity

The similarity between two objects is the average, over the p descriptors, of the similarities calculated for all descriptors. For each descriptor j, the *partial similarity value* s_{12j} between objects $\mathbf{x}_1$ and $\mathbf{x}_2$ is computed as follows.

• For *binary* descriptors, $s_j = 1$ (agreement) or 0 (disagreement). Gower proposed two forms for this coefficient. The form used here is symmetrical, giving $s_j = 1$ to double-zeros. The other form, used in Gower's asymmetrical coefficient S_{19} (Subsection 4), gives $s_j = 0$ to double-zeros.

• *Qualitative* and *semiquantitative* descriptors are treated following the simple matching rule stated above: $s_j = 1$ when there is agreement and $s_j = 0$ when there is disagreement. Double-zeros are treated as in the previous paragraph.

• *Quantitative* descriptors (real numbers) are treated in an interesting way. For each descriptor, one first computes the difference between the states of the two objects $|y_{1j} - y_{2j}|$, as in the case of distance coefficients belonging to the Minkowski metric group (Section 7.4). This value is then divided by the largest difference (R_j) found for this descriptor across all sites in the study — or, if one prefers, in a reference population*. Since this ratio is actually a normalized distance, it is subtracted from 1 to transform it into a similarity:

$$s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j]$$

Missing values

Kronecker delta

Gower's coefficient may be programmed to include an additional element of flexibility: no comparison is computed for descriptors where information is *missing* for one or the other object. This is obtained by a value w_j, called *Kronecker's delta*, describing the presence or absence of information: $w_j = 0$ when the information about y_j is missing for one or the other object, or both; $w_j = 1$ when information is present for both objects. The final form of *Gower's coefficient* is the following:

$$S_{15}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \qquad \textbf{(7.20)}$$

Coefficient S_{15} produces similarity values between 0 and 1 (maximum similarity). One last touch of complexity, which was not suggested in Gower's paper but is added here, provides weighting to the various descriptors. Instead of *0 or 1*, one can assign to w_j a value *between 0 and 1* corresponding to the weight one wishes each descriptor to have in the analysis. Giving a weight of 0 to a descriptor is equivalent to removing it from the analysis. A missing value automatically changes the weight w_j to 0.

* In most applications, the largest difference R_j is calculated for the data table under study. In epidemiological studies, for example, one may proceed to the analysis of a subset of a much larger data base. To insure consistency of the results in all the partial studies, it is recommended to calculate the largest differences (the "range" statistic of data bases) observed throughout the whole data base for each descriptor j and use these as values R_j when computing S_{15} or S_{19}.

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

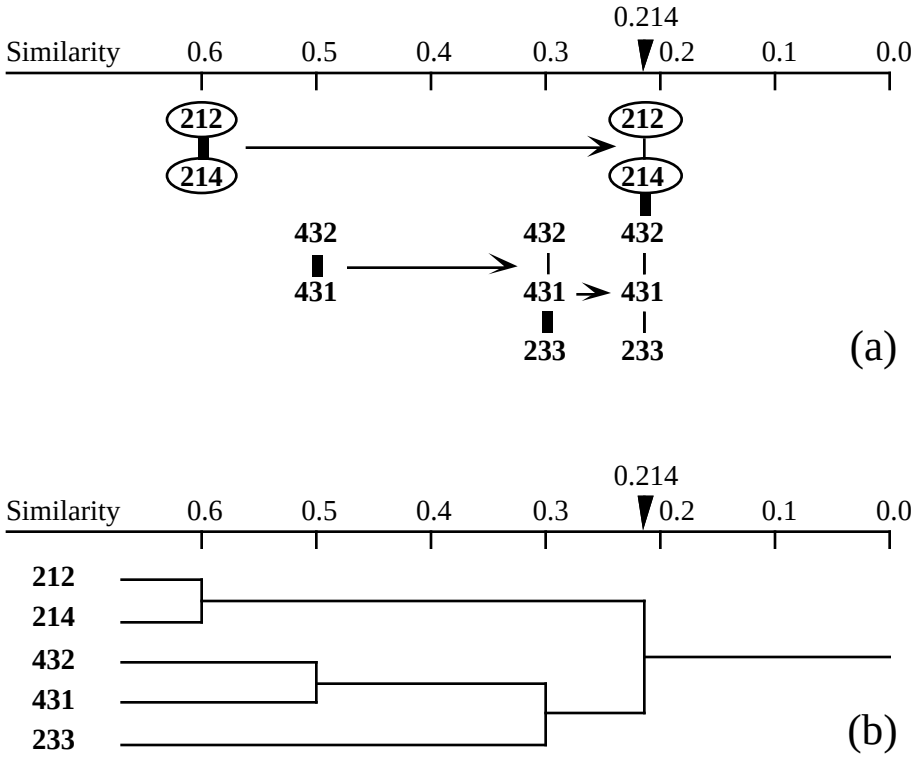


Figure 8.2 Illustrations of single linkage agglomerative clustering for the ponds of the example. (a) Connected subgraphs: groups of objects are formed as the similarity level is relaxed from left to right. Only the similarity levels where clusters are modified by addition of objects are represented. 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.

