Vegetatio **81**: 61–77, 1989.

New combinatorial clustering methods

János Podani
Department of Plant Taxonomy and Ecology, L. Eötvös University, Kun B. tér 2, Budapest, H-1083, Hungary; and Research Institute of Ecology and Botany, Hungarian Academy of Sciences, Vácrátót, H-2163, Hungary

Accepted 12.12.1988

Keywords: Cluster homogeneity, Dendrogram, Flexible method, Hierarchical classification, Rock grassland, Ultrametric

Abstract

Sixteen clustering methods are compatible with the general recurrence equation of combinatorial SAHN (sequential, agglomerative, hierarchical and nonoverlapping) classificatory strategies. These are subdivided into two classes: the d-SAHN methods seek for minimal between-cluster distances the h-SAHN strategies for maximal within-cluster homogeneity. The parameters and some basic features of all combinatorial methods are listed to allow comparisons between these two families of clustering procedures. Interest is centred on the h-SAHN techniques; the derivation of updating parameters is presented and the monotonicity properties are examined. Three new strategies are described, a weighted and an unweighted variant of the minimization of the increase of average distance within clusters and a homogeneity-optimizing flexible method. The performance of d- and h-SAHN techniques is compared using field data from the rock grassland communities of the Sashegy Nature Reserve, Budapest, Hungary.

Abbreviations: CP = Closest pair; RNN = Reciprocal nearest neighbor; SAHN = Sequential, agglomerative, hierarchical and nonoverlapping

Nomenclature of syntaxa follows Soó, R. 1964. Synopsis systematico-geobotanica florae vegetationisque Hungariae I. Akadémiai, Budapest.

Introduction

The sequential, agglomerative, hierarchical and nonoverlapping clustering techniques (the so-called SAHN methods, Sneath & Sokal 1973) are commonly used procedures of numerical classification in vegetation science (see Orlóci 1978; van der Maarel 1979; and Greig-Smith 1983, for review), including synsystematics (e.g., Orlóci & Stanek 1979; Mucina 1982; Moreno-Casasola & Espejel 1986). A family of these methods requires only a symmetric distance (dissimilarity, similarity, etc.) matrix **W** to be stored in computer memory during computations; the raw data may be released once this matrix has been calculated (*stored matrix* approach, Anderberg 1973). The original data are not needed because there is a *combinatorial* solution to recompute between-cluster measures using the information contained in **W** and in an array of cluster sizes. Lance &

Table 1. Parameters for combinatorial d-SAHN (1–8) and h-SAHN (9–16) clustering methods.

Clustering method	α_i	α_j	β	γ	λ_h	λ_i	λ_j
1. Single linkage (SL)	1/2	1/2	0	− 1/2	0	0	0
2. Complete linkage (CL)	1/2	1/2	0	1/2	0	0	0
3. Unweighted average (UPGMA)	$\frac{n_i}{n_h + n_i}$	$\frac{n_j}{n_h + n_j}$	0	0	0	0	0
4. Weighted average (WPGMA)	1/2	1/2	0	0	0	0	0
5. Centroid (UPGMC)	$\frac{n_i}{n_h + n_i}$	$\frac{n_j}{n_h + n_j}$	$-\frac{-n_i n_j}{(n_i + n_j)^2}$	0	0	0	0
6. Median (WPGMC)	1/2	1/2	− 1/4	0	0	0	0
7. β-Flexible (β-FLEX)	(1−β)/2	(1−β)/2	<1	0	0	0	0
8. (β, γ)-Flexible ((β, γ)-FLEX)	(1−β)/2	(1−β)/2	unrestricted	unrestricted	0	0	0
9. Minimum increase of sum of squares (MISSQ)	$\frac{n_h + n_i}{n.}$	$\frac{n_h + n_j}{n.}$	$-\frac{n_h}{n.}$	0	0	0	0
10. Minimum sum of squares of new cluster (MNSSQ)	$\frac{n_h + n_i}{n.}$	$\frac{n_h + n_j}{n.}$	$\frac{n_i + n_j}{n.}$	0	$-\frac{n_h}{n.}$	$-\frac{n_i}{n.}$	$-\frac{n_j}{n.}$
11. Minimum increase of variance (MIVAR)	$\left(\frac{n_h + n_i}{n.}\right)^2$	$\left(\frac{n_h + n_j}{n.}\right)^2$	$-\frac{n_h(n_i + n_j)}{n.^2}$	0	0	0	0
12. Minimum variance of new cluster (MNVAR)	$\left(\frac{n_h + n_i}{n.}\right)^2$	$\left(\frac{n_h + n_j}{n.}\right)^2$	$\left(\frac{n_i + n_j}{n.}\right)^2$	0	$-\left(\frac{n_h}{n.}\right)^2$	$-\left(\frac{n_i}{n.}\right)^2$	$-\left(\frac{n_j}{n.}\right)^2$

13. Minimum increase of weighted average distance (WMIDIS)	$\frac{b_{hi}}{b.}$	$\frac{b_{hj}}{b.}$	$\frac{b_{ij}}{b.} - \frac{1}{2}$	0	$-\frac{b_h + n_i n_j}{2b.}$	$-\frac{b. - 2b_i - 2n_h n_j}{4b.}$	$-\frac{b. - 2b_j - 2n_h n_i}{4b.}$
14. Minimum increase of unweighted average distance (UMIDIS)	$\frac{b_{hi}}{b.}$	$\frac{b_{hj}}{b.}$	$\frac{b_{ij}}{b.} - \frac{b_{ij}}{b_h + b_{ij}}$	0	*	*	*
15. Minimum average distance within new cluster (MNDIS)	$\frac{b_{hi}}{b.}$	$\frac{b_{hj}}{b.}$	$\frac{b_{ij}}{b.}$	0	$-\frac{b_h}{b.}$	$-\frac{b_i}{b.}$	$-\frac{b_j}{b.}$
16. λ-Flexible (λ-FLEX)	$(1-3\lambda)/3$	$(1-3\lambda)/3$	$(1-3\lambda)/3$	0	$\leqq 0$	$= \lambda_h$	$= \lambda_h$

$b_i = \binom{n_i}{2}$, n_i = number of objects in C_i, $n. = n_h + n_i + n_j$, * expression too long, see formula (14) in Appendix.

Williams (1966, 1967) presented the classical example for this approach; they suggested the recurrence formula

$$w_{h,\,ij} = \alpha_i w_{hi} + \alpha_j w_{hj} + \beta w_{ij} + \\ + \gamma \,|\, w_{hi} - w_{hj} \,| \qquad (1)$$

to update the values of **W** for the single linkage, complete linkage, group average, centroid and median (Gower 1967) clustering algorithms (Table 1). If clusters C_i and C_j are merged in a clustering cycle, then $w_{h,\,ij}$ gives the updated criterion value to be used in the next cycle for cluster $C_i \cup C_j$ with any other cluster C_h. The simple average (or weighted average linkage, WPGMA) method originally suggested by Sokal & Michener (1958) is also combinatorial. Wishart (1969) developed first the parameters for the incremental sum of squares agglomeration technique. The β-flexible strategy of Lance & Williams (1967) has no fixed parameters, the values of α_i, α_j, and β may be changed under certain conditions to provide transitions between extremely space-dilating and space-contracting algorithms (see also Sneath & Sokal 1973). Less constrained is the (β, γ)-flexible technique introduced by DuBien & Warde (1979) which in fact contains the single linkage, complete linkage, simple average and β-flexible strategies as special cases. Diday *et al.* (1982) presented the parameters for the minimization of the increase of variance, so that the number of methods formally compatible with equation (1) increased to ten.

