A COMPARISON OF SOME METHODS OF SELECTING SPECIES IN VEGETATION ANALYSIS[1]

M. B. Dale CSIRO, Division of Computing Research, Carmody Rd., St Lucia 4067 Australia

M. Beatrice, Istituto di Botanica, Università di Bologna, Italy

R. Venanzoni, Istituto di Botanica, Università di Camerino, Italy

and

C. Ferrari Istituto di Botanica, Università di Bologna, Italy

Keywords: Species reduction, Constancy, Fidelity, Discrimination, Uniformity, Correlation, Specificity, Cost, Concordance, Compositional similarity, Order similarity, Partition comparison, Matrix comparison, Compatibility, Robust, Resistant.

Abstract: This paper examines the reasons for, and methods of, species selection which have been proposed for phytosociological use. It attempts to evaluate these methods in terms of cost and effectiveness, by determining if the procedures can identify constant, faithful and discriminating species. Using a single set of data, 5 classes of methods were identified and typical methods from three of the classes were then applied to a second data set to determine their efficacy in recovering patterns similar to those in the full data.

The results indicate that two methods, one based on the two-parameter model, the other on a double Poisson model, perform well, and at reasonable computational cost. In addition various proposed tests of resemblance of similarity matrices and partition comparison are examined and shown to be of limited usefulness. Some possible extensions to the double Poisson method are indicated, and some relationships to robust and resistant estimation procedures noted.

Introduction

There are three major reasons for wanting to select subsets of species. The first is simply necessity, usually because computational facilities are inadequate. Such limitations have largely disappeared for large computer systems and most of the simpler methods of analysis, although a method such as Johnson and Goodall's (1979) maximum likelihood ordination may still meet problems. The increasing usage of microcomputers has tended to reintroduce such constraints on the size of problem which can be handled, but such restrictions are probably only temporary.

The second reason is to reduce a complex distinction by approximating it with a simpler one. Discriminant functions, for example, produce complex polythetic distinctions, while taxonomic keys (see e.g. Pankhurst 1978) replace these with a simpler monothetic choice. Much effort has been put into *a posteriori* selection for discrimination, but methods such as Association Analysis (Williams and Lambert 1959) provide the simple solution *a priori*. Indicator species analysis (Hill, Bunce and Shaw 1975) provides an interesting combination of polythetic division with oligothetic characterisation, trying to get the best of both worlds.

The third, and most interesting, reason is to reduce the effects of irrelevant species, *i.e.* to remove species which do not contribute to the major patterns. Such species in effect provide a noisy background in which interesting patterns are embedded, so that their removal should accentuate the clarity of the patterns. In this paper we shall examine various methods which have been proposed in the literature for performing such reduction.

Before doing this it should be noted that the effects of such interfering, 'noisy', species is also related to the estimation procedures used in the pattern finding programs. In fact most of the ordination procedures in use employ least squares algorithms, and these are known to be very sensitive to outlying values. This is true for principal component analysis, principal coordinates analysis, and correspondance analysis. The currently fashionable detrended correspondance analysis is likely to suffer problems both in identifying the axes and in the subsequent detrending procedure. It would be possible to use robust or resistant methods, as proposed by Gabriel and Odoroff (1983) but these are neither well known nor widely available at present. Even with robust methods it is necessary to obtain a good estimate of the weights before starting the estimation procedure. Such estimation of weights is, in effect, what species selection procedures do. Thus, if the weighting procedures are effective, they can be used to initialise robust estimation procedures.

Most such selection methods start by ordering the species according to some more or less complex criterion, and then accept the higher ranking species as relevant, rejecting the others. Thus the ordering provides a weighting which is followed by a dichotomy. Here we shall not be particularly concerned with the problem of dividing the set, for which several methods are available. Equally we shall not be examining the possibility of directly using the weights in a robust procedure. Through the courtesy of D. Faith we have recently studied a new proposal of this kind which calculates weights for a species from similarity measures for the stand pairs where the species occurs in

[1] This paper is dedicated to the memory of Monica Beatrice, our coauthor, friend, collaborator and excitement centre.

both stands, then recalculates the similarities, and iterates until the procedure converges. Here the similarities will be directly usable although the computational cost will be high; in fact proportional to the square of the number of sites and of species. This may be acceptable for moderate size problems using the latest vector processors. It should be noted that this iterative reweighting procedure is closely related to an earlier proposal of Feoli (1973) to use the ratio of similarities between sites containing a species, and similarities of sites with the species absent to provide a more effective monothetic divisive classification algorithm.

Although a variety of methods have been described in the literature, often under the heading of feature selection, few have had wide use, and most have been applied only by their initial inventors. We have written a number of programs for calculating weights, some of which will be available in the TAXON program package (Ross 1983). Since the cost of computing weights can be significant, the first problem is to determine if the different methods do indeed order the species differently. If they do then we can reduce the number of methods which need more detailed assessment to determine if they significantly modify the information recovered by subsequent analysis. We are of course seeking interesting species and we start by attempting to identify what makes a species interesting in a phytosociological context. This is difficult to do, but even if others disagree with our arguments it may encourage them to provide more exact treatment themselves. We have also to consider the robustness of the estimation of the weights itself, for if this is markedly affected by the noisy species then the weighting will itself be biased. Finally the computational cost is of significance since there is little merit in paying very highly unless the advantages to be gained are comparatively large.

It should be noted that there are other reasons for requiring selection of characters. Estabrook, Johnson and McMorris (1976) have proposed a method of recovering putative phylogenies based on assessing compatibility of character state distributions. The model employed is rather specific to taxonomy, and to cladistics, and will not be further discussed here.

Definition, Discrimination and Assessment

If we start with a single set of releves and no *a priori* defined types, then we would like a species selection procedure to identify at least four different kinds of species. First there are the uninteresting, *noise*, species whose rejection will help us in identifying residual patterns. Second there are species whose presence is necessary in at least one type, though not necessarily restricted to it. These form the class of *constant* species which was emphasised by Scandinavian phytosociological approaches. Third is the class of species restricted to a single type, and hence sufficient to identify it. These are the *faithful* species of the Braun-Blanquet approach. Note that they need not be present in all the relevés of the type. Fourth, and finally, there are *discriminatory,* or differential, species which enable us to distinguish one type from another. Like the constant species these are not necessarily restricted to a single type, but in specific cases enable us to distinguish one type from another. Such discriminatory species will not exist, of course, if our data contains only a single type, and, where they do exist, may not be distinguishable from constant or faithful species; the species classes, other than the first, are not mutually exclusive.

We must seek, at the least, a method which will distinguish the first class from the others. Hopefully we can do better and also distinguish in some manner between the threee classes of interesting species as well. However any analysis based on a subset of the species can involve the loss of information, which may not all be noise. But how can we assess the significance of such losses? There are several possible answers to this question, depending on where in the subsequent analysis we choose to make the assessment. But the answer also, and obviously, depends on the objectives of the reduction.

We could start by asking if the different methods of selection agree among themselves, *i.e.* are concordant, and if so then we can choose the cheapest alternatives for further assessment. Several classes of methods may exist so that a simple test of overall concordance is not sufficient. One means of assessing concordance is to classify the methods using the subset of species they select as descriptors. We can then choose representative members of each class for further examination.

Having so chosen we proceed to further examine the relationships between subsets and the full data by calculating similarity matrices for each. These can be compared directly to determine if the resemblance structure contained in them is markedly modified by the selection procedures. A test such as Mantel's (1967) is strictly valid only if the two matrices being compared are independently generated; this is clearly not true if we are using subsets of the same data. Such a test can still be used to assess relationship, but significance levels must be regarded as suspect. Hubert (1979) shows a generalisation of Mantel's test permitting the comparison of k matrices, and his paper also shows some relationships with rank-based measures of concordance.