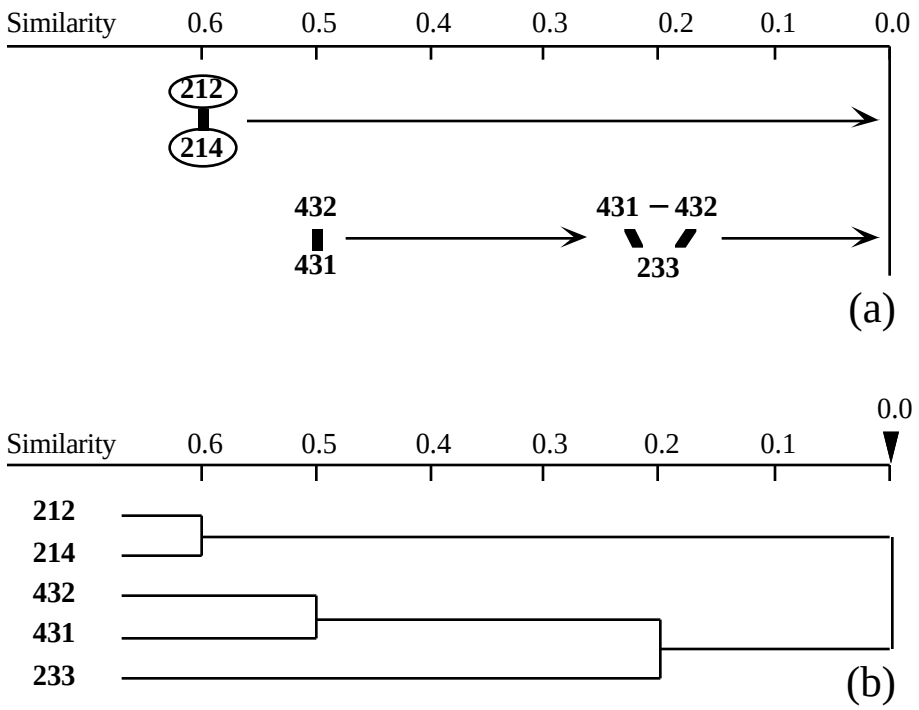


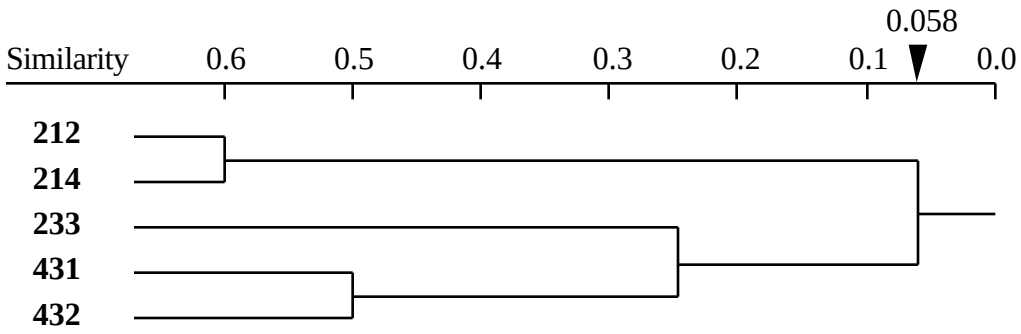
Figure 8.3 Complete linkage clustering of the ponds of Ecological application 8.2. Symbols as in Fig. 8.2.

)

Unweighted arithmetic average clustering (UPGMA)

Average the similarities between all members of the two clusters about to fuse, giving equal weight **to all original similarities**.

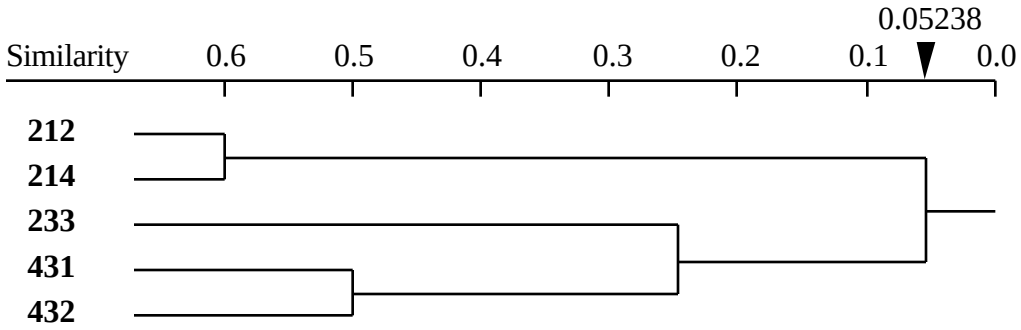
	212	**214**	**233**	**431**	**432**
212	-----				
214	**0,600**	-----			
233	0,000	0,071	-----		
431	0,000	0,063	0,300	-----	
432	0,000	0,214	0,200	0,500	-----
212-214		-----			
233		0,0355	-----		
431		0,0315	0,300	-----	
432		0,1070	0,200	**0,500**	-----
212-214		-----			
233		0,0355	-----		
431-432		0,06925	**0,250**	-----	
212-214		-----			
233-431-432		**0,058**	-----		



Weighted arithmetic average clustering (WPGMA)

Average the similarities between all the members of the two clusters about to fuse, giving the same weight **to each branch.**

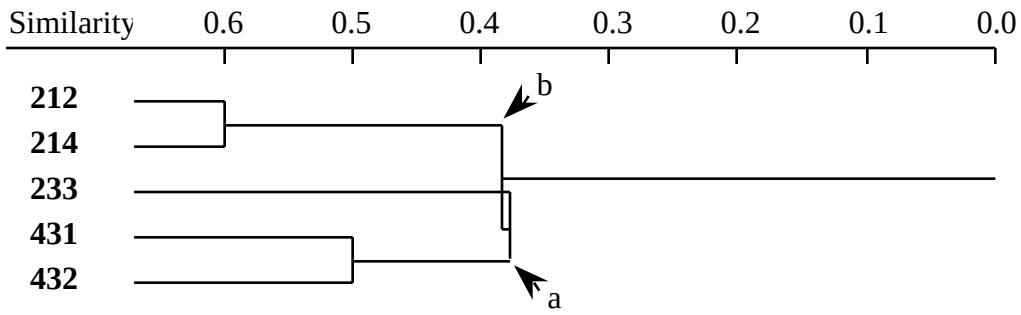
	212	**214**	**233**	**431**	**432**
212	-----				
214	**0,600**	-----			
233	0,000	0,071	-----		
431	0,000	0,063	0,300	-----	
432	0,000	0,214	0,200	0,500	-----
212-214		-----			
233		0,0355	-----		
431		0,0315	0,300	-----	
432		0,1070	0,200	**0,500**	-----
212-214		-----			
233		0,0355	-----		
431-432		0,06925	**0,250**	-----	
212-214		-----			
233-431-432		**0,05238**	-----		



Unweighted centroid clustering (UPGMC)

Find the group centroids, giving equal weight **to all original similarities**.