In the meantime, it was revealed that further SAHN methods have combinatorial solutions. Jambu (1978) and Podani (1978, 1979) independently derived updating parameters for two techniques which had formerly been described by Anderberg (1973) as representatives of the *stored data* approach. These methods minimize the error sum of squares or the variance of the newly formed clusters. Podani (1978, 1979) also showed that the method of average linkage within the new group (Anderberg 1973, p. 139), in which cluster homogeneity is defined as the average of within-cluster similarities, is also combinatorial. These three methods differ from those compatible with equation (1) in that within-cluster measures are also used in recalculating between-cluster measures. The updated value, $w_{h,ij}$, is determined according to six values of **W**:

$$\begin{matrix} w_{hh} & w_{hi} & w_{hj} \\ & w_{ii} & w_{ij} \\ & & w_{jj} \end{matrix}$$

where w_{hh}, w_{ii} and w_{jj} (replaced by w_h, w_i, and w_j, for simplicity) denote either sum of squares, variance or average similarity within clusters C_h, C_i, and C_j, respectively. Jambu (1978, see also Jambu & Lebeaux 1983) suggested that the scope of the Lance-Williams formula be extended to these methods by adding three terms to equation (1) so that a more general recurrence relation is obtained:

$$w_{h,ij} = \alpha_i w_{hi} + \alpha_j w_{hj} + \beta w_{ij} + \gamma \,|\, w_{hi} - w_{hj} \,| + + \lambda_h w_h + \lambda_i w_i + \lambda_j w_j \qquad (2)$$

(see Table 1 for parameters of the above three methods).

This paper shows that three more SAHN techniques fit into equation (2): the minimization of the increase of weighted and unweighted average distance within clusters, and a new flexible strategy. A classification of the 16 known combinatorial techniques is suggested considering whether inter-cluster distances or within-cluster homogeneities are optimized during the clustering process. Those concerned with cluster homogeneity are discussed and the derivation of their updating parameters is presented. Two tables summarize basic information on the combinatorial SAHN methods compatible with the general recurrence formula given by equation (2). 23 different algorithms of these methods are applied to a phytosociological data set from grassland communities, and the resulting dendrograms are assessed by a multiple comparison method.

Combinatorial h-SAHN methods

Distance- versus homogeneity-optimizing strategies

The SAHN procedures have been the subject of intensive research for many years, but an interesting aspect recognized by Lance & Williams (1967) remains largely overlooked. Those authors distinguished among three basic types of measures used in cluster analysis; these types will serve as the starting point in this paper for a more comprehensive categorization of combinatorial SAHN methods.

For a number of SAHN techniques inter-cluster distances (or dissimilarities, similarities, etc.) are defined (the (i, j)-measures in Lance & Williams' terminology). Distances are geometrically interpretable in a Euclidean space and compatible with all methods to be discussed; therefore they will be used in the sequel unless otherwise stated. Two clusters, C_i and C_j, are fused if their distance, $d(C_i, C_j)$, is minimal in the given clustering step. One entry in **W** is defined in two ways to ensure compatibility with equation (1):

$$w_{ij} = d(C_i, C_j) \text{ or } w_{ij} = d^2(C_i, C_j)\,.$$

There are no restrictions on within-cluster homogeneity; and the fusion levels indicated in the dendrogram are between-cluster distances containing no information on within-cluster structure. Typical examples are the group average, centroid and single linkage methods.

In other SAHN procedures the fusion criterion relies on some measure of within-cluster homogeneity, even if inter-object distances are calculated first in the analysis. Using an appropriate homogeneity measure, $h(C_i)$, the analysis may proceed in two different ways. One possibility is to maximize the homogeneity of the newly formed clusters, that is,

$$w_{ij} = h(C_i \cup C_j)\,.$$

This criterion corresponds with the (i)-measures of Lance & Williams (1967) although they considered such measures to have relevance in nonhierarchical clustering only. An example is the minimization of sum of squares in new clusters, a method already mentioned in connection with the general recurrence formula (2). The other optimization procedure involves minimization of the change of homogeneity upon the fusion of two clusters, so that one entry of **W** will have the following general form,

$$w_{ij} = h(C_i \cup C_j) - \rho_i h(C_i) - \rho_j h(C_j)\,.$$

where ρ_i and ρ_j are weights specific to each procedure. Such criteria were called the (*ij*, *k*)-measures by Lance and Williams, with *k* referring to the union of C_i and C_j. The incremental sum of squares technique which amalgamates clusters so as to minimize the increase of within-cluster sum of squares is an example. The increments w_{ij} are not used directly as fusion levels in the dendrogram; a more appropriate level is $h(C_i \cup C_j)$ so that the results of the alternative homogeneity-optimizing strategies become directly comparable.

I think that the distinction between the *d*istance- and *h*omogeneity-optimizing SAHN strategies is important and facilitates the discussion of combinatorial methods. I suggest the use of abbreviations d-SAHN and h-SAHN, respectively, to cover these two main groups of procedures. Within the second category, further distinction is made between the nh-SAHN and ch-SAHN techniques depending on whether the homogeneity of *n*ew clusters or the *c*hange of homogeneity is minimized. It is noted that the use of this terminology is not restricted to combinatorial methods; there are d-SAHN methods (e.g., the *U*-statistic clustering method proposed by d'Andrade 1978) and h-SAHN methods (e.g., those utilizing information theoretic criteria to measure cluster homogeneity, Lance & Williams 1967, Sneath & Sokal 1973) which do not satisfy the recurrence relation (2).

The close relationship among the h-SAHN combinatorial methods is that the derivation of updating parameters for equation (2) follows the same logic (see Appendix). This further supports the importance of an at least technical distinction between d-SAHN and h-SAHN combinatorial procedures.

Fusion criteria in combinatorial h-SAHN methods

Four definitions of within-cluster homogeneity have been proposed in association with combinatorial SAHN methods. These are the error sum of squares (*SSQ*), variance (*VAR*), and average distance (*DIS*) or similarity (*SIM*) within clusters. The latter measure is especially important in various approaches to phytosociological classification (e.g., Popma *et al.* 1983). The clustering process may proceed in two basically different ways: 1) maximization of the homogeneity of new clusters, and 2) minimization of the decrease of homogeneity. The combination of homogeneity measures and strategy types gives rise to 6 h-SAHN methods, one of them with weighted and unweighted variants. These include both widely used and less known methods of numerical classification, as well as two procedures for which the updating parameters are presented for the first time in this paper. This section gives a summary of fusion criteria; the derivation of parameters is presented in the Appendix. The parameters are shown in Table 1 while other useful information is summarized in Table 2 which allows for comparing the basic features of d-SAHN and h-SAHN combinatorial procedures. A practical importance of the subsequent discussion is that several publications do not specify exactly the fusion criterion actually used; reference to terms such as minimum variance clustering and Ward's method is a potential source of confusion.

Optimization of new within-cluster heterogeneity

Minimization of error sum of squares (dispersion) within the new cluster (MNSSQ)

Cluster homogeneity is expressed in terms of error sum of squares calculated from pairwise distances, d_{ij}, of objects. One element of the starting matrix is $d_{ij}^2/2$. At each stage of the analysis any C_r and C_s are fused provided that

$$w_{rs} = min\{SSQ(C_i \cup C_j)\colon\; 1 \leqq i < j \leqq n\}\,,$$

(Anderberg 1973: p. 148).

Table 2. Some properties of combinatorial SAHN clustering methods.