Proceeding, we can now derive an ordination from each similarity matrix and attempt to use some measure of canonical correlation to indicate similarity and dissimilarity in the patterns obtained. Direct use of canonical correlation is not possible of course. Again significance tests would be suspect. Furthermore a method such as Procrustes rotation (Gower 1975) gives considerable freedom since it permits shifts of origin, changes of scale and rotation all to be applied to attain maximal correspondance. The measure of similarity obtained will thus depend on the set of transformations which we permit the algorithm to employ. Similarity will thus depend both on the species selected and on the algorithm, which is not really desirable.

We can make a weaker test by simply requiring that

some estimate of the topological dimensionality of the sets be the same. The topological, or intrinsic, dimensionality can be defined as the minimum number of parameters required by a continuous function which adequately predicts the data. Trunk (1968) for example gives a method of estimating this value by a method he terms maximum likelihood. This seems quite attractive as long as we do not want to be very precise in our identification of the function. Random data should show a greater variability in the estimates, which should be reduced if our selection procedure is working well. Thus we are not dependent on taking the patterns identified in the full data as our target in the assessment procedure. The estimate of variance for Trunk's estimate of dimensionality can be obtained from the components which he combines to obtain his estimate.

Alternatively we can use classification methods and compare the partitions, and possibly hierarchies if hierarchical methods are used. The number of groups need not be identical in any pair of analyses, and objective stopping rules would be desirable. There are various possible measures in the literature for providing stopping rules, and for comparing partitions and hierarchies. Note that we are not really interested here in obtaining a tree which combines the information from several analyses, the so-called consensus trees, since our objective is to select effective methods, not to combine several alternative estimates.

Finally what we require is that the patterns we induce from an effective and robust analysis of the full data should also be recoverable from the reduced data. It is not clear that we already have robust methods of analysis and, even if we had, not all classes of pattern might be equally resistant to the selection procedures. The only final appeal is to human interpretation. Logical theories of induction do exist, but are rather unhelpful especially as our methods are heuristic rather than optimal. It is thus possible that failure of the selection procedure could be due to lack of robustness in the search procedures.

Subset selection *a posteriori* is possible and much simpler. Webb, Tracey, Williams and Lance (1967) reduced 818 species to 266 without loss of pattern and to 65 with remarkably little effect. Here the selection criteria were biological not numeric, the 65 species being those which could reach the canopy of the forest, which the authors term the 'big-trees'. Such selection of species on biological grounds is ubiquitous. In collecting data we often reject certain categories as irrelevant or uninteresting. We might, for example, choose to ignore bryophytes, lichens, epiphytes or ground flora. As a consequence of Watanabe's (1969) *Theorem of the Ugly Duckling,* some such selection is always necessary and all phytosociologists practice it. What is less clear is the nature of the rules used to make such selection. Pragmatic reasons often predominate; difficulties of collection and identification are often the reason for rejecting certain taxonomic groups. There can, however, be ecological reasons for rejection. Bryophytes are commonly regarded as responding to a very local environment, especially in areas where they do not form the major component of biomass. Their rejection is then presumably due to a decision that the scale of the investigation is too large for bryophytes to contribute any useful information. Such beliefs are often difficult to justify, as in the case of the general use of species as descriptors where the justification often seems to be an argument *ad populum* rather than a reasoned choice.

Data and Analysis

To provide some variety of material we have used two sets of data. The first is a composite of several published phytosociological tables, with various Associations and subassociations distinguished. In addition several less well defined groupings exist within the data. There are 129 relevés (12 from Ferrari 1971, 6 from Ferrari and Speranza 1976, 58 relevés in two subsets (25 and 34) from Ferrari and Grandi 1974, and finally 54 relevés in 4 transects of 13, from Ferrari and Galanti 1972). We can expect to find discriminating species for at least some of the Associations, and we would hope, of course, to distinguish constant and faithful species too. We can also expect some species to be removable without much loss in information. This data set was used in establishing the concordance of the various methods, and it was discovered after this had been completed that two species had been erroneously entered twice. This was corrected for subsequent analyses, and makes no difference to the concordance assessment, where the usefulness of the subset is not being assessed, only its compatibility with other subsets. Since these data are all published they are not repeated here.

The second data set consists of 13 relevés and 95 species (Table 1) from Venanzoni (unpublished). Since this is an example of a single type, we should not expect to find discriminating species. However it is clearly over-defined and a considerable reduction in species numbers should be possible.

In reducing the number of species, given the weights, a subjective assessment was used to determine a useful rejection level. Objective means could have been used but most of the obvious candidate procedures are strongly affected by outlying values and would thus lead to very small subsets. Where the original author defined selection procedures, these have been used. In most cases we elected to reject *c.* 75% of the species. This figure was chosen from the results of the *a posteriori* selection method examined, where a value of 0.5 seemed an appropriate cutpoint and led to retention of 25% of the species. A 75% rejection rate is fairly severe, though not so harsh as the 90% obtained by Webb *et al* (1967). For two methods, as we shall see later, wholesale rejection on this scale proved impossible, because of tied values.

Twenty-five selection methods were investigated in all, and the first step is to assess their concordance. This was attempted using classification methods. Two similarity coefficients were used, one a Manhattan metric (Minkowski distance of order 1) which measure how far the selected subsets had the same content. The other was a Levenshtein distance (see e.g. Lu and Fu 1978) which measures similarity in ordering as well as in content. It ef-

Table 1. Second data set. Species marked with * are characteristic for the Class Rhamno-Prunetea

Shrubs														
Prunus spinosa*	+.2	1.2	+	+	1.2	1.2	2.2	4.4	2.2	2.2	1.1	+	1.1	12
Crataegus oxyacantha*	1.2	1.1	+	+	1.1	+	1.1	+	.	+	2.2	3.4	3.3	12
Rubus caesius	.	+	+	+	+	1.1	+	+	+	+	+	+	1.1	12
Ligustrum vulgare*	1.2	.	+	1.2	1.1	+	1.1	.	1.2	1.2	1.2	1.1	+	11
Acer campestre	+	+	+	+	+	+	.	.	1.2	1.2	±	+	+	11
Lonicera etrusca	.	+	+	+	.	+	+	+	+	+	+	+.2	+	11
Cornus sanguinea*	+	.	+	.	1.2	+	+	+	+.2	+	+	+	.	10
Rhus coriaria	3.4	4.4	5.5	4.4	3.4	5.5	1.2	1.2	1.2	.	.	.	.	9
Clematis vitalba*	+	.	+	+	+	1.1	1.1	.	.	.	+	+	1.1	9
Quercus pubescens	1.1	+	+	+	+	+	+	.	.	.	.	+	+	9
Euonimus europeaeus*	.	+	.	+	+	+	+	+	.	.	+	.	1.1	8
Rosa canina*	+	.	.	.	+	+	+	+	+	.	+	+	.	8
Ulmus campestris	.	.	.	+	+	+	+	+	.	+	+	.	+	8
Tamus communis	.	.	+	+	.	.	.	.	.	+	.	+	.	4
Cercis siliquastrum	+	+	.	.	.	.	.	.	1.2	.	.	.	.	3
Fraxinus ornus	+	+	.	.	.	.	.	.	.	.	.	+	.	3
Osiris alba	+	+	.	.	.	.	.	.	.	.	.	.	.	2
Paliurus australis	.	.	.	+	.	.	.	1.2	.	.	.	.	.	2
Prunus avium	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Lonicera xylosteum	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Spartium junceum	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Pistacia terebinthus	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Malus sylvestris	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Juniperus oxycedrus	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Cornus mas	.	.	+	.	.	.	.	.	.	.	.	.	.	1
Lonicera caprifolium	.	.	+	.	.	.	.	.	.	.	.	.	.	1
Laburnum anagyroides	.	.	.	+	.	.	.	.	.	.	.	.	.	1
Sorbus domestica	.	.	.	.	+	.	.	.	.	.	.	.	.	1
Corylus avellana	.	.	.	.	+	.	.	.	.	.	.	.	.	1
Acer monspessulanum	.	.	.	.	.	.	+	.	.	.	.	.	.	1
Sambucus ebulus	.	.	.	.	.	.	.	.	.	.	.	.	+	1
Herbs:														
Brachypodium pinnatum	+.2	2.3	+	+	+	+	+	1.2	+	+	1.1	+	+	13
Hedera helix	+.2	+.2	3.3	+	1.2	+	+	1.2	1.1	1.1	+	2.2	.	12
Asparagus acutifolius	1.1	1.1	+	+	+	+	+	+	+.2	.	.	+.2	+	11
Cruciata glabra	.	.	+	.	+	+	+	+	+	+	+	+	+	10
Arum italicum	.	+	+	+	.	+	+	.	1.2	+	+	+	+	10
Galium aparine	.	.	.	+	+	+	+	+	+	+	.	.	+	8
Chaerophyllum aureum	.	.	.	+	.	.	+	+	.	.	+	+	+	6
Leopoldia comosa	.	+	+	+	.	+	.	.	.	.	.	+	.	5
Dactylis glomerata	.	.	+	.	.	+	.	+.2	.	+	+	.	.	5
Lamium maculatum	.	.	+	+	.	+	+	.	.	.	.	.	1.2	5
Alliaria petiolata	.	.	.	+	+	.	+	+	+	.	.	.	.	5
Rumex sanguineum	.	.	.	+	+	.	.	.	.	+	+	.	+	5
Galium mollugo	.	.	+	.	.	.	.	.	.	+	1.2	+	1.1	5