	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	-----
212-214		-----			
233		0,1355	-----		
431		0,1315	0,300	-----	
432		0,2070	0,200	**0,500**	-----
212-214		-----			
233		0,1355	-----		
431-432		0,29425	**0,375** (=a)	-----	
212-214		-----			
233-431-432		**0,3802** (=b)	-----		



As this example shows, reversals may occur when fusion similarity b > c.

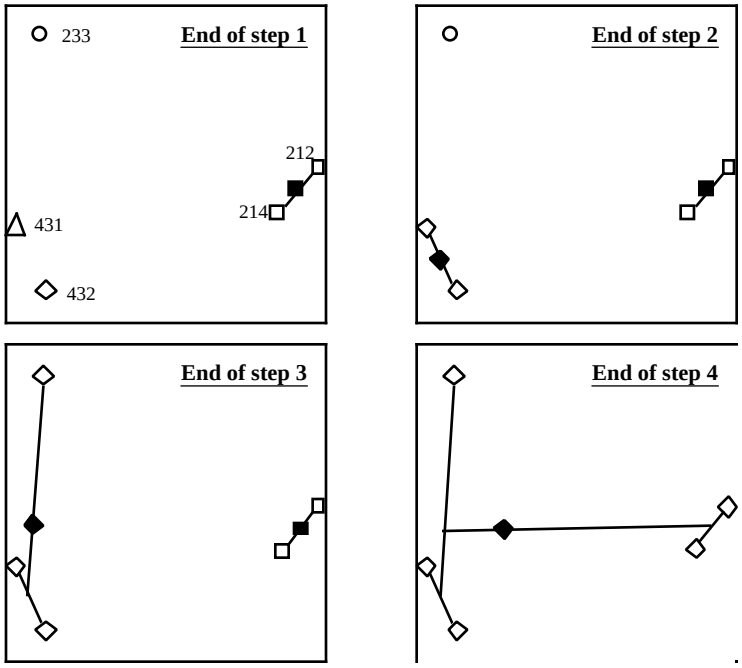
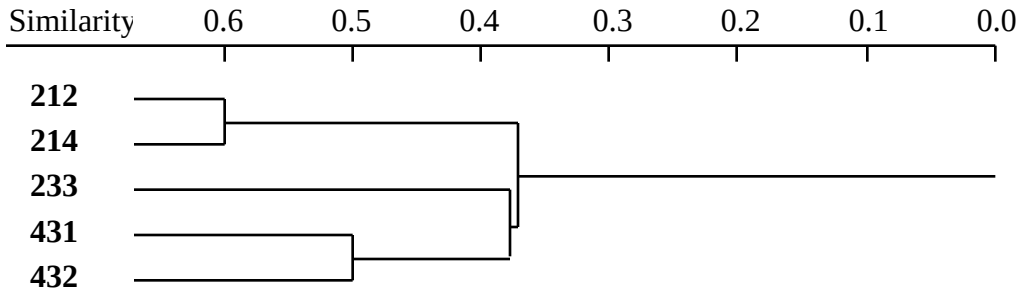


Figure 8.8 The four UPGMC clustering steps of Fig. 8.7 are drawn in A-space. Objects are represented by open symbols and centroids by dark symbols; object identifiers are shown in the first panel only. Distinct clusters are represented by different symbols. The first two principal coordinates, represented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

Weighted centroid clustering (WPGMC)

Find the group centroids, giving the same weight **to each branch**.

	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	-----
212-214		-----			
233		0,1355	-----		
431		0,1315	0,300	-----	
432		0,2070	0,200	**0,500**	-----
212-214		-----			
233		0,1355	-----		
431-432		0,29425	**0,375**	-----	
212-214		-----			
233-431-432		**0,37113**	-----		



Reversals may also occur with this method.

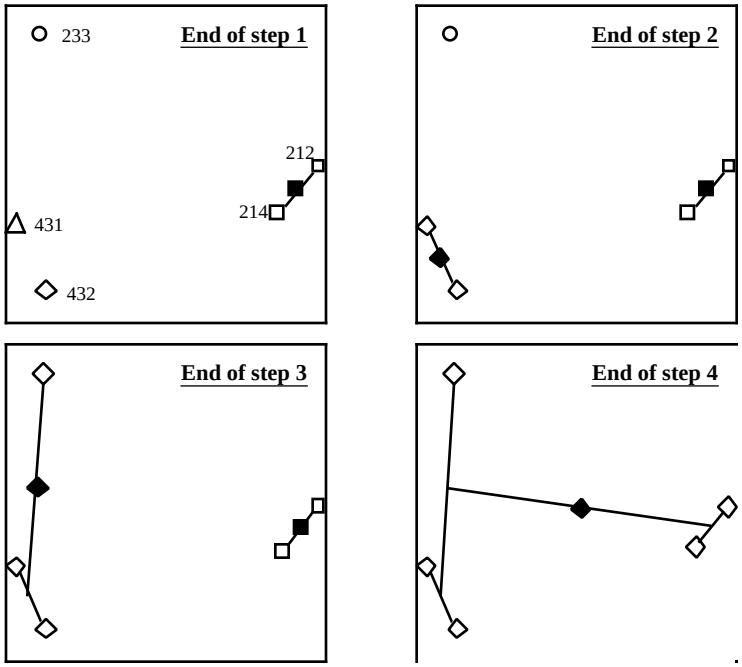


Figure 8.10 The four WPGMC clustering steps of Fig. 8.9 are drawn in A-space. Objects are represented by open symbols and centroids by dark symbols; object identifiers are shown in the first panel only. Distinct clusters are represented by different symbols. The first two principal coordinates, represented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

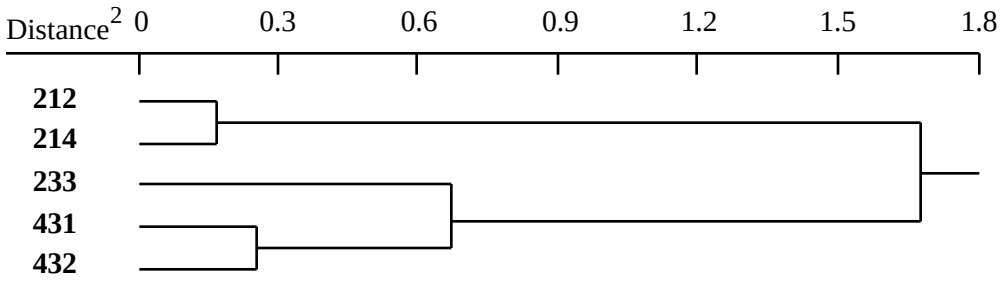
Ward's minimum-variance method

Merge the objects or clusters that minimize the sum of squared distances to the cluster centroids. The computations can be done using the general matrix updating algorithm, starting with the matrix of **squared** distances.