Clustering method	Initialization of **W**	Fusion level of $C_i \cup C_j$ * (and update of w_i)	$\alpha_i + \alpha_j + \beta +$ $+ \lambda_h + \lambda_i + \lambda_j$	Monotone fusion levels		Results of CP and RNN agree
				CP	RNN	
1. SL	d_{ij}	w_{ij}	1	yes	yes	yes
2. CL	d_{ij}	w_{ij}	1	yes	yes	yes
3. UPGMA	d_{ij}	w_{ij}	1	yes	yes	yes
4. WPGMA	d_{ij}	w_{ij}	1	yes	yes	yes
5. UPGMC	d_{ij}^2	$\sqrt{w_{ij}}$	<1	no	no	no
6. WPGMC	d_{ij}^2	$\sqrt{w_{ij}}$	.75	no	no	no
7. β-FLEX	d_{ij}	w_{ij}	1	yes	yes	yes
8. (β, γ)-FLEX	d_{ij}	w_{ij}	1	yes/no	yes/no	yes/no
9. MISSQ	$d_{ij}^2/2$	$w_{ij} + w_i + w_j$	1	yes	yes	yes
10. MNSSQ	$d_{ij}^2/2$	$w_{ij}*$	1	yes	yes	yes
11. MIVAR	$d_{ij}^2/4$	$w_{ij} + \frac{n_i}{n_i + n_j} w_i + \frac{n_j}{n_i + n_j} w_j$	$\leqq 1 \leqq$	no	no	no
12. MNVAR	$d_{ij}^2/4$	$w_{ij}*$	1	yes	yes	yes
13. WMIDIS	d_{ij}	$w_{ij} + \frac{1}{2} w_i + \frac{1}{2} w_j *$	$\leqq 1 \leqq$	no	no	no
14. UMIDIS	d_{ij}	$w_{ij} + \frac{b_i}{b_i + b_j} w_i + \frac{b_j}{b_i + b_j} w_j *$	$\leqq 1 \leqq$	yes?	no	no
15. MNDIS	d_{ij}	$w_{ij}*$	1	yes	no	no
16. λ-FLEX	d_{ij}	$w_{ij}*$	1	yes	no	no

Minimization of variance within the new cluster (MNVAR)

Cluster homogeneity is measured by the average contribution of objects to the total sum of squares of the cluster (i.e., variance). One element of the starting matrix is $w_{ij} = d_{ij}^2/4$. The condition for the fusion of two clusters C_r and C_s is that the variance of the new cluster be minimal:

$$w_{rs} = min\{VAR(C_i \cup C_j)\colon\ 1 \leqq i < j \leqq n\}$$

(cf. Anderberg 1973: p. 148).

Optimization of average distance or (dis)similarity within the new cluster (MNDIS).
Originally, I suggested that cluster homogeneity be measured by the simple matching coefficient generalized to more than 2 objects (Podani 1978, 1979). In this case, homogeneity was defined as the number of agreements among objects divided by the possible number of agreements. This ratio is simply the average of all pairwise similarity coefficients within the cluster. However, the strategy equally applies to other types of similarity measures, as well as to dissimilarity and distance coefficients. If cluster homogeneity is defined as the average of pairwise similarities (SIM), then C_r and C_s are selected for fusion if

$$w_{rs} = max\{SIM(C_i \cup C_j):\ 1 \leqq i < j \leqq n\}\ .$$

For dissimilarities and distance, the criterion is

$$w_{rs} = min\{DIS(C_i \cup C_j):\ 1 \leqq i < j \leqq n\}\ .$$

This method, termed as average linkage within the new group, was considered formerly by Anderberg (1973: p. 139) as a representative of the stored matrix approach: only the distances have to be retained in computer memory during calculations even if the combinatorial algorithm is not used. However, the combinatorial procedure is much faster than the algorithm suggested by Anderberg.

Minimization of the increase of heterogeneity

Minimization of the increase of sum of squares (MISSQ)
This technique has been referred to under various and often misleading names (e.g., Ward's method, minimum variance (!) clustering, sum of squares agglomeration, and a better one: incremental sum of squares clustering) and belongs to the most widely used clustering algorithms. As clusters to be fused any C_r and C_s are chosen so that

$$w_{rs} = min\{SSQ(C_i \cup C_j) - SSQ(C_i) - SSQ(C_j):\ 1 \leqq i < j \leqq n\}\ ,$$

(see e.g., Anderberg 1973; Orlóci 1967; Wishart 1969).

Minimization of the increase of variance (MIVAR)
In this strategy clusters C_r and C_s are fused provided that

$$w_{rs} = min\{VAR(C_i \cup C_j) - \frac{n_i}{n_i + n_j} VAR(C_i) - \frac{n_j}{n_i + n_j} VAR(C_j):\ 1 \leqq i < j \leqq n\}\ ,$$

(Diday *et al.* 1982). Jambu & Lebeaux (1983) stated that the parameters for this method agree with those of MISSQ, but this is not the case. Diday *et al.* (1982, p. 89) listed first the correct parameters for this technique without showing the derivation of parameters which is presented in the Appendix.

Minimization of the increase of average within-cluster distances (MIDIS)
This is a new strategy with two alternative variants. The average within-cluster distances for clusters C_h and $C_i \cup C_j$ may be calculated in two different ways, i.e., with and without considering cluster sizes. Accepting the terminology of Sneath & Sokal (1973), these variants are termed as unweighted and weighted MIDIS, respectively, because when cluster sizes are neglected the smaller cluster receives greater weight. In this sense, these alternatives are analogous to the pair of group average and weighted average (UPGMA and WPGMA) methods as well as to the pair of the median and centroid strategies from the group of d-SAHN methods.

In the weighted case (WMIDIS), clusters C_r and C_s are amalgamated provided that

$$w_{rs} = min\{DIS(C_i \cup C_j) - \frac{1}{2} DIS(C_i) - \frac{1}{2} DIS(C_j):\ 1 \leqq i < j \leqq n\}\ .$$

The condition for the fusion of C_r and C_s in the unweighted strategy (UMIDIS) is as follows:

$$w_{rs} = min\{DIS(C_i \cup C_j) - \\ - \frac{b_i}{b_i + b_j} DIS(C_i) - \frac{b_j}{b_i + b_j} DIS(C_j): \\ 1 \leqq i < j \leqq n\},$$

with $b_i = \binom{n_i}{2}$. For similarities, *max* should replace *min* in both formulae.

Ultrametric properties of combinatorial SAHN methods

A hierarchy produced by a clustering algorithm may be described in terms of a matrix **D** in which δ_{ij} is the lowest hierarchical level at which objects i and j belong to the same cluster. If any triplet δ_{ij}, δ_{hi}, and δ_{hj} of such values satisfies the relations

$$\delta_{ij} \leqq max\{\delta_{hi}, \delta_{hj}\},$$
$$\delta_{hi} \leqq max\{\delta_{ij}, \delta_{hj}\}, \quad and$$
$$\delta_{hj} \leqq max\{\delta_{ij}, \delta_{hi}\},$$

the output values are *ultrametric distances* (cf. Johnson 1967). In this case the fusion levels monotonically increase: the fusion level of any cluster C_h and $C_i \cup C_j$ cannot be lower than that of cluster C_i with C_j. Failure to satisfy the above relationships is manifested as "reversals" in the tree diagram. The methods compatible with formula (1) will produce monotone increasing fusion levels provided that

$$\alpha_i + \alpha_j + \beta \geqq 1, \tag{3a}$$
$$\alpha_i + \alpha_j \geqq 0, \tag{3b}$$
$$\gamma \geqq -min\{\alpha_i, \alpha_j\}, \tag{3c}$$

(see Milligan 1979; Batagelj 1981). Therefore, 5 d-SAHN strategies are monotonic, the exceptions being UPGMC, WPGMC and, depending on the choice of parameters, the (β, γ)-flexible method (Table 2).

The h-SAHN methods require separate and more thoroughful scrutiny for monotonicity. First, the nh-SAHN methods are considered. Diday (1983) suggested four necessary and sufficient conditions to ensure monotonicity of methods compatible with equation (2). These conditions include 3a–c and

$$\lambda_1, \lambda_2, \lambda_3 \geqq 0. \tag{3d}$$

From Table 2 it is easily seen that neither nh-SAHN methods discussed meet the requirement expressed by condition (3d). One should observe, however, that $\alpha_i + \alpha_j + \beta + \lambda_h + \lambda_i + \lambda_j = 1$ for MNSSQ, MNVAR and MNDIS, therefore it is worth examining if the constraints

$$\alpha_i + \alpha_j + \beta + \lambda_h + \lambda_i + \lambda_j \geqq 1, \tag{4a}$$
$$\lambda_h, \lambda_i, \lambda_j \leqq 0, \tag{4b}$$
$$\alpha_i, \alpha_j, \beta \geqq 0, \quad \text{and} \tag{4c}$$
$$\gamma = 0 \tag{4d}$$

are sufficient to prove the ultrametric feature of these methods. The hierarchical levels monotonically increase if it can be shown that $w_{h,ij} \geqq w_{ij}$ for every clustering step. The proof below utilizes some elements of Milligan's (1979) proof applied to d-SAHN procedures.