fectively measures the number of insertions and deletions necessary to convert one list of species into another. Both similarity matrices were classified hierarchically using Lance-Williams (1966) flexible sorting strategy with *Beta* value set to -0.25. A Principal Coordinates analysis (Gower 1966) was also performed. The number of groups was fixed from the Manhhattan analysis using the Ratkowsky-Lance (1978) criterion. This employs a between/within variance measure normalised for the number of groups to indicate optimal grouping. In addition an additive similarity tree was calculated using the method of Sattath and Tversky (1977). It was thought possible that the methods might show some hierarchical structure, and this procedure reflects the distance between subsets in the length of branches on the tree. This is unlike most other classification methods which result in ultrametric trees where only connectivity is of interest. In the results the difference between the two similarity measures was not pronounced and we here concentrate on the Levenshtein results, as the more sensitive measure. Only the first set of data was analysed in this way, and typical members of the classes of selection method formed were then applied to

Bromus erectus	.	+	+	.	.	.	+	.	.	.	+	+.2	.	5
Inula conyza	.	+	.	+	.	+	+	.	.	.	.	.	.	4
Stellaria media	.	.	+	+	.	.	.	.	.	+	.	.	+	4
Buglossoides purpurocaerulea	.	.	.	.	+	+	.	.	.	.	.	1.1	.	3
Rubia peregrina	.	+	.	.	+	.	.	+	.	.	.	.	.	3
Bryonia dioica	.	.	.	.	+	.	.	.	+	+	.	.	.	3
Teucrium chamaedrys	.	.	.	+	.	.	.	.	.	.	+	+	.	3
Digitalis micrantha	+	+	.	.	.	.	.	.	.	.	.	.	.	2
Orchis purpurea	+	.	.	+	.	.	.	.	.	.	.	.	.	2
Ranunculus bulbosus	+	.	.	.	+	.	.	.	.	.	.	.	.	2
Silene vulgaris	.	.	+	+	.	.	.	.	.	.	.	.	.	2
Bunium bulbocastanum	.	.	.	.	+	+	.	.	.	.	.	.	.	2
Viola hirta	.	.	+	.	.	.	.	.	.	.	.	+	.	2
Poa trivialis	.	.	.	.	.	.	.	.	.	+	.	.	+	2
Agropyrum sp.	+.2	.	.	.	.	.	.	.	.	.	.	.	.	1
Galium lucidum	+	.	.	.	.	.	.	.	.	.	.	.	.	1
Euphorbia cyparissias	+	.	.	.	.	.	.	.	.	.	.	.	.	1
Dorycnium hirsutum	+	.	.	.	.	.	.	.	.	.	.	.	.	1
Saponaria ocymoides	+	.	.	.	.	.	.	.	.	.	.	.	.	1
Teucrium flavum	.	+.2	.	.	.	.	.	.	.	.	.	.	.	1
Veronica hederifolia	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Ferula communis	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Cephalaria leucantha	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Geranium robertianum	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Anemone coronaria	.	+	.	.	.	.	.	.	.	.	.	.	.	1
Crepis vesicaria	.	.	+	.	.	.	.	.	.	.	.	.	.	1
Geranium sp.	.	.	+	.	.	.	.	.	.	.	.	.	.	1
Arabis turrita	.	.	.	+	.	.	.	.	.	.	.	.	.	1
Geum urbanum	.	.	.	.	+	.	.	.	.	.	.	.	.	1
Carex divulsa	.	.	.	.	.	+	.	.	.	.	.	.	.	1
Glaechoma hederacea	.	.	.	.	.	+	.	.	.	.	.	.	.	1
Eranthis hyemalis	.	.	.	.	.	.	+	.	.	.	.	.	.	1
Campanula trachelium	.	.	.	.	.	.	+	.	.	.	.	.	.	1
Nigella damascena	.	.	.	.	.	.	.	+	.	.	.	.	.	1
Thlaspi arvensis	.	.	.	.	.	.	.	+	.	.	.	.	.	1
Ranunculus ficaria	.	.	.	.	.	.	.	+	.	.	.	.	.	1
Ranunculus lanuginosus	.	.	.	.	.	.	.	.	.	+	.	.	.	1
Geranium pusillum	.	.	.	.	.	.	.	.	.	.	+	.	.	1
Malva hirsuta	.	.	.	.	.	.	.	.	.	.	+	.	.	1
Chaerophyllum temulum	.	.	.	.	.	.	.	.	.	.	+	.	.	1
Arabis sagittata	.	.	.	.	.	.	.	.	.	.	.	+	.	1
Silene alba	.	.	.	.	.	.	.	.	.	.	.	+	.	1
Helleborus foetidus	.	.	.	.	.	.	.	.	.	.	.	+	.	1
Ruscus aculeatus	.	.	.	.	.	.	.	.	.	.	.	+.2	.	1
Fragaria vesca	.	.	.	.	.	.	.	.	.	.	.	+	.	1
Melandrium album	.	.	.	.	.	.	.	.	.	.	.	.	+	1
Carex flacca	.	.	.	.	.	.	.	.	.	.	.	.	+	1
Artemisia vulgaris	.	.	.	.	.	.	.	.	.	.	.	.	+	1
Symphytum tuberosum	.	.	.	.	.	.	.	.	.	.	.	.	+	1
Pastinaca urens	.	.	.	.	.	.	.	.	.	.	.	.	+	1

both data sets, to develop the subsets to be further evaluated, and compared to the full data.

The subsequent analysis for the first data set consisted of a classification using the Bray-Curtis similarity measure, again with flexible sorting, and using the Ratkowsky-Lance criterion to determine number of groups. Comparisons of the partitions formed were made using the Fowlkes-Mallowes (1983) B_k statistic.

Topological dimensionality was estimated using Trunk's (1968) procedure and the similarity of the similarity matrices was determined using Mantel's (1967) statistic. Both the Fowlkes-Mallowes and Mantel methods allow to significance tests, but these, as noted earlier, can only be regarded as indicative.