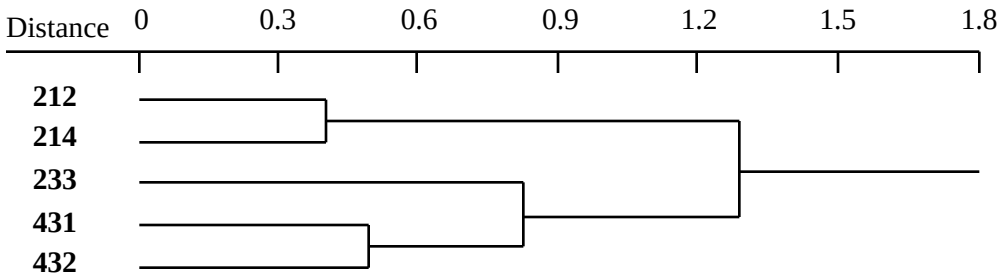
Distances^2	**212**	**214**	**233**	**431**	**432**
212	-----				
214	**0.16000**	-----			
233	1.00000	0.86304	-----		
431	1.00000	0.87797	0.49000	-----	
432	1.00000	0.61780	0.64000	0.25000	-----
212-214		-----			
233		1.18869	-----		
431		1.19865	0.49000	-----	
432		1.02520	0.64000	**0.25000**	-----
212-214		-----			
233		1.18869	-----		
431-432		1.54288	**0.67000**	-----	
212-214		-----			
233-431-432		**1.67952**	-----		

Dendrograms for Ward's results can be represented with a variety of scales. All of them have the same topology, though.

1) Squared fusion distances (from fusion table above; Jain & Dubes, 1988):



2) Fusion distances (i.e., the square roots of the squared distances above: Legendre & Vaudor, 1991). This solution is especially suitable when one wants to compare the fusion distances to the original distances (Shepard's diagram, 'cophenetic' correlation, etc.)



3) The *total error sums of squares* (TESS; ESS in Everitt, 1980; E_k^2 in Jain & Dubes, 1988) at the various fusion levels, calculated as follows:

TESS = $\sum_k e_k^2$ over the various clusters k, where

e_k^2 = sum of squared dist. of objects in cluster k to their common centroid

$e_k^2 = \sum_{dimensions\ j} \sum_i [y_{ijk} - m_{jk}]^2$

m_{jk} representing the mean value of the objects along dimension j in cluster k, while the y_{ik} are the raw coordinates of the members (i) of that cluster. In our example, TESS could be calculated from the coordinates of the pools in Euclidean space:

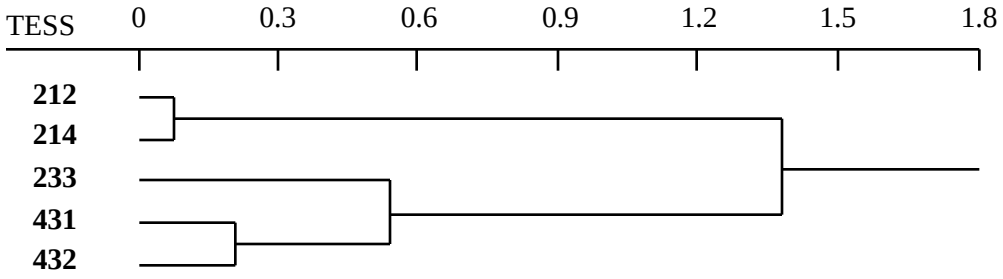
212	0.56741	0.05762	-0.16120	-0.06918
214	0.43595	-0.08026	0.15288	0.08914
233	-0.30433	0.46980	0.09632	-0.00713
431	-0.40400	-0.12248	-0.25849	0.05106
432	-0.29503	-0.32468	0.17049	-0.06389

Reconstructed data: principal coordinates (PCoA) computed from the distance matrix.

Alternatively, the within-cluster sums of squares e_k^2 may be computed as:

$$e_k^2 = \sum_i (D^2)/n_k$$

where the D^2's are the squared distances among objects in cluster k and n_k is the number of objects in that cluster. This value may be computed directly from the matrix of squared distances (top, two pages back).



4) The SAS package recommends using either the sum of within-cluster sums of squares (TESS above) divided by the total sum of squares ($1–R^2$), to obtain proportions of variance (SAS User's Guide: Statistics, p. 267), or else the *semipartial R-squared* (R^2) which is the among-cluster sum of squares divided by the total sum of squares (SAS User's Guide: Statistics, p. 281).

Computing TESS for dendrogram on previous page

212-214	$\sum_i (D^2)/n_k$	0.16/2	= 0.08
212-214 431-432	$\sum_k \sum_i (D^2)/n_k$	0.16/2 + 0.25/2	= 0.205
212-214 233-431-432	$\sum_k \sum_i (D^2)/n_k$	0.16/2 + 1.38/3	= 0.54
5 objects	$\sum_i (D^2)/n_k$	6.8988/5	= 1.38

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).