Proof. The assumed constraint (4a) may be rewritten as $\beta \geqq 1 - \alpha_i - \alpha_j - \lambda_h - \lambda_i - \lambda_j$. Substituting this constraint into (2) yields

$$w_{h,ij} \geqq \alpha_i w_{hi} + \alpha_j w_{hj} + \\ + (1 - \alpha_i - \alpha_j - \lambda_h - \lambda_i - \lambda_j) w_{ij} + \\ + \lambda_h w_h + \lambda_i w_i + \lambda_j w_j.$$

After rearrangement we have

$$w_{h,ij} \geqq w_{ij} + \alpha_i(w_{hi} - w_{ij}) + \alpha_j(w_{hj} - w_{ij}) + \\ + \lambda_h(w_h - w_{ij}) + \lambda_i(w_i - w_{ij}) + \\ + \lambda_j(w_j - w_{ij}). \tag{5}$$

Since in the first clustering step w_h, w_i, $w_j = 0$, it must be that w_{ij}, w_{hi}, $w_{hj} \geqq w_i$, w_j, w_h. The clustering procedure always selects the smallest value in **W**, so that $w_{ij} \leqq w_{hj}$, w_{hi}. Since the constraints (4b–c) require that the λ-s are non-positive and that α_i, α_j cannot be negative, the last five terms in (5) must be greater than or equal to zero and may be deleted from the inequality without losing its validity, and we end up with the desired inequality:

$$w_{h,ij} \geqq w_{ij}\,. \tag{6}$$

After the fusion, w_i is set equal to w_{ij} and row and column j of **W** are masked. It is apparent that w_i will not be greater than any off-diagonal value of **W**. According to (6), the off-diagonal values of **W** cannot decrease in the subsequent steps, therefore the newly computed fusion levels are never lower than the earlier values. Thus, the monotonicity of levels holds, no matter whether *SSQ*, *VAR* or *DIS* are used as the homogeneity measure.

In the proof above it was assumed that the classical paradigmatic SAHN clustering algorithm is employed, i.e., a single fusion is performed in every clustering step (*c*losest *p*air or CP algorithm). When reciprocal nearest neighbors are fused in each cycle to accelerate the analysis (*r*eciprocal *n*earest *n*eighbor or RNN algorithm, see e.g., Anderberg 1973; Murtagh 1983; Day & Edelsbrunner 1984), the constraints (4a–d) are insufficient to ensure monotonicity and the nature of the homogeneity measure will be of primary concern. It is enough to examine whether the reducibility condition.

$$w_{h,ij} \geqq min\{w_{hi}, w_{hj}\} \text{ for all } h$$

(Bruynooghe 1978), holds for all reciprocal nearest neighbors i and j. Since this condition is satisfied for MNSSQ and MNVAR, their results (and thus their ultrametric properties) are unaffected by the choice between the CP and RNN algorithm. However, for the MNDIS criterion the reducibility condition is not satisfied as the following simple example demonstrates. Let the matrix of Euclidean distances of four objects be given by

$$\mathbf{W} = \begin{array}{cccc} 0 & 1 & 2+\varepsilon & 2+\varepsilon \\ & 0 & 2+\varepsilon & 2+\varepsilon \\ & & 0 & 2 \\ & & & 0 \end{array}$$

Object pairs 1–2 and 3–4 are reciprocal nearest neighbors if $0 < \varepsilon$, but $w_{1,34}$, $w_{2,34} = (6 + 2\varepsilon)/3 < 2 + \varepsilon$. As a consequence, the RNN algorithm does not exclude the possibility of reversals. Thus, whereas the ultrametric properties of d-SAHN methods remain the same for the CP and RNN algorithms (Gordon 1987; see Table 2), there is at least one counter-example among the h-SAHN strategies.

The ch-SAHN methods are less similar to one another in general properties than the nh-SAHN strategies. There are differences in the way of calculating the fusion levels and the sum of parameters is not a constant, except in MISSQ (Table 2). The increment of MISSQ is monotonic, because the parameters satisfy relation (3). Also, it is easy to see that $SSQ(C_i \cup C_j) = w_{ij} + w_i + w_j$ cannot be lower than w_i or w_j (i.e., the sums of squares are additive). The hierarchy produced by MISSQ is monotonic regardless whether the increments or the new sums of squares are indicated as fusion levels. This is not so with MIVAR, because from the inequality

$$\left(\frac{n_h + n_i}{n_.}\right)^2 + \left(\frac{n_h + n_j}{n_.}\right)^2 - \frac{n_h(n_i + n_j)}{n_.^2} \geqq 1$$

it should follow that

$$n_h^2 \geqq n_h n_i + n_h n_j + 2 n_i n_j$$

which does not always hold, so the increments are not monotonic. The result of MIVAR is very strongly influenced by the algorithm employed. For example, the RNN algorithm produces a tree which might suggest the existence of some clusters even in random data, while the dendrogram of the CP algorithm exhibits extensive chaining (Fig. 1).

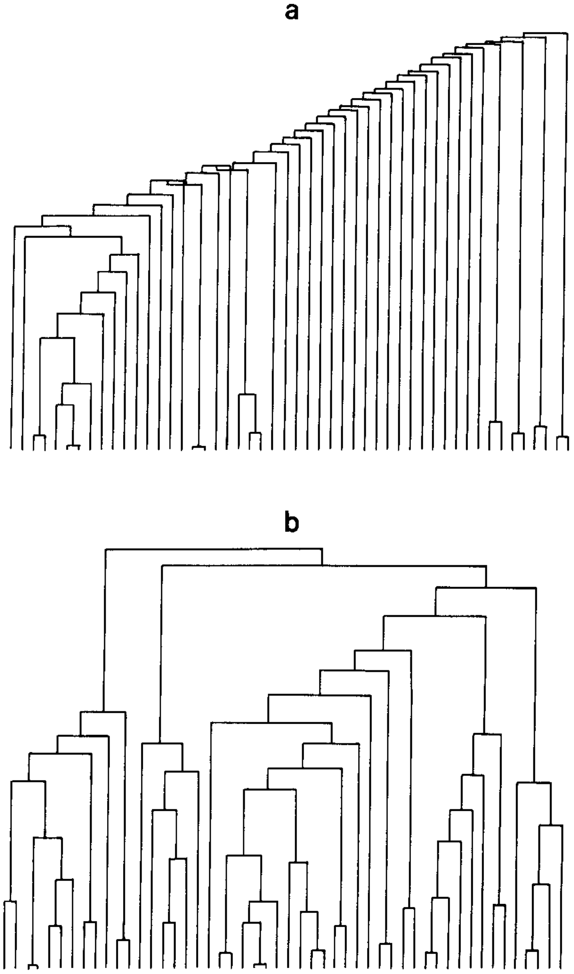

Fig. 1. Comparison of dendrograms produced by alternative algorithms of the MIVAR clustering strategy from a random data set. a) CP algorithm, b) RNN algorithm.

Although in this example only the CP algorithm produced reversals, in other analyses (not shown) both algorithms of MIVAR failed to satisfy monotonicity. This is in apparent contradiction with Diday's (1983) view that the reducibility condition holds for MIVAR and its results are always monotonic (see the table in his Appendix 2).

Analyses of random data (not illustrated) revealed that neither algorithm of WMIDIS has the property of producing reversal-free dendrograms. Also, the RNN algorithm of UMIDIS is also liable to failure of monotonicity. Whether the CP version of UMIDIS may also yield reversals, or it is always monotonic, is not known, however.

I was unable to construct artificial or random data which led to reversals for this strategy; in the worst case complete chaining of objects with very small but monotonic increases of levels resulted. A proof is needed to substantiate the statement that UMIDIS-CP is always monotonic.