For the second set of data, Principal coordinate analyses of the Bray-Curtis similarity matrices was followed by a comparison using canonical correlation analysis. Trunk's procedure was also applied. As noted earlier Trunk's procedure would give a very general level of comparison. Unfortunately both data sets returned an intrinsic dimensionality of 1, with no variance! This could hardly be reduced. We had hoped that the subsets and the

Table 2. Feature Selection Methods

Reference.	Measure.	Pair.	Single.	Prior.	Data Type.	Site	Iterative.	Program.	Note No.
Orlóci (1973)	Sum Squares and Cross products	Y	–	Y	A (?P)	–	–	FOBSCP	1
Orlóci (1978)	Communality	Y	–	Y	A (?P)	–	–	FOBSCP	2,3
Orlóci (1978)	Squared Mult. Correlation	Y	–	Y	A (?P)	–	–	FOBSCP	2
Orlóci (1978)	Specificity	–	Y	Y	A (?P)	–	–	FOBSCP	3
Wong-Liu (1975)	Typicality	Y	Y	Y	S	Y	Y	FOBCAI	4
Orlóci (1976)	Joint Inf.	Y	–	Y	F	–	–	FOBINF	5
and	Mutual Inf.	Y	–	Y	F	–	–	FOBINF	5
(1978)	Equivocation	–	Y	Y	F	–	–	FOBINF	5
Feoli, Lagonegro &	Entropy	Y	–	–	F	–	–	FOBINF	5
Orlóci (1982)	Joint Entropy	Y	–	–	F	–	–	FOBINF	5
Dahl (1960)	Uniformity	Y	–	–	P	–	–	FOBUNI	6
Dale, Beatrice & Venanzoni	Uniformity	Y	–	Y	P/A	–	–	FOBUNI	6
Goodall (1953)	Frequency	–	Y	Y	P	–	–	GSTAT	7
Williams, and Lambert (1959)	Sum abs. Correlation	Y	–	Y	P	–	–	(ASSOC)	8
Williams, Dale, Macnaughton-Smith & Mockett (1964)	Sum Squared Correlation	Y	–	Y	P/A	–	–	FOBCOR	9
Williams & Dale (1964)	Correlation	Y	–	Y	P/A	–	–	FOBCOR	9
Dale & Williams (1976)	Eident	–	Y	Y	F	Y	?Y	FOBEID	10
Lance & Williams (1971)	Cramer	–	Y	–	P/A/S	–	–	CRAMER	11
Hogeweg (1976)	Kruskal Wallis	–	Y	–	A	–	–	(KRUSWAL)	12
Boulton & Wallace (1973)	Coding	–	Y	–	P/A/S	–	–	SNOB	13
Harter (1975)	Double Poisson	–	Y	Y	P/F	Y	–	FOBDPW	14

full data for both data sets would have shown some differences. If the full data set have had a slightly higher dimensionality than the subsets, then some loss of information might be inferred. If the subsets had higher dimensionality than the full sets there would have been some indication that removal of noisy species was in fact improving the results. In the absence of any interesting result, further investigation of Trunk's method was discontinued until more suitable data become available.

Species Selection Methods

The list of methods examined is shown in Table 2. We shall not give details of all the methods, since these are largely available in the literature for any interested reader. Instead we shall provide notes on items of significance or interest. The acronyms are unfortunate, perhaps, but are only used here to identify the methods quickly.

In the table, methods which rely on interaction between species are indicated in the *pair* column, while those using only one species at a time are indicated in the *single* column. The *prior* column indicates that which methods use *a priori* assessment while the *data types* indicated are *A* for abundance, *P* for presence, *S* for state data and *F* for frequency data. The *site* column indicates if the method also provides an ordering of the sites jointly with the species ordering, while *iterative* indicates that the method requires an interative solution, whose convergence properties must needs be examined.

Note 1. The results obtained with this method depend on the standardisation of the data; we have used unstandardised values. Unlike correlation based methods, discussed later, species, other than the first selected, are rated using residuals remaining after previously selected species have been removed. The FOBSCP program is based on that of Orlóci (1978).

Note 2. Communality is that part of the variance of one species which is 'shared' by other species and is the com-

plement of specificity. The squared multiple correlation between one species and all others is a well known lower bound on the total communality, although various procedures have been developed for generating more exact values, such as maximum likelihood procedures and Guttman's (1953) Image analysis. Orloci's program was again the basis for ours but it is, at first sight, a somewhat odd and expensive method (Rohlf, 1977). Most methods of estimating communalities are iterative, not robust and liable to numerical instability problems. Orlóci (pers. commun.) regards his program as didactic, and also points out that his suggested procedure partitions the total communality. He first finds the species with the largest communality, then removes the effects of this species before calculating the communality for the second, and so on. This is inherently a somewhat costly computational procedure.

Note 3. The estimation of specificity requires pairwise calculations, since it is the complement of communality. Again Orloci's program was used as the basis for ours.

Note 4. Wong and Liu (1975) combine indices of species composition and species correlation in a single index, and also derive, iteratively, a weighting for typicality of relevés. They accept only *significant* correlations, setting others to a zero value, but this is simply a heuristic to reduce the effects of large numbers of small values. We developed 2 species orderings, one based on composition, here called CAIu, and the other based on correlation, here called CAIfd. For both methods the species selected were those with highest weights, which in facts selects the less typical species. We are thus selecting against constant species. It is possible that reversing the order of selection would be more acceptable as a means of characterising the data, which was the objective of the original paper.

Note 5. These Information measures play a role for frequency data analogous to sums of squares for abundance data. Equivocation plays the same role as specificity, being a measure of unrelatedness. The FOBINF program is again based on Orlóci's (1978) programs. The Feoli, Lagonegro and Orlóci (1984) joint information is closely related, being the sum of mutual and equivocation information.

Note 6. Dahl (1960) suggested his measure of uniformity of a set of relevés as the ratio of the mean number of species per sample to the Fisher *alpha* index of diversity (Fisher, Corbett and Williams 1943), which is defined as:

$$\frac{(\text{total number of species} - \text{mean number of species})}{\log(\text{number of sites})}$$

Dahl associates a uniformity with a single species by defining the group of relevés for which the uniformity is calculated as those containing the species. Dahl, Prestvik and Toftaker (1981) applied the measure to determine character species for associations. In practice the measure seems very sensitive to group size; a minimal group size must be specified, as a group consisting of a single relevé is necessarily uniform. Some normalisation for group size differences thus seems appropriate and we have simply used the ratio of size of group to size of population. It would also be possible to determine uniformity in groups of stands identified by the absence of a species, or to combine both presence and absence estimates into a single composite index. The program used calculates a geometric mean of the indicies, possibly size adjusted, for the presence and the absence groups, but also reports the single group measures, with and without size adjustment.

An extension, developed for this paper, permits the use of abundance data instead of presence/absence data as used by Dahl originally. This means of course that no direct relationship to the index of diversity is present. The measure used is the ratio of the sum of mean abundances of species in the selected group of sites times the logarithm of number of sites (the numerator) and the difference between the maximal abundances and the sum of mean abundances (the denominator). Thus:

$$\frac{\text{sum mean abundances} * \log(\text{number of sites})}{\text{sum maximal abundances} - \text{sum mean abundance}}$$

The measure is a function related to the mean-range ratio. We experimented by using median abundances but this gave no apparent advantage and slightly higher costs of computing. Group size adjustment is still possible with the extended measure. The 'presence' group was determined using a threshold value for abundance, although in the present example only strict presence/absence was employed.

Note 7. Besides Goodall (1953) who combined frequency with correlation, Williams and Lambert (1959) rejected rare species, while Hill and Gauch (1980) downweight such species because they interfere with their algorithm for finding eigenvalues. Since sufficient, or faithful, species are typically rare, this selection procedure must emphasise necessary (constant) species, or discriminating species.

Note 8. The original proposal for Association Analysis (Williams and Lambert 1959) attempted to relate the summed correlations to an averoid factor analysis, but other workers have since used other coefficients with no such interpretation. We did not pursue this alternative as the sum of squared correlations seems preferable for our purposes.

Note 9. The original paper suggested that weights be recalculated during the course of division into groups (see Note 12, and also Macnaughton-Smith, Williams, Dale and Mockett, 1964). Lambert *et al.* (1973) have explored some other alternatives in somewhat similar classification procedures. The Williams and Dale (1962) partitioning is an obvious extension. Thus we have in all 5 correlation measures, Total, LL (Presence/Presence), LN (Presence/Abundance-given-presence), NL (Abundance-given-Presence/Presence) and NN (Abundance-given-presence/Abundance-given-presence).