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. 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:

$$\mathbf{i}, \mathbf{g}) = (1 - \alpha_h - \alpha_i - \beta) + \alpha_h S(\mathbf{h}, \mathbf{g}) + \alpha_i S(\mathbf{i}, \mathbf{g}) + \beta S(\mathbf{h}, \mathbf{i}) - \gamma |S(\mathbf{h}, \mathbf{g}) - S(\mathbf{i}, \quad \textbf{(8.11)}$$

When using distances, the combinatorial equation becomes:

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

Clustering proceeds in the same way for all combinatorial agglomerative methods. Table 8.8 gives the values of the four parameters for the most commonly used combinatorial agglomerative clustering strategies.

Ultrametric property

Any dendrogram can be uniquely represented by a matrix where the similarity (or distance) for a pair of objects is given by the similarity (or distance) level where these objects become members of the same group in the dendrogram. Consider the UPGMA clustering results presented above. The clustering levels lead to the following similarity and distance matrices:

Clustering

S	212	214	233	431	432
212	------				
214	0.600	------			
233	0.058	0.058	------		
431	0.058	0.058	0.250	------	
432	0.058	0.058	0.250	0.500	------

Clustering

D=1-**S**	212	214	233	431	432
212	------				
214	0.400	------			
233	0.942	0.942	------		
431	0.942	0.942	0.750	------	
432	0.942	0.942	0.750	0.500	------

Such a matrix is often called a *cophenetic matrix* (Sokal & Rohlf, 1962; Legendre & Legendre, 1983; Jain & Dubes, 1988; etc.).

If there are no 'reversals' in the clustering (below), the classification has the following *ultrametric property*, and the matrix may be called *ultrametric*:

$$D(a,b) \leq \text{Max}\,[D(a,c), D(b,c)]$$

or, in terms of similarities:

$$S(a,b) \geq \text{Min}\,[S(a,c), S(b,c)]$$

The property can be verified for all triplets of similarities or distances in the above matrices summarizing our UPGMA clustering results.

With a *partition* of the data (as in the *k*-means method, below), groups of objects are obtained that are not related through a dendrogram. A cophenetic matrix may nevertheless be calculated. Consider two groups of objects (1, 2) and (3, 4, 5). The cophenetic matrices would be the following:

Partition

S	1	2	3	4	5
1	-----				
2	1	-----			
3	0	0	-----		
4	0	0	1	-----	
5	0	0	1	1	-----

Partition

D=1-**S**	1	2	3	4	5
1	-----				
2	0	-----			
3	1	1	-----		
4	1	1	0	-----	
5	1	1	0	0	-----

Reversals

Reversals may occasionally occur in the clustering structure when using UPGMC or WPGMC, or with some unusual combinations of the parameters of the general agglomerative model of Lance & Williams. When this happens, the cophenetic matrix violates the ultrametric property.

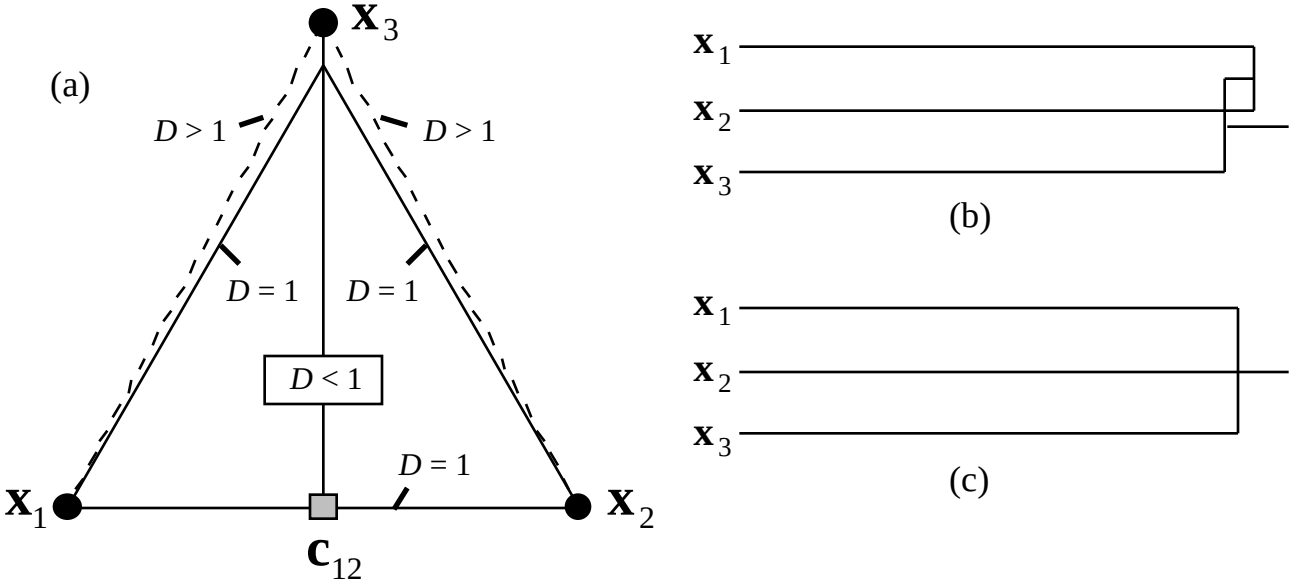


Figure — A reversal may occur in situations such as (a), where $\mathbf{x}_1$ and $\mathbf{x}_2$ cluster first because they represent the closest pair, although the distance from $\mathbf{x}_3$ to the centroid $\mathbf{c}_{12}$ is smaller then the distance from $\mathbf{x}_1$ to $\mathbf{x}_2$. (b) The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

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

$$d(\mathbf{x}_1 \quad \mathbf{x}_2 , \mathbf{x}_3) \geq d(\mathbf{x}_1 , \mathbf{x}_2)$$

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:

$$\alpha_h + \alpha_i + \beta \geq 1$$

and

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

(Milligan, 1979; Jain & Dubes, 1988: 85). Departures from ultrametricity are never important in practice. As an example, a reversal was produced in our UPGMC run (above).