A flexible h-SAHN clustering strategy

Starting from the conditions (4a–d), which guarantee a monotonic fusion strategy, a new flexible method is defined by imposing the following constraints upon equation (2):

$$\alpha_i + \alpha_j + \beta + \gamma + \lambda_h + \lambda_i + \lambda_j = 1\,,$$

$$\lambda_h = \lambda_i = \lambda_j \leqq 0\,,$$

$$\alpha_i = \alpha_j = \beta, \quad \text{and}$$

$$\gamma = 0\,.$$

The change of parameters under these conditions provides an infinite number of results for the same set of objects. For $\lambda = 0$ (so that $\alpha = \beta = 1/3$), the firstly formed clusters will tend to attract single

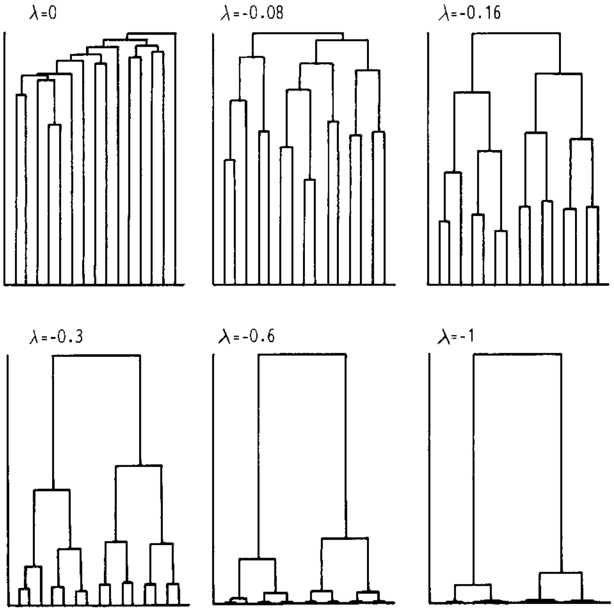

Fig. 2. The effect of λ on the results of a λ-flexible strategy (the CP algorithm is employed).

objects because cluster-to-object measures will generally be lower than object-to-object measures. The strategy seems to be space-contracting and the hierarchy has some degree of chaining so characteristic of single linkage dendrograms. Note that complete chaining observed for high values of β in β-flexible sorting (Lance & Williams 1967) does not occur for this maximum value of λ. As λ becomes large negative, the objects are more intensely grouped and the differences between the first and last fusion levels greatly increase. This is because cluster-to-cluster measures become usually much larger than object-to-object measures. This is demonstrated in Fig. 2 by a series of dendrograms obtained at 6 values of λ for a small data set (15 plots taken randomly from the whole set described below).

Application of combinatorial clustering methods to phytosociology

Data

80 sample plots, each of 4×4 m^2 size, were taken in the rock grassland communities of the Sashegy Nature Reserve, within the city limits of Budapest, Hungary, in 1976. The percentage cover of species was estimated in each plot, but in this paper only presence/absence scores will be used for classifications. The total number of species is 123. The 80 by 123 phytosociological table is not presented here; a copy is available from the author upon request.

Previous classifications

The grassland communities on the dolomite substrate of the study area have long been the subject of intensive phytosociological research. Based on the methods of the Zürich-Montpellier school, Zólyomi (1958) reported 4 community types (associations) from the area. These are: 1) *Festucetum pallentis hungaricum* mostly in southern exposition on rocks and steep slopes; 2) *Caricetum humilis balatonicum* and 3) *Festuco pallenti-Brometum pratensis*, both on hilltops and gentle slopes with some accumulated rendzina, and finally 4) *Seslerietum sadlerianae* on the northeastern slopes with relatively cool microclimate. The objective of my surveys in the reserve was to examine whether this classification can be confirmed by cluster analysis and multidimensional scaling methods. Published results (Podani 1985, 1986, 1988a) seem to suggest that in the presence/absence situation there are 3 vegetational noda in the study area along a combined gradient of species richness and plant cover. The first nodum roughly corresponds to the open *Festucetum pallentis hungaricum*, but the distinction of the other types is less clear in the binary case.

In this paper the classification study mentioned above will be extended by applying all combinatorial methods to the Sashegy data. Euclidean distances between the 80 plots are used because this measure is compatible with every algorithm discussed here. d-SAHN clustering (SL, CL, UPGMA, WPGMA, UPGMC-CP, WPGMC-CP, and β-FLEX with $\beta = -.25$) was performed by program NCLAS2. h-SAHN clustering (MISSQ, MNSSQ, MIVAR-RNN and -CP, MNVAR, WMIDIS-RNN and -CP, UMIDIS-RNN and -CP, MNDIS-CP and λ-FLEX with $\lambda = 0, -.08, -.16, -.30, -.60$, and -1) was carried out by program HMCL2. The resulting dendrograms were compared in every pair based on three dendrogram descriptors (cluster membership divergence, subtree membership divergence, and cladistic difference, Podani & Dickinson 1984). Cophenetic difference and partition membership divergence were excluded from the comparison because of the presence of reversals in some dendrograms and the lack of commensurability in hierarchical levels. The distance matrix of dendrograms as prepared by program DENDAT was subjected to principal coordinates analysis (PCoA, program PRINCOOR) and complete linkage clustering (CL, program NCLAS2) to reveal structural relationships among dendrograms. All the programs used here are included in the SYN-TAX III package (Podani 1988b). The computations were performed on an IBM370 mainframe computer and an IBM AT compatible machine.

Clustering results

The 23 dendrograms (not shown) represent a wide range of classifications of the 80 sample plots. One extreme is the complete chaining of objects (MIVAR-CP, UMIDIS-CP, and WMIDIS-CP) without any groups indicated. This reflects an undesirable property of the CP algorithm of these strategies: the firstly formed cluster will tend to be fused with individual objects one by one in the subsequent clustering steps. Therefore, these extremely space-contracting strategies are not recommended for use in phytosociological classification and their results will be excluded from further comparisons in this paper. Other dendrograms (SL and WMIDIS-RNN) also exhibit a relatively high degree of chaining with the presence of nuclei for some interpretable clusters. UPGMC, WPGMC and λ-FLEX (with $\lambda = 0$) also produced chains, but a large cluster containing relevés from the species rich section of the study area is quite distinct in the hierarchy. The sample plots are subdivided into two large clusters by MIVAR-RNN, UMIDIS-RNN and λ-FLEX ($\lambda = -.08$); the first cluster interpretable as one of the three noda (either the richest or the poorest in species) and the other containing the remaining plots. 3 clusters, clearly identifiable as noda representing 3 levels of species richness (low, intermediate, and high) and total cover (open, transitional, and closed grassland), are depicted by MNVAR, MISSQ and β-FLEX ($\beta = -.25$). The 3 groups may be easily delineated in the map of the study area (see Podani 1985, 1986, 1988a) suggesting good phytosociological interpretability. The dendrograms obtained by λ-FLEX ($\lambda = -.16$, $-.3$, $-.6$, and -1), MNDIS, CL and MNSSQ imply a more refined group structure (four or five clusters) that can be derived by breaking the 3-cluster MISSQ or MNVAR classifications. Finally, the UPGMA and WPGMA dendrograms suggest the existence of even more small clusters.