Note 10. Eident values are deviations from fit of the two-parameter model (see Dale and Anderson, 1973), summed in absolute value. The fitting of the model should strictly be iterative, but Dale (unpublished) has evidence that the initial approximations are quite good; iteration produces changes of less than 10% in value, with the rank order of the species remaining largely invariant.

Note 11. Cramer values (Lance and Williams 1971) are

scaled chi-square and/or variance ratios, calculated for some fixed number of groups. Thus the full data must be used to obtain the groups. The number of groups was determined using the Ratkowsky-Lance criterion, which chooses that number of groups which maximises discrimination, as measured by a mean Cramer value normalised for number of groups. We employed 2 similarity measures, one Bray-Curtis, the other a quantitative Information measure (Dale, 1971), thus giving 2 orderings, Cramer-Bray-Curtis, and Cramer-NIASM.

Note 12. Hogeweg (1976) employs weights derived from a Kruskal-Wallis test in an iterative reweighting procedure. This varies the weights during the course of a series of classifications. The Cramer values can be used in essentially the same manner and represent a parametric analysis of variance analogue, whereas Kruskal-Wallis is based on ranks. Since we were not employing the reweighting procedure we have used only the Cramer values here. We would expect the Cramer values to emphasise species *sufficient* to define groups, whereas the Kruskal-Wallis procedure should be more sensitive to *necessary* species. A combination of the two methods would be quite attractive.

Note 13. The SNOB procedure uses a theory of inductive inference based on coding theory to develop fuzzy classifications, that is classifications where elements may belong in some degree to several groups. At various stages it can identify species which can be coded without further group recognition, thus performing a sort of dynamic species selection procedure. However the computation is expensive, and removal of species before analysis would seem preferable, at least until very fast vector-processing computers are widely available. We have examined the method solely as a means of species reduction in this paper.

Note 14. The problems of identifying index terms for document collections and identifying documents relevant to a search request later are perhaps more severe even than species selection in phytosociology. The number of words and documents can be very large indeed, running to several tens of thousands. If we equate relevés with documents and species with words, we can adapt some of the approaches which have been developed in information retrieval to species selection. The description given here is in phytosociological terms, although the original papers are, of course, phrased for information retrieval.

Harter (1975a, 1975b, see also Bookstein and Kraft 1977) has suggested a model in which every species is related to 2 classes of sites, an *elite* class in which it is important, and a *residual* class in which it is unimportant. We might interpret the elite class as containing those sites in which the species has some indicator value. However the classes are **NOT** to be regarded as directly equivalent to any vegetation type. Such a relationship in possible, for faithful species for example, but it is not necessary. Harter assumes that in both elite and residual classes, a species will have a Poisson distribution, but that the means for the two classes will differ. He then uses moment estimates to obtain both the means and a mixing coefficient, which indicates the degree of overlap of the classes. Maximum likelihood estimation would also be possible, but involves a more expensive iterative algorithm. He then calculates the probability that a site in which a species occurs one or more times, is a member of the elite class, and from this derives a weight for the species.

Van Rijsbergen, Robertson and Porter (1980) have further developed this approach, noting that there is a probabilistic relationship between a site falling in the elite class, and the site belonging to a particular type. The type can be regarded as an Association. No single species need necessarily have an elite class identical to a type but it may indicate the type with higher or lower probability. They then show how, using Harter's weights, species likely to represent a type can be identified. This represents a conflation of individualistic and Association views somewhat akin to the fuzzy sets used in the SNOB program.

Given the first three moments of the observed distribution of a species, we proceed as follows. Let the moments be m_1, m_2, m_3. Then define:

$$a = m_{12} + m_1 - m_2$$
$$b = m_3 + 2\,m_1 - 3\,m_2 - m_1\,(m_2 - m_1)$$
$$c = (m_2 - m_1) - m_1\,m_3$$

Now solve the equation

$$ax^2 + b.x + c$$

for x. Using the standard formula set

$$L = -b - \text{sqrt}\,(b^2 - 4ac)$$

and set M to the other root. L is the mean for the elite group, M that for the residual group. If the roots are complex, then set

$L = m_1$ and $M = 0$,

If $M < K$ then set

$L = (m_2 - m_1) / m_1$ and $M = 0$

The mixing coefficient h has the value

$h = (m_1 - M) / (L - M)$.

If $h < 0$ or $h > 1$ then set

$L = m_1$, $M = 0$, and $h = 1$.

The separation factor used by Harter is then calculated as

$(L - M) / (L + M)^{0.5}$.

If the mixing coefficient $h = 1$ then only a single population is present.

Van Rijsbergen *et al.* proceed by estimating the relationship between 'being elite' and 'being in a type'. The weight to be given to k occurences of a species in a site is then calculated from

$$w_k = -\log (p_k q_0)/(q_k p_0)$$

where

$p_k = pe^{-L}L^k + (1-p)e^{-M}M^k$

$q_k = qe^{-L}L^k + (1-q)e^{-M}M^k$

$p_0 = pe^{-L} + (1-p)e^{-M}$

$q_0 = qe^{-L} + (1-q)e^{-M}$

The value of q is estimated by the mixing coefficient *h*, while p is given values reflecting the stochastic relationship of the elite set of the particular species and the type. The value of p can be estimated only if we know the types; otherwise only plausibile values can be suggested. If p=1 then the elite set is identified with the type, which would possibly be acceptable in phytosociology, though not in information retrieval.

In the absence of other information it is also possible that some environmental indicator value could be used as an estimate of p. The resulting site weight would then represent the evidence for a particular environmental interpretation. The value given to p would reflect belief in the indicator value of the species for some particular environmental feature, a value estimated from other data. We hope to pursue this possibility in a later study. Our program permits the user to set values for p, but we have, in this study used only the single value 0.5, which Van Rijsbergen *et al. (1980)* found optimal for their purposes. The sensitivity of the w_k values to the choice of value for p also requires investigation.

Results

Our first question concerns the computational cost of the methods, since *inter pares* we should clearly choose the cheapest method to accomplish our task. The FOBINF and FOBSCP programs proved excessively expensive to use, with three solutions not being obtained at all due to excessive computer time requirements. More effective algorithms are required. The SNOB procedure is also expensive, but it does of course also provide a classification at the same time. FOBCAI uses an iterative procedure and is as expensive as the FOBINF calculations. In comparison the other methods were comparitively cheap, FOBCOR and FOBUNI being slightly more expensive than the others. FOBCOR produces all the correlation weights so this is a little unfair, but much of the calculation is duplicated whichever specific correlation values are used.

In terms of our computer, a CDC CYBER 76, these cheap methods occupied at most 5 secs, whereas the more expensive methods could use some hundreds of seconds! Note that a numeric classification of the full data required between 4 and 20 secs. depending on the specific method chosen.

All the programs except SNOB provide a ranking of the entire list of species. SNOB in fact rejects only 11 species, all rare, by noting that they can be regarded as invariant over the entire population. By examining the discriminatory lists for each class further species might be rejected, but this is a non-trivial task. We have therefore rejected SNOB as a convenient method for species reduction, which is not to be construed as a comment on its merits as a classification method.

We are left with 22 methods (see Fig. 1) which we apply to the first data set, to obtain the selection lists which are then further classified using the Levenshtein metric. We treat the results at the 3 and 5 group levels (Figs. 1-3). The groups are as follows:

1. A group based on composition, comprising Specificity, Entropy, Equivocation, the quantitative Uniformity measures and the Eident values. The Double Poisson is somewhat remote from the others, which can be seen in the Principal Coordinates analysis and the Additive similarity tree as well.

2. The group of *a posteriori* methods, Cramer-Bray-Curtis and Cramer-Niasm. This isolation suggests that none of the other selection procedures have emulated these methods at selecting discriminating species. If the objective of recovering patterns similar to those in the entire data is accepted, then none of the other selection procedures is apparently performing particularly well. They may of course be rejecting outliers and uncovering novel information but as yet we have no evidence for this.