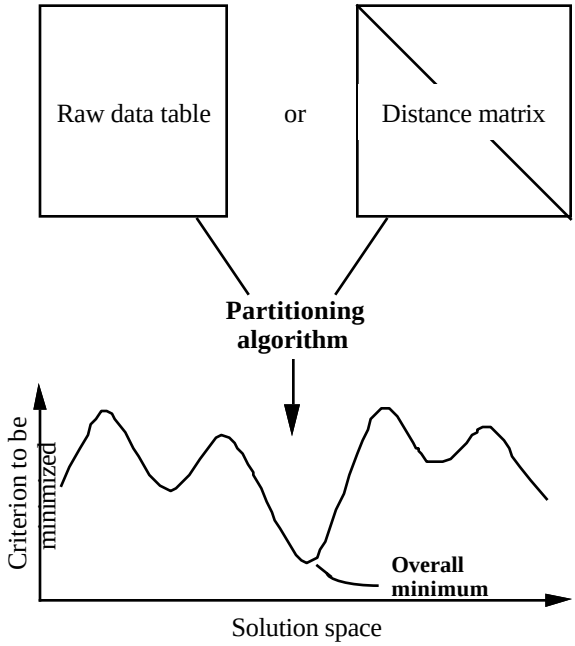


Figure 8.17 *K*-means algorithms search the space of solutions, trying to find the overall minimum (arrow) of the objective criterion to be minimized, while avoiding local minima (troughs).

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.

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*.

Overall minimum

Several solutions may be used to help a *K*-means algorithm converge towards the overall minimum of the objective criterion E^2_K. 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.

• Provide an initial configuration corresponding to the result of a hierarchical clustering, obtained from a space-conserving method (Table 8.8).

• 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.

• Attribute the objects at random to the various groups. Find a solution and note the E^2_K value. Repeat the procedure a number of times (for example, 100 times), starting every time from a different random configuration.

_K_-means: Objective functions
References: MacQueen (1967); Anderberg (1973)

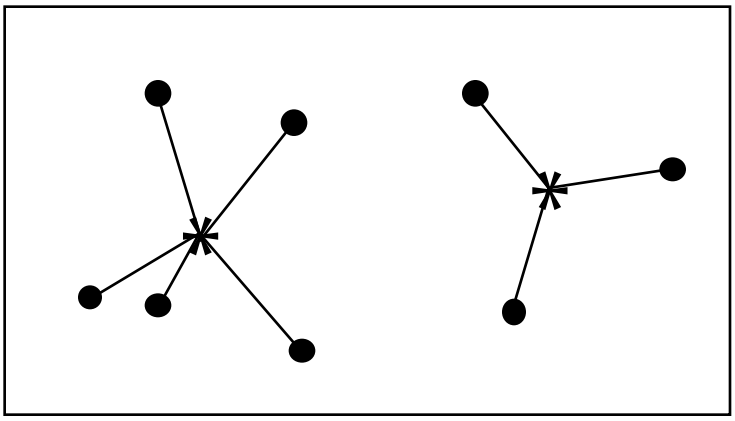
The objective functions to be minimized are the same as in Ward's method:

On raw quantitative (metric) data, one minimizes the total error sums of squares (TESS):

TESS = $\sum_k e_k^2$ over clusters k, where

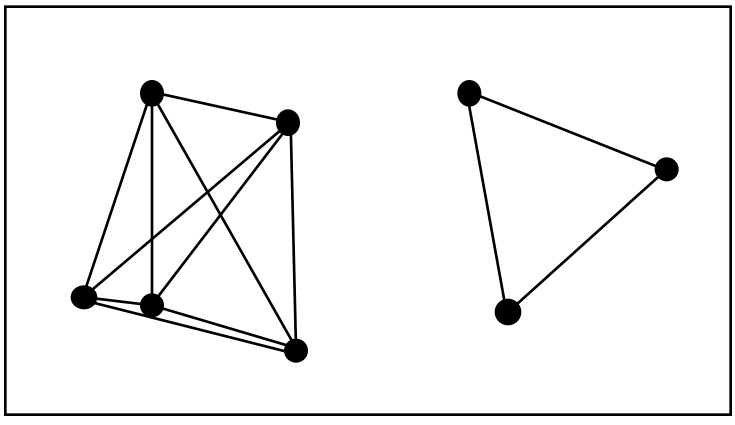
$e_k^2 = \sum_{dimensions\ j} \sum_i [y_{ijk} - m_{jk}]^2$

m_{jk} representing the mean value of the points i for dimension j in cluster k.



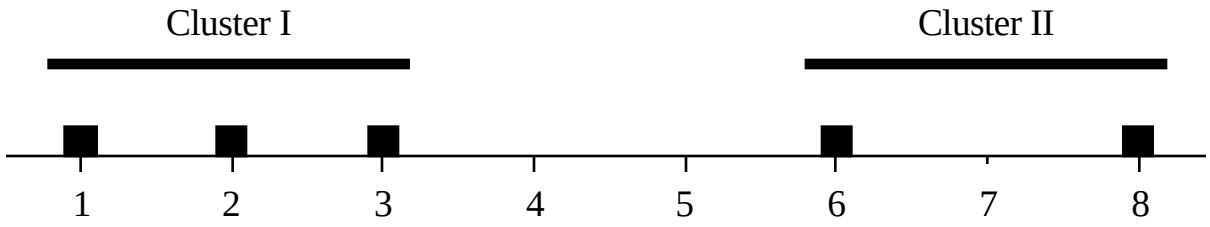
On a distance matrix, one minimizes TESS, the sum, over all clusters (k), of the sums of the squared distances between the members (i) of each cluster and their centroids (same formula as above); TESS is also equal to the sum of the mean squared within-group distances:

TESS = $\sum_k \sum_i (D^2)/n_k$ where n_k is the number of *objects* in cluster k



K-means: Simple examples

1) Divide the following 5 points (1 dimension) in two clusters:

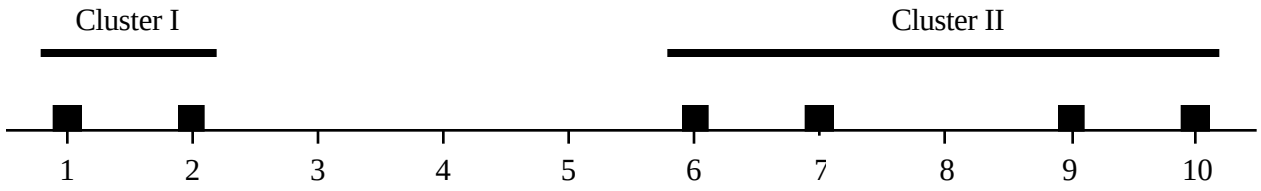


$e_I^2 = (1^2 + 0^2 + 1^2) = 2$ $\quad\quad$ $\Sigma(D^2)/3 = (2^2 + 1^2 + 1^2)/3 = 2$

$e_{II}^2 = (1^2 + 1^2) = 2$ $\quad\quad$ $\Sigma(D^2)/2 = 2^2/2 = 2$

TESS = 4 $\quad\quad$ TESS = 4

2) Divide the following 6 points (1 dimension) in two clusters



<u>"Best" solution</u>: (1, 2) (6, 7, 9, 10)

$e_I^2 = (0.5^2 + 0.5^2) = 0.5$ $\quad\quad$ $\Sigma(D^2)/2 = 1^2/2 = 0.5$

$e_{II}^2 = (2^2 + 1^2 + 1^2 + 2^2) = 10.0$ $\quad\quad$ $\Sigma(D^2)/4 = (1^2+3^2+4^2+2^2+3^2+1^2)/4 = 10.0$

TESS = 10.5 $\quad\quad$ TESS = 10.5

<u>Another solution</u>: (1, 2, 6, 7) (9, 10)

$e_I^2 = (3^2 + 2^2 + 2^2 + 3^2) = 26.0$ $\quad\quad$ $\Sigma(D^2)/4 = (1^2+5^2+6^2+4^2+5^2+1^2)/4 = 26.0$

$e_{II}^2 = (0.5^2 + 0.5^2) = 0.5$ $\quad\quad$ $\Sigma(D^2)/2 = 1^2/2 = 0.5$

TESS = 26.5 $\quad\quad$ TESS = 26.5

Cophenetic correlations and other measures of adjustment

The Pearson correlation between the values in the cophenetic matrix and those in the original resemblance matrix (excluding the values on the diagonal) is called *cophenetic correlation* (Sokal & Rohlf, 1962), *matrix correlation* (Sneath & Sokal, 1973) or *standardized Mantel* (1967) *statistic.* It measures to what extent the clustering results correspond to the original resemblance matrix; if the clustering perfectly rendered the coefficients in the original matrix, the cophenetic correlation would be 1.

Cophenetic correlations cannot be tested for significance, since the cophenetic matrix is not independent of the original similarity matrix; one comes from the other through the clustering algorithm. To test this correlation, one would have to pretend that the two matrices are independent from one another under H_0; in other words, that the clustering algorithm is likely to have a null efficiency… The similarity between hierarchical classifications obtained from different data sets and measured by matrix correlation or other measures of consensus (Rohlf, 1974, 1982) may be tested for significance, though (Lapointe & Legendre 1990, 1991, 1992a, 1992b).

Correlations take values in the interval [-1,1]. The cophenetic correlation is expected to be positive if the original similarities are compared to cophenetic similarities (or distances to distances), and negative if similarities are compared to distances. The higher the absolute value of the cophenetic correlation is, the better the adjustment.

Other measures of adjustment have been proposed. For instance, Gower's (1983) distance is the sum of the squared differences between the values in the cophenetic similarity matrix and in the original similarity matrix. That measure of adjustment takes values in the interval [0, ∞]. The smaller the value of Gower 's distance, the better the adjustment is. As in the case of the cophenetic correlations, this measure simply has a comparative value among clustering results obtained for the same original similarity matrix. This measure, also called *stress 1* (Kendall, 1938), is used as a measure of goodness-of-fit in nonmetric multidimensional scaling (MDS).

Following are three measures of adjustment for UPGMA clustering of the 5 pools:

Pearson's *r* cophenetic correlation = 0.95111
Kendall's tau b cophenetic correlation = 0.77364
Gower's distance = 0.03962

Shepard-like diagrams

A *Shepard diagram* is a scatter plot of distances in a space of reduced dimension, obtained by ordination methods, compared to distances in the original distance matrix. This type of diagram has been proposed by R. N. Shepard (1962) in the paper where he first described nonmetric multidimensional scaling (MDS). *Shepard-like diagrams* can be constructed to compare the similarities (or distances) of the cophenetic matrix to the similarities (or distances) of the original resemblance matrix. Such a plot may help decide between parametric and nonparametric cophenetic correlation coefficients. If the relationship between original and cophenetic similarities is curvilinear in the Shepard-like diagram, a nonparametric correlation coefficient should be used.