Comparison of dendrograms

The quick evaluation of results revealed high similarities as well as considerable differences among the alternative classifications. However, a

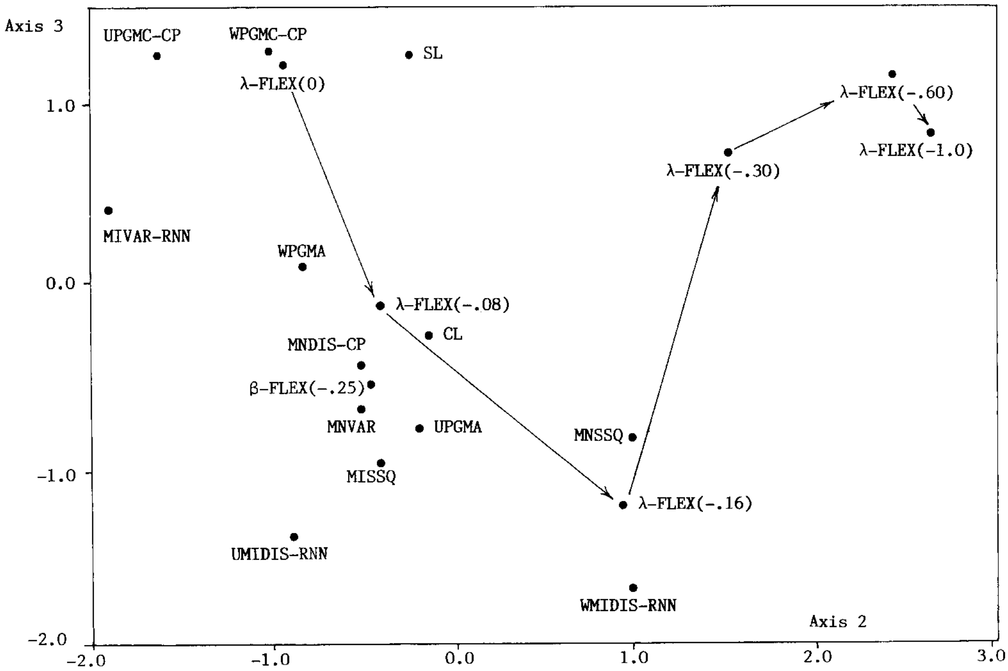

Fig. 3. Principal co-ordinates ordination of 20 dendrograms representing classifications of 80 sample plots from the rock grasslands of Sashegy Nature Reserve (see text for symbols).

more thoroughful analysis of the performance of methods calls for objective assessment. The comparison of partitions of relevés at a nearly constant number of clusters, as done by Gauch & Whittaker (1981) would be less useful in the present study because of the excessive differences in the topological structure of dendrograms. Instead, the multiple comparison strategy of Podani & Dickinson (1984) was adopted as it avoids problems of defining clusters in the hierarchy.

The PCoA ordination of dendrograms reveals that the most important underlying factor implies a tendency towards chaining. On the first component (15.5%) WMIDIS-RNN and SL have large positive scores (4.9 and 2.0, respectively). The other dendrograms are positioned around the centroid (scores ranging from − 1.0 to 0.7); therefore the first axis is not illustrated. The relationships among dendrograms are best explained by the next two axes (12.3% and 7.7%, respectively) shown in Fig. 3. SL, UPGMC, WPGMC and λ-FLEX (λ = 0) form a group of dendrograms already recognized by subjective scrutiny, but they considerably differ from WMIDIS (see also Fig. 4). Note the effect of changing the value of λ in λ-FLEX: for small negative values the dendrograms are close to the majority of results. However, further decrease of λ leads to a rather different classification. The phytosociologically most interpetable results form a cluster, from CL to MNDIS, in the dendrogram of Fig. 4.

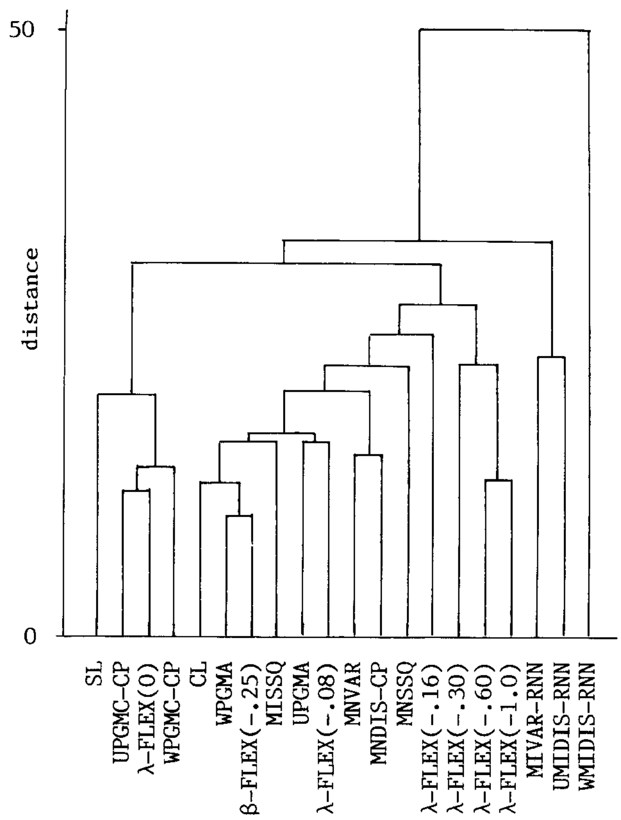

Fig. 4. Classification of 20 dendrograms of the Sashegy relevés by complete linkage clustering (see text for symbols).

Concluding remarks

The combinational SAHN clustering methods may be logically divided into two classes. The d-SAHN methods seek for minimal between-cluster distances and the h-SAHN techniques for maximal within-cluster homogeneity. The d-SAHN clustering methods are compatible with the well-known recurrence formula of Lance & Williams (1966) and make use of parameters α, β and γ. Podani (1978, 1979) used a separate formula for the h-SAHN methods using parameters α, β and λ, whereas Jambu (1978) suggested that all combinatorial methods should be included in the same general equation. The latter suggestion is elegant but a little unfortunate in the sense that the λ-s have no meaning for the d-SAHN methods and the γ parameter is never used by the h-SAHN strategies.

One aim of the present paper is to provide a comprehensive list of combinatorial clustering procedures, with relatively more emphasis placed on the h-SAHN methods. Within-cluster sum of squares, variance and average distance are used as measures of cluster homogeneity. The optimization of homogeneity may be achieved in two basically different ways: the fusion criterion is either the maximization of homogeneity of the new cluster created in a clustering step, or the minimization of the decrease of homogeneity. The combinations of homogeneity measure and fusion criterion define 6 clustering methods, one of them with weighted and unweighted alternatives. 5 of these clustering methods have been suggested earlier, only UMIDIS and WMIDIS appear to be new.

Based on the monotonicity constraints of MNSSQ, MNVAR and MNDIS, a new flexible strategy is defined. This method is apparently similar to the β-flexible method of Lance & Williams (1967) in producing a series of trees that represent transitions between space-contracting and space-dilating clusterings. To establish its potential utility in applied studies, this λ-flexible method deserves future research along the lines of DuBien and Warde's (1979) study on (β, γ)-flexible methods and Milligan's (1987) work evaluating the β-flexible method.

Special attention was paid to the algorithmic and monotonicity properties of combinatorial h-SAHN methods. The CP and RNN algorithm of the same method may produce radically different results, as examples based on random and actual data demonstrated.

The use of the RNN algorithm of MNDIS may destroy the ultrametric structure output by the CP algorithm of the same procedure. As far as I know, it has not been reported earlier that the ultrametric properties of a method depend on the algorithm employed. The explanation of this feature merits a detailed mathematical analysis of these algorithms.

The analysis of field data from the rock grassland communities of the Sashegy Nature Reserve served as a basis for the comparison of d- and h-SAHN procedures. The classifications were considerably different suggesting that there is no clear-cut group structure in the data. The phytosociological implication of the results is that the existence of the community types formerly described from the study area cannot be confirmed in the presence/absence case. Instead, there is a species richness gradient from the open grassland towards the completely closed communities of northern exposition. The congruence among many clustering results still suggests that three noda are worth distinguishing for descriptive purposes. The multiple comparison of dendrograms may help phytosociologists to select clustering procedures that reflect different aspects in the data. However, since the comparisons were based only on a single actual data set, a more extensive simulation study is needed to compare the performance characteristics of combinatorial clustering procedures.

Acknowledgements

I am grateful to A. D. Gordon (University of St Andrews) and P. Juhász-Nagy (L. Eötvös University) for their comments and suggestions. I would like to thank the anonymous referees for their thorough criticism of the manuscript.

References

Anderberg, M.R. 1973. Cluster analysis for applications. Wiley, New York.

Batagelj, V. 1981. Note on ultrametric clustering algorithms. Psychometrika 46: 351–352.

Bruynooghe, M. 1978. Classification ascendante hiérarchique des grands ensembles de données: une algorithme rapide fondé sur la construction des voisinages réductibles. Les Cahiers de l'analyse des Données 3: 7–33.