3. A group based on correlation, which can be further subdivided into 3 subgroups.

a) Total, LL, LN and CAI-fd which all use pairwise correlation measures and all standardise data.
b) Joint entropy, CAI-u, NL and NN correlation and both Dahl's measures. These are again pairwise measures, except CAI-u, but standardisation appears to have had little effect on the two correlation measures.
c) A group composed of mutual and joint information which are unstandardised and somewhat unselective. They are remote from the *a posteriori* group and we have ignored them hereafter.

The ordination displays these patterns quite well using the first 2 axes. The third axis appears to capture some elements of a presence-abundance distinction but is not overly important. While it is not necessary to interpret these axes, which are simply display devices, it is perhaps comforting to show some kind of possible meaning. The 3 axes account for about 70% of the trace. The additive similarity tree also suggests three lines, corresponding with the three major groups, but the fit, as measured by Kruskal's (1964) S^2 stress measure, was not good (**c.** 22%).

We now select methods from each of groups 1,3a and 3b – these groups are clearly distinct. In the compositional group we chose Harter's method because of its extra information and, because this was a little atypical, we also included the Eident values. For the correlation methods we chose the total correlation as the simplest measure, but the Wong-Liu approach seemed to merit further examination because of the extra information on stand typicality, and because the two elements of which it is composed, CAI-u and CAI-fd, are widely separated. Thus we have 5 methods to examine in greater detail, to which we can add the *a posteriori* CRAMER methods.

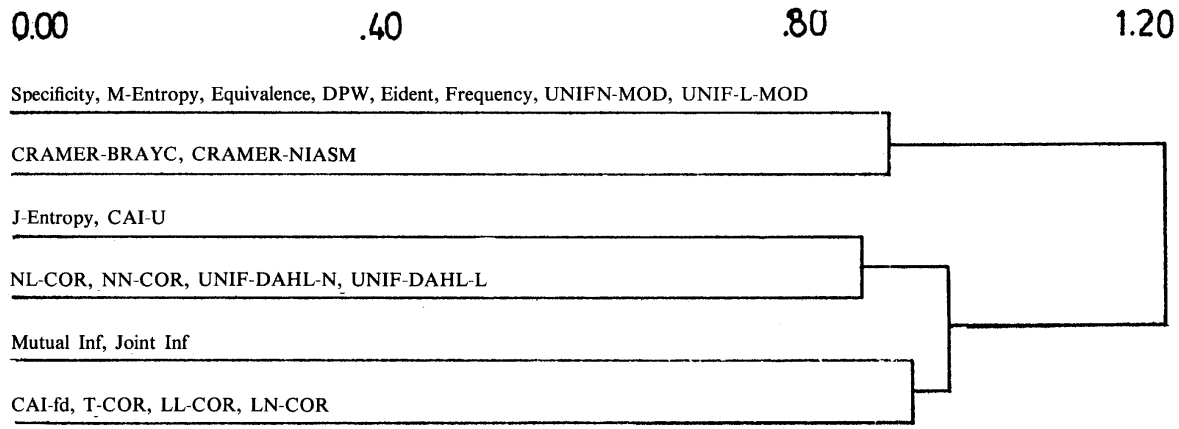
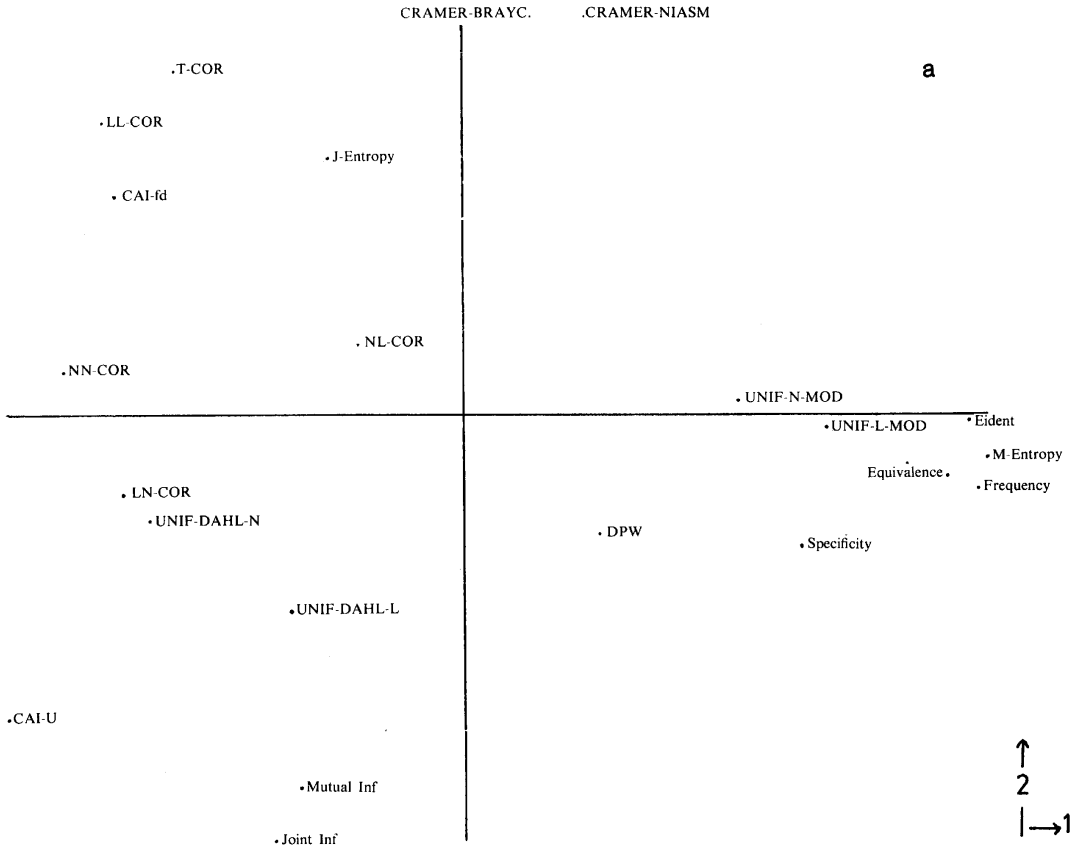


Fig. 1 Classification of the 22 methods. Abbreviations used in Fig. 1, Fig. 2 and Fig. 3:

- **Specificity**
- **Joint Entropy**
- **Mutual Entropy (M-Entropy)**
- **Equivalence information (Equivalence)**
- **Mutual information (Mutual Inf)**
- **Cramer values Bray-Curtis classification (CRAMER-BRAYC)**
- **Cramer values NIASM classification (CRAMER-NIASM)**
- **Double Poisson method (DPW)**
- **Wong-Liu method composition (CAI-u)**
- **Wong-Liu method feature dependence (CAI-fd)**
- **Eident values**
- **Frequency**
- **Uniformity extension presence data only (UNIF-L-MOD)**
- **Uniformity extension quantitative data (UNIF-N-MOD)**
- **Total correlation sums of squares (TCOR)**
- **Presence-presense correlation (LLCOR)**
- **Presence-abundance correlation (LNCOR)**
- **Abundance-presence correlation (NLCOR)**
- **Abundance-abundance correlation (NNCOR)**
- **Joint information (Joint Inf)**
- **Dahl Uniformity Size adjusted and presence-absence groups (UNIF-DAHL-L)**
- **Dahl uniformity presence only no size adjustment (UNIF-DAHL-M)**