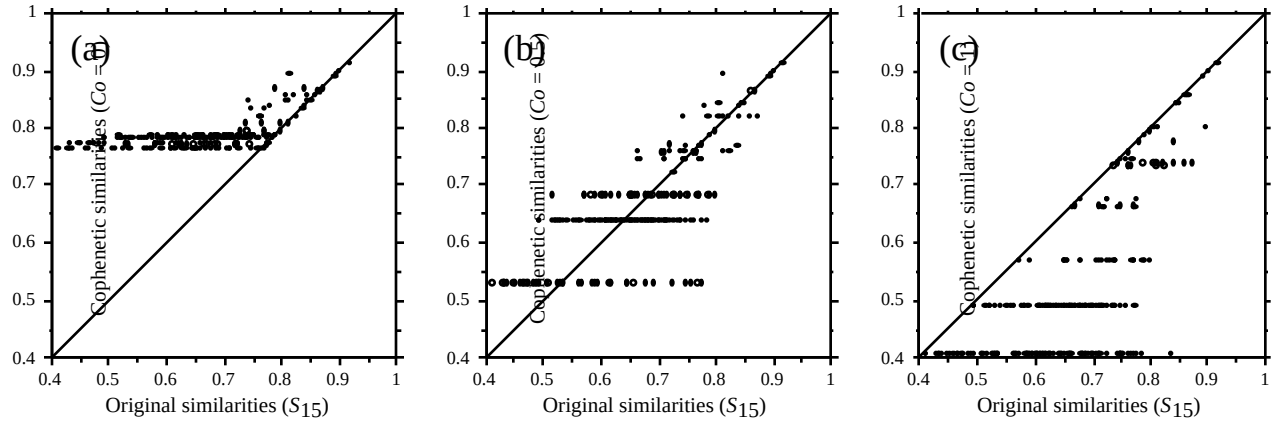


Figure — Shepard-like diagrams comparing cophenetic similarities to original similarities for 21 lakes clustered using (a) single linkage (*Co* = 0, cophenetic *r* = 0.64, τ = 0.45), (b) proportional link linkage (*Co* = 0.5, cophenetic *r* = 0.75, τ = 0.58), and (c) complete linkage clustering (*Co* = 1, cophenetic *r* = 0.68, τ = 0.51). There are 210 points (i.e. 210 similarity pairs) in each graph. The diagonal lines serve as visual references.

The Figure also helps understand the space-contraction effect of single linkage clustering, in which cophenetic similarities are always larger than or equal to the original similarities; the space-conservation effect of intermediate linkage clustering with connectedness values around *Co* = 0.5; and the space-dilation effect of complete linkage clustering, in which cophenetic similarities can never exceed the original similarities. There are (*n* – 1) clustering levels in a dendrogram. This limits to (*n* – 1) the number of different values that can be found in a cophenetic matrix and, hence, along the ordinate of a Shepard-like diagram. This is why points form horizontal bands in the Figure.

Mathematical models behind clustering methods

Model: Geometric

- Unweighted centroid clustering (UPGMC)
- Weighted centroid clustering (WPGMC)
- Dissimilarity analysis (Macnaughton-Smith *et al.*, 1964)
- Gravitational

Model: Arithmetic average

- Unweighted arithmetic average (UPGMA; "average linkage" in SAS, SYSTAT or SPSS)
- Weighted arithmetic average (WPGMA)
- Flexible clustering

Model: Graph theory

- Single linkage
- Proportional-link linkage
- Complete linkage

Model: Probabilistic

- Clifford & Goodall (1967)

Model: Statistical

- Minimum variance (Ward, 1963)
- *K*-means (MacQueen, 1967; Anderberg, 1973)
- Maximum-likelihood hierarchical clustering for distribution mixtures
- Nonparametric probability density estimates ("density linkage" in SAS)

Model: Information theory

- Edwards & Cavalli-Sforza (1965)
- Information analysis
- Association analysis

Model: Fuzzy sets [Review paper by Bezdek, 1987]

Model: ???

- Flexible clustering (Lance & Williams, 1966, 1967)

Etc.

References

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York. xiii + 35p.

Bezdek, J. C. 1987. Some non-standard clustering algorithms. Pp. 225-287 *in*: P. Legendre & L. Legendre [eds.] Developments in numerical ecology. NATO ASI Series, Vol. G 14. Springer-Verlag, Berlin.

Clifford, H. T. & D. W. Goodall. 1967. A numerical contribution to the classification of the Poaceae. Aust. J. Bot. 15: 499-519.

Day, W. H. E. 1986. Foreword: Comparison and consensus of classifications. J. Classif., 3: 183-185.

Edwards, A. W. F. & L. L. Cavalli-Sforza. 1965. A method for cluster analysis. Biometrics 21: 362-375.

Everitt, B. 1980. Cluster analysis, 2nd edition. Halsted Press, John Wiley & Sons, New York.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27: 857-871.

Gower, J. C. 1983. Comparing classifications. Pp. 137-155 *in*: Felsenstein, J. [ed.] Numerical taxonomy. NATO ASI Series, Vol. G 1. Springer-Verlag, Berlin.

Jain, A. K. & R. C. Dubes. 1988. Algorithms for clustering data. Prentice Hall, Englewood Cliffs, New Jersey.

Lance, G. N. & W. T. Williams. 1966a. A generalized sorting strategy for computer classifications. Nature (Lond.) 212: 218.

Lance, G. N. & W. T. Williams. 1967. A generalized theory of classificatory sorting strategies. I. Hierarchical systems. Computer Journal 9: 373-380.

Lapointe, F.-J. & P. Legendre. 1990. A statistical framework to test the consensus of two nested classifications. Syst. Zool. 39: 1-13.

Lapointe, F.-J. & P. Legendre. 1991. The generation of random ultrametric matrices representing dendrograms. J. Class. 8: 177-200.

Lapointe, F.-J. & P. Legendre. 1992a. A statistical framework to test the consensus among additive trees (cladograms). Syst. Biol. (in press).

Lapointe, F.-J. & P. Legendre. 1992b. Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees. Syst. Biol. (in press).

Legendre, P. & L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam.

Legendre, P. and A. Vaudor. 1991. The R Package: Multidimensional analysis, spatial analysis. Département de sciences biologiques, Université de Montréal. iv + 142 p.

Macnaughton-Smith, P., W. T. Williams, M. B. Dale & L. G. Mockett. 1964. Dissimilarity analysis: a new technique of hierarchical sub-division. Nature (Lond.) 202: 1034-1035.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Pp. 281-297 *in*: L. M. Le Cam & J. Neyman [eds.] Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1. University of California Press, Berkeley.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209-220.

Rohlf, F. J. 1974. Methods of comparing classifications. Annu. Rev. Ecol. Syst. 5: 101-113.

Rohlf, F. J. 1982. Consensus indices for comparing classifications. Math. Biosci. 59: 131-144.

SAS. 1985. SAS user's guide: statistics. SAS Institute Inc., Cary, North Carolina.

Sneath, P. H. A. & R. R. Sokal. 1973. Numerical taxonomy — The principles and practice of numerical classification. W. H. Freeman, San Francisco.

Sokal, R. R. & F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. Taxon 11: 33-40.

Ward, J. H. Jr. 1963. Hierarchical grouping to optimize an objective function. J. Amer. Stat. Ass. 58: 236-244.