D'Andrade, R. 1978. U-statistic hierarchical clustering. Psychometrika 43: 59–68.

Day, W.H.E. & Edelsbrunner, H. 1984. Efficient algorithms for agglomerative hierarchical clustering. J. Classif. 1: 7–24.

Diday, E. 1983. Inversions en classification hiérarchique: application à la construction adaptive d'indices d'agrégation. Rev. Stat. Appl. 31: 45–62.

Diday, E., Lemaire, J., Pouget, J. & Testu, F. 1982. Eléments d'analyse de données. Dunod, Paris.

DuBien, J.L. & Warde, W.D. 1979. A mathematical comparison of an infinite family of agglomerative clustering algorithms. Can. J. Stat. 7: 29–38.

Gauch, H.G. & Whittaker, R.H. 1981. Hierarchical classification of community data. J. Ecol. 69: 537–557.

Greig-Smith, P. 1983. Quantitative plant ecology. 3rd ed. Blackwell, Oxford.

Gordon, A.D. 1987. A review of hierarchical classification. J. Roy. Stat. Soc., Ser. A. 150: 119–137.

Gower, J.C. 1967. A comparison of some methods of cluster analysis. Biometrics 23: 623–638.

Jambu, M. 1978. Classification automatique pour l'analyse des données. Tome 1. Dunod, Paris.

Jambu, M. & Lebeaux, M.-O. 1983. Cluster analysis and data analysis. North Holland Publ. Company, Amsterdam.

Johnson, S.C. 1967. Hierarchical clustering schemes. Psychometrika 32: 241–254.

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

Lance, G.N. & Williams, W.T. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. Comput. J. 9: 373–380.

Morena-Casasola, P. & Espejel, I. 1986. Classification and ordination of coastal sand dune vegetation along the Gulf and Caribbean Sea of Mexico. Vegetatio 66: 147–182.

Mucina, L. 1982. Numerical classification and ordination of ruderal plant communities (Sisymbrietalia, Onopordetalia) in the western part of Slovakia. Vegetatio 48: 267–275.

Milligan, G.W. 1979. Ultrametric hierarchical clustering algorithms. Psychometrika 44: 343–346.

Milligan, G.W. 1987. A study of the beta-flexible clustering method. Working paper No WPS 87–61. College of Business, Ohio State University.

Murtagh, F. 1983. A survey of recent advances in hierarchical clustering. Comput. J. 26: 354–359.

Orlóci, L. 1967. An agglomerative method for classification of plant communities. J. Ecol. 55: 193–205.

Orlóci, L. 1978. Multivariate analysis in vegetation research, 2nd ed. Junk, The Hague.

Orlóci, L. & Stanek, W. 1979. Vegetation survey of the Alaska Highway, Yukon Territory: types and gradients. Vegetatio 41: 1–56.

Podani, J. 1978. Hierarchical clustering methods for the analysis of binary phytosociological data. Ph. D. thesis, L. Eötvös University, Budapest (manuscript in Hungarian).

Podani, J. 1979. Generalized strategy for homogeneity–optimizing hierarchical classificatory methods. In: Orlóci, L., Rao, C.R. & Stiteler, W.M. (eds), Multivariate methods in ecological work. pp. 203–209. International Co-operative Publishing House, Burtonsville, Maryland.

Podani, J. 1985. Syntaxonomical congruence in a small-scale vegetation survey. Abstr. Bot. 9: 99–128.

Podani, J. 1986. Comparison of partitions in vegetation studies. Abstr. Bot. 10: 235–290.

Podani, J. A method for generating consensus partitions and its application to community classification. Coenoses (in press).

Podani, J. 1988. SYN-TAX III. A package of programs for data analysis in ecology and systematics. Coenoses 3: 111–119.

Podani, J. & Dickinson, T.D. 1984. Comparison of dendrograms: a multivariate approach. Can. J. Bot. 62: 2765–2778.

Popma, J., Mucina, L., van Tongeren, O. & van der Maarel, E. 1983. On the determination of optimal levels in phytosociological classification. Vegetatio 52: 65–75.

Sneath, P.H.A. & Sokal, R.R. 1973. Numerical taxonomy. 2nd ed. Freeman, San Francisco.

Sokal, R.R. & Michener, C.D. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38: 1409–1438.

van der Maarel, E. 1979. Multivariate methods in phytosociology, with reference to the Netherlands. In: Werger, M.J.A. (ed.), The study of vegetation. pp. 163–225. Junk, The Hague.

Wishart, D. 1969. An algorithm for hierarchical classifications. Biometrics 25: 165–170.

Zólyomi, B. 1958. The natural vegetation of Budapest and its surroundings. In: Pécsi, M. (ed.), Budapest természeti képe. pp. 508–642. Akadémiai, Budapest (in Hungarian).

Appendix

The derivation of updating parameters for h-SAHN clustering methods.

nh-SAHN strategies

As far as the derivation of parameters is concerned, this is the simpler situation. In general, a total of all inter-object distances is reproduced from the six w-s, and then this total is divided by some function of $n.$ ($n. = n_h + n_i + n_j$) to yield a homogeneity measure for cluster $C_h \cup C_i \cup C_j$.

MNSSQ

Since w_{ij} is obtained as the sum of all squared within-cluster distances divided by the number of objects in the cluster, we can write that

$$w_{ij} = \frac{1}{n_i + n_j} \sum_{p,q \in C_i \cup C_j} d_{pq}^2 \text{ and}$$

$$w_i = \frac{1}{n_i} \sum_{p,q \in C_i} d_{pq}^2 .$$

$n_i w_i$ is contained in both $(n_h + n_i)w_{hi}$ and $(n_i + n_j)w_{ij}$, and $n_h w_h$ and $n_j w_j$ are also present twice, therefore the sum of squared distances within $C_h \cup C_i \cup C_j$ is

$$\sum_{p,q \in C_h \cup C_i \cup C_j} d_{pq}^2 = (n_h + n_i)w_{hi} + (n_h + n_j)w_{hj} + $$
$$+ (n_i + n_j)w_{ij} - n_h w_h - n_i w_i - n_j w_j .$$

This sum divided by $n.$ yields the sum of squares in cluster $C_h \cup C_i \cup C_j$.

MNVAR

The total of squared distances in cluster $C_i \cup C_j$ is reproduced as follows,

$$w_{ij} = \frac{1}{(n_i + n_j)^2} \sum_{p,q \in C_i \cup C_j} d_{pq}^2 \text{ and}$$

$$w_i = \frac{1}{n_i^2} \sum_{p,q \in C_i} d_{pq}^2 .$$

Then

$$\sum_{p,q \in C_h \cup C_i \cup C_j} d_{pq}^2 = (n_h + n_i)^2 w_{hi} + (n_h + n_j)^2 w_{hj} + \\ + (n_i + n_j)^2 w_{ij} - n_h^2 w_h - n_i^2 w_i - n_j^2 w_j \ .$$

This sum divided by $n_.^2$ yields the variance in cluster $C_h \cup C_i \cup C_j$.

MNDIS

The derivation of the updating parameters, shown for distances, starts by reproducing the sum of distances within clusters:

$$w_{ij} = \frac{1}{b_{ij}} \sum_{p,q \in C_i \cup C_j} d_{pq} \text{ and}$$

$$w_i = \frac{1}{b_i} \sum_{p,q \in C_i} d_{pq} \ ,$$

where the binomial coefficient $b_i = \binom{n_i}{2}$ is the number of pairs within C_i, and $b_{ij} = \binom{n_i + n_j}{2}$. Then, by similar justification as in the above cases, the sum of all distances within cluster $C_h \cup C_i \cup C_j$ is

$$\sum_{p,q \in C_h \cup C_i \cup C_j} d_{pq} = b_{hi} w_{hi} + b_{hj} w_{hj} + \\ + b_{ij} w_{ij} - b_h w_h - b_i w_i - b_j w_j \ .$$

This sum divided by $b_. = \binom{n_.}{2}$ yields the average within-cluster distance for $C_h \cup C_i \cup C_j$.

ch-SAHN strategies

The derivation of parameters starts by reproducing the total of pairwise distances for the new cluster $C_h \cup C_i \cup C_j$. This quantity is divided by a function of $n_.$ yielding the homogeneity measure for that cluster. Then, subtraction of unweighted or weighted homogeneity measures of clusters C_h and $C_i \cup C_j$ from this quantity gives the increment.