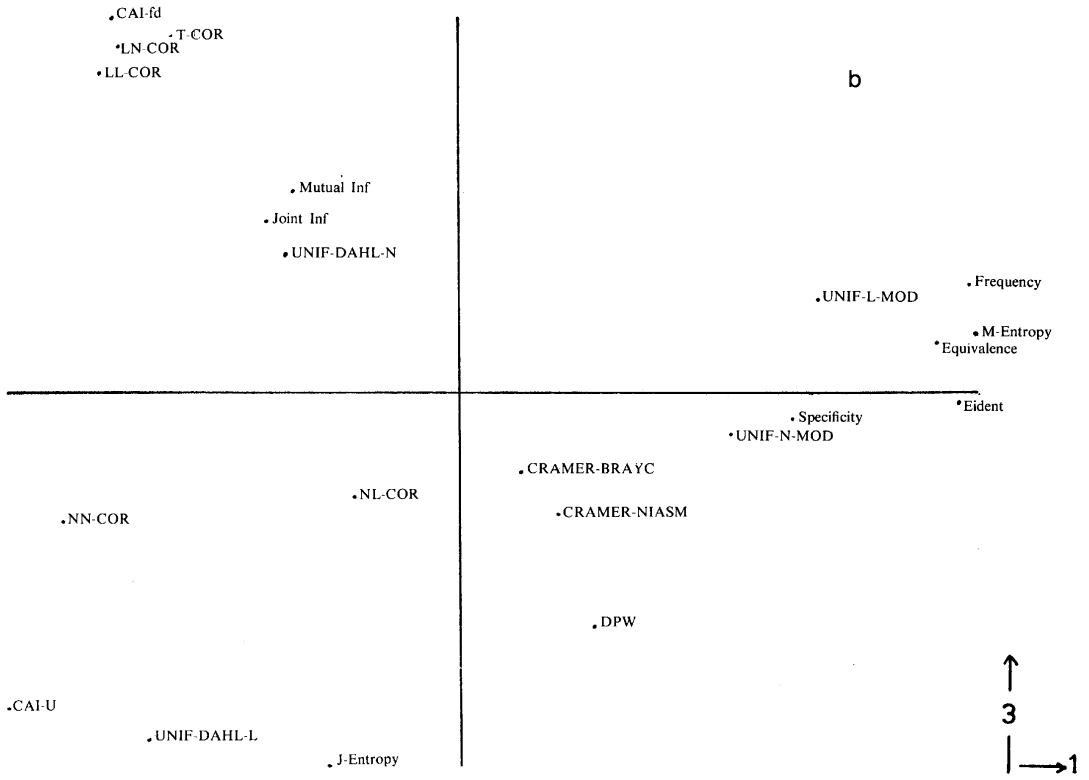


Fig. 2. Principal Coordinates analysis . a axes 1 & 2 . b axes 1 & 3

Efficacy of Reduction

For the first set of data we classified both the full set and the reduced sets and examined the correspondence between the results at all levels between 2 and 6 groups. To make this comparison we used both the RAND (1971) and the Fowlkes-Mallowes (1983) statistic; the RAND statistic proved much less sensitive and the results are not discussed here. Next we used Mantel's (1967) test to compare the matrices of Bray-Curtis similarities. For the second set of data, we ordinated using principal coordinates analysis, and compared results using a canonical correlation method.

First Set. The Mantel test results are shown in Table 3, and are largely unhelpful since all methods showed strong similarities to the full data. The significance levels are of course indicative only, although we used the weaker, non-parametric Chebyshev limits instead of the normal approximation. There is some indication that the Eident, TCOR and DPW results are more strongly related to the complete analysis than the CAI results, but that is about the limit of discrimination.

For the Fowlkes-Mallowes test we included Wallace's (1983) corrections but this had no effect on the results shown in Table 4. Again the significance tests are indicative only. In addition they can be calculated only if the number of groups being compared is equal. This is rather unsatisfactory as a simple example will show. Suppose 100 stands divide into 2 group of fifty relevés in one analysis. In another analysis they might first divide in 2 groups of 99 and 1, then the group of 99 divide appropriately into 2 groups of 49 and 50. At the 2 group level the result is particularly bad, but at the 3 group level for the second analysis the result will be near perfect concordance, and we have learnt of an outlying value. To investigate this possibility we show, in Table 5, the number of groups giving the maximal value to the Fowlkes-Mallowes statistic. We present both the value of the statistic itself and the number of groups at which this maximum was obtained. Tied values are found, indicating unclear relationships between the two sets of groupings, but it can be seen from Table 4 that both CAI-u and CAI-fd results are poor, with the others in the order Eident, DPW and TCOR. Table 5 strengthens this result, for the CAI methods show correspondence only at 2 or 3 group levels. The Eident method shows a maximum at 5, TCOR at 6, while DPW finds only 4. Using internal tests for number of groups, which are known to be rather weak but are all that are available, we obtain for the Ratkowsky-Lance test 2 groups for DPW and Eident, 4 for TCOR, while another test due to Mojena (1977) gives 5 groups for Eident, and 4 for both DPW and TCOR.

In Table 6 we present compositional correspondence of groups between the 4 group full data result, and various levels for the reduced sets; 5 for Eident, 4 for DPW. The TCOR result shows much less correspondence, but the other two show reasonable recovery of most of the

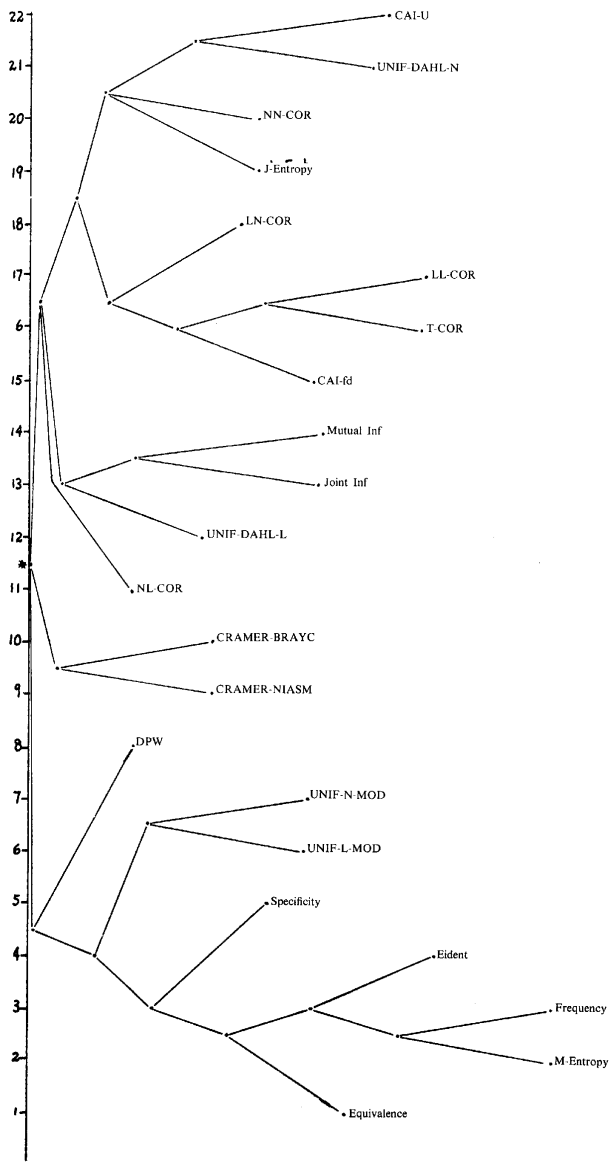


Fig. 3 Additive similarity tree. Note that line lenghts are significant in this analysis.

groups. The differences lie only in the shifting of a few relevés and the isolation of a single site in the Eident analysis. Thus for the first set of data we can conclude that Eident, TCOR or DPW are acceptable, with perhaps Eident and DPW being slightly better on the Fowlkes-Mallowes results and on simplicity of computation. TCOR appears to be identifying some different pattern, perhaps due to the standardisation of the data.

Second Set. The results are shown in Table 7. We record for each selection method the number of species obtained, and the effectiveness (% Trace) of the ordination using a 3 dimensional solution. Although none of these, except CAI-fd at 87%, shows remarkable recovery it was felt that inclusion of a large number of other axes would be confusing, especially given the known problems associated with nonlinearity and the lack of robustness of least squares fitting. In any case if the methods do not recover the most important axes they are likely to provide only confusing information, provided of course that the main axes are not dominated by outliers. We relate the 3-dimesional solutions to the solution for the full data, using canonical correlation, and in Table 7 we show the 3 correlations and the relationship of the canonical axes to the principal coordinates axes. We regard a coordinate axis as related to the canonical axis if a loading of >0.5 in magnitude was found.