MISSQ

w_i and w_j are the sum of squares for clusters C_i and C_j, respectively, w_{ij} is the increase upon their fusion, therefore the sum of squares of cluster $C_i \cup C_j$ will be $w_{ij} + w_i + w_j$. Thus

$$\sum_{p,q \in C_i \cup C_j} d_{pq}^2 = (n_i + n_j)(w_{ij} + w_i + w_j)$$

and

$$\sum_{p,q \in C_h \cup C_i \cup C_j} d_{pq}^2 = (n_h + n_i)(w_{hi} + w_h + w_i) + \\ + (n_h + n_j)(w_{hj} + w_h + w_j) + \\ + (n_i + n_j)(w_{ij} + w_i + w_j) - \\ - n_h w_h - n_i w_i - n_j w_j \ .$$

This quantity divided by $n_.$ gives the sum of squares of cluster $C_h \cup C_i \cup C_j$. Then, the increase of sum of squares is obtained by subtracting the heterogeneity of clusters $C_i \cup C_j$ and C_h,

$$w_{h,ij} = \frac{1}{n_.} \sum_{p,q \in C_h \cup C_i \cup C_j} d_{pq}^2 - (w_{ij} + w_i + w_j + w_h) =$$

$$= \frac{1}{n_.}[(n_h + n_i)w_{hi} + (n_h + n_j)w_{hj} + ((n_i + n_j) - n_.)w_{ij} + \\ + ((n_i + n_j) + (n_h + n_i) - n_i - n_.)w_i + \\ + ((n_i + n_j) + (n_h + n_j) - n_j - n_.)w_j + \\ + ((n_h + n_i) + (n_h + n_j) - n_h - n_.)w_h] \ .$$

The last three terms cancel, so the formula reduces to

$$w_{h,ij} = \frac{1}{n_.}[(n_h + n_i)w_{hi} + (n_h + n_j)w_{hj} - n_h w_{ij}] \ ,$$

therefore parameters λ_h, λ_i and λ_j are zero.

MNVAR

The variance of cluster $C_i \cup C_j$ may be determined according to the formula

$$VAR(C_i \cup C_j) = w_{ij} + \frac{n_i}{n_i + n_j} w_i + \frac{n_j}{n_i + n_j} w_j \ ,$$

and similar relations hold for $VAR(C_h \cup C_i)$ and $VAR(C_h \cup C_j)$. Then, the variance of $C_h \cup C_i \cup C_j$ will be

$$VAR(C_h \cup C_i \cup C_j) = \frac{1}{n_.^2}[(n_i + n_j)^2 \, VAR(C_i \cup C_j) + \\ + (n_h + n_i)^2 \, VAR(C_h \cup C_i) + (n_h + n_j)^2 \, VAR(C_h \cup C_j) - \\ - n_h^2 w_h - n_i^2 w_i - n_j^2 w_j] \ . \qquad (7)$$

The variance before the fusion of C_h with $C_i \cup C_j$ is

$$VAR(C_h; C_i \cup C_j) = \frac{n_h}{n_.} w_h + \frac{n_i + n_j}{n_.} VAR(C_i \cup C_j) \ . \qquad (8)$$

Substitution of variances into (7) and (8), and subtraction of (8) from (7) yields the increment sought:

$$w_{h,ij} = \frac{(n_h + n_i)^2}{n_.^2} w_{hi} + \frac{(n_h + n_j)^2}{n_.^2} w_{hj} + \\ + \left(\frac{(n_i + n_j)^2}{n_.^2} - \frac{n_i + n_j}{n_.} \right) w_{ij} + \\ + \left(\frac{n_h^2 + n_j n_h + n_j n_i}{n_.^2} - \frac{n_j}{n_.} \right) w_h +$$

$$+\left(\frac{n_i^2 + n_i n_h + n_i n_j}{n_.^2} - \frac{n_i}{n.}\right) w_i +$$

$$+\left(\frac{n_j^2 + n_j n_h + n_j n_i}{n_.^2} - \frac{n_j}{n.}\right) w_j \,.$$

The last three terms cancel and the formula reduces to:

$$w_{h,ij} = \frac{(n_h + n_i)^2}{n_.^2} w_{hi} + \frac{(n_h + n_j)^2}{n_.^2} w_{hj} - \frac{n_h(n_i + n_j)}{n_.^2} w_{ij} \,.$$

WMIDIS

Cluster sizes are disregarded so that the average distance within cluster $C_i \cup C_j$ is

$$DIS(C_i \cup C_j) = w_{ij} + \frac{1}{2} w_i + \frac{1}{2} w_j \,, \tag{9}$$

and similar relations exist for $DIS(C_h \cup C_i)$ and $DIS(C_h \cup C_j)$. The average distance within $C_h \cup C_i \cup C_j$ is calculated as

$$\begin{aligned} DIS(C_h \cup C_i \cup C_j) = \frac{1}{b.}[&b_{hi} DIS(C_h \cup C_i) + \\ &+ b_{hj} DIS(C_h \cup C_j) + b_{ij} DIS(C_i \cup C_j) - \\ &- b_h w_h - b_i w_i - b_j w_j] \,. \end{aligned} \tag{10}$$

The average distance before the fusion of C_h with $C_i \cup C_j$ is

$$DIS(C_h;\, C_i \cup C_j) = \frac{1}{2} w_h + \frac{1}{2} DIS(C_i \cup C_j) \,. \tag{11}$$

Substitution of average distances into (10) and (11), subtraction of (11) from (10) and subsequent rearrangement of the formula give the recurrence relation:

$$\begin{aligned} w_{h,ij} = \frac{1}{b.}[&b_{hi} w_{hi} + b_{hj} w_{hj} + \\ &+ (b_{ij} - b./2) w_{ij} - ((b_h + n_i n_j)/2) w_h + \\ &+ ((b. - 2b_i - 2n_h n_j)/4) w_i + \\ &+ ((b. - 2b_j - 2n_h n_i)/4) w_j] \,. \end{aligned}$$

UMIDIS

Formula (9) is replaced by

$$DIS(C_i \cup C_j) = w_{ij} + \frac{b_i}{b_i + b_j} w_i + \frac{b_j}{b_i + b_j} w_j \tag{12}$$

and expression (11) is rewritten as

$$DIS(C_h;\, C_i \cup C_j) = \frac{b_h}{b_h + b_{ij}} w_h + \frac{b_{ij}}{b_h + b_{ij}} DIS(C_i \cup C_j) \,. \tag{13}$$

Substitution of average distances into (12) and (13) and subtraction of (13) from (12) leads to a very complicated formula which cannot be brought into a much simpler form because sums of binomial coefficients are present in the denominators. It is left to the reader to show that the recurrence relation for UMIDIS takes the form

$$\begin{aligned} w_{h,ij} = \frac{1}{b.}[&b_{hi} w_{hi} + b_{hj} w_{hj} + (b_{ij} - b.b_{ij}/(b_h + b_{ij})) w_{ij} + \\ &+ (b_h b_{hi}/(b_h + b_i) + b_h b_{hj}/(b_h + b_j) - \\ &- b_h b./(b_h + b_{ij}) - b_h) w_h + (b_i b_{ij}/(b_i + b_j) + \\ &+ b_i b_{hi}/(b_h + b_i) - b_i b.b_{ij}/((b_h + b_{ij})(b_i + b_j)) - b_i) w_i + \\ &+ (b_j b_{ij}/(b_i + b_j) + b_j b_{hj}/(b_h + b_j) - \\ &- b_j b.b_{ij}/((b_h + b_{ij})(b_i + b_j)) - b_j) w_j] \,. \end{aligned} \tag{14}$$