Ideally we should recover the same axes in the same order. The Eident solution shows strong correlations and identifies the same space, but rotates it so that the first and third canonical axes are composites of the first and third coordinate axes. The CAI-fd result shows little correlation except for the first which relates the first canincal with the third coordinate, while CAI-u is apparently unrelated to anything. Significance tests, for what they are worth, suggest that both TCOR and CAI-fd are recovering only a single dimension, CAI-fd the third, TCOR the first. Such would perhaps be adequate if the weights were being used in a classification procedure with recalculation after each division.

The other good result is DPW, which shows correlations with 2 of the full data axes, and in fact identifies the same axes as the full set and in the same order. We lose some of the third dimension when using this reduction. Since we obtain other information from the DPW method, it would on these data seem reasonably effective without undue cost.

Given the perfomance of the DPW method, it is worth looking at the other information which it provides in more detail. We have also to se if the classes obtained on classification of the reduced sets reflect the subjective interpretations given in the original papers.

Additional Information and Original Structure

In Tables 8 and 9 we show the results for both data sets of the DPW method in more detail, together with the *Beta* values from the Eident solution. In addition for the first set we indicate the species most important at the 4-group level as assessed by the *a posteriori* CRAMER values. Since this last identifies sufficient species which can be rare, several other species attained high CRAMER values, but were not included in the reduced list, e.g. *Picris echioides* (0.96), *Rapistrum rugosum* (0.83), *Bupleurum tenuis* (0.79), and *Atriplex litoralis* (0.76). About half the species with CRAMER values exceeding 0.5 are included in the reduced list.

The question to be answered here is whether we can distinguish between constant and faithful species in these data. The *Beta* values reflect commonness and rarity in the entire data, but are not really very selective, while the CRAMER values show discriminating species and inspection of data tables for each group would allow constancy or fidelity to be determined. However the mixture column of the DPW results seems to differentiate quite well, with

Table 3. Results of Mantel test for similarity matrices

Reduced Set	Observed Value	Expected Value	St. Dev.	Chebyshev Probability
EIDENT	2817	5156	45.33	.00037
DPW	2755	4986	44.82	.00039
TCOR	3715	6862	59.67	.00036
CAI-u	1805	3306	77.75	.00268
CAI-fd	2758	5135	78.15	.00108

Table 4. Comparison of reduced and full analyses using the Fowlkes-Mallowes statistic

Reduced	Value	Number of Groups (equal in both). Entries × 1000				
Set		2	3	4	5	6
EIDENT	Obs.	1000	1000	994	804	635
	Exp.	615	395	378	304	255
	S.D.	17	16	20	16	10
DPW	Obs.	877	850	844	578	604
	Exp.	621	407	391	279	254
	S.D.	16	15	20	16	10
TCOR	Obs.	703	539	580	634	496
	Exp.	703	461	333	274	214
	S.D.	0	17	16	11	10
CAI-u	Obs.	582	606	558	479	406
	Exp.	597	492	463	411	369
	S.D.	14	14	18	15	10
CAI-fd	Obs.	548	507	489	427	385
	Exp.	525	440	406	360	312
	S.D.	9	12	16	18	12

Table 5. Maximal Fowlkes-Mallowes B_k statistics for comparing reduced and full data sets with unequal number of groups. Entries are maximal B_k values and number of groups in the reduced data set.

Reduced Set	Number of groups in full set.				
	2	3	4	5	6
EIDENT	100;2	100;3	99;4	100;5	77;4
DPW	88;2	85;3	84;4	72;4	72;4
TCOR	70;2	63;6	61;6	63;3/4/5	59;3/4/5
CAI-	60;3	61;3	59;3	56;3	52;3
CAI-fd	55;2	55;2	52;2	49;2	47;2

values close to 0 indicating 2 classes, those close to 1 a single population. The second set of data, as we might expect since it is a single Association, has many essentially constant species ($h = 1$), while the first set has many fewer mixture values of untiy, and many smaller values for other species. Thus there does seem to be some differentiation being made by the DPW method, with high mixture values indicating constant species, low mixture values faithful or perhaps discriminating species. Without a formal definition of the Association, or any other syntaxomic unit, it is not possible to be sure of the relationship of the elite class of any species to a syntaxon. For an abstract type it seems unilikely that any relevé will be truly representative, or that the elite class of a single species will of itself be a sufficient definition. Indeed the elite class is a stochastic concept to which relevés may only be assigned probabilistically. That an Association can be regarded as an idealisation to which real relevés partially belong is an ineresting notion which has implications for vegetation mapping (Dahl, pers. commun.).

For the first data set the comparison with the original structure is shown in Table 10. The DPW analysis seems

to accord best. Of the 12 samples of Ferrari (1971) all remain together, and the 6 sites of Ferrari and Grandi (1976) are also kept together, except in the Eident analysis where one is removed. The two sub-associations are relatively clearly distinguished, with some relevés shifting even in the full data analysis. The 52 sites of Ferrari and Galanti (1972) had not been assigned to Associations, but in most cases seem to associate with one or other of the defined subassociations. However some of the sites which are geographically close, as reported in the original descriptions, appear to form fragments of different Associations. The DPW result seems as good as the full data in assigning these 52, the Eident results deviating somewhat more. However it is also possible that some redefinition of the sub-associations is desirable.

For the second data set the analyses order the 13 relevés, so as to separate 1, 2, 4 and 6 from 10, 11, 12 and 13 on the first axis, and 7, 8, 9 and 10 from the remainder on the second axis. This simply indicates that there is variation within the syntaxon, but investigation of this would require more extensive data. Certainly Table 11,

Table 6. Comparison of partitions from reduced and full sets.

EIDENT

Reduced Groups	Groups from Full data analysis			
	1	2	3	4
1	12	–	–	–
2	–	1	–	–
3	–	5	–	–
4	–	–	4	49
5	–	–	58	–

DPW

Reduced Groups	Groups from Full data analysis			
	1	2	3	4
1	12	–	–	–
2	–	6	–	–
3	–	–	8	49
4	–	–	54	–

Table 7. Comparison of ordinations, second data set.

Subset Name	No. Sps.	% Var. retained in 3D.	Canon. corr. 1			Canon. corr. 2			Canon. corr. 3	
			Value	Axes Subset	Axes Full	Value	Axes Subset	Axes Full	Value	Axes Subset	Axes Full
Eident	28	49	.997	-1,-3	-1,-3	.99	2	2	.90	-1,3	1,-3
DPW	13	55	.99	1	1	.92	2	2	.29	3	3
TCOR	27	70	.94	1	-1,3	.71	-2	-1,-3	.17	3	-2
CAI-u	22	52	.55	3	3	.19	1,3	1	.01	2	-2
CAI-fd	21	87	.88	3	1	.48	2	2	.16	1	-3

Table 8. DPW analysis, first data set, with additional information, and CRAMER and Eident *Beta* values.

Species name	Elite	Resid.	Diff.	Mixture	Beta	CRAMER
Agropyron littorale	3.23	0	1.8	1	5.64	0.50
Spartium junceum	2.98	0	1.7	0	–	–
Agropyron repens	2.98	0	1.7	0	–	0.40
Artemisia cretacea	2.41	0	1.6	1	3.24	0.87
Daucus carota	2.04	0	1.4	1	2.48	0.72
Rumex crispus	1.98	0	1.4	0	–	0.40
Festuca rubra	1.98	0	1.4	0	–	–
Polygonum crispus	1.98	0	1.4	0	–	0.40
Dorycnium hirsutum	1.89	0	1.4	0.2	0.26	0.45
Atriplex patula	1.52	0	1.2	0.3	0.39	–
Hordeum marinum	1.51	0	1.2	0.4	0.58	0.66
Aster linosyris	1.5	0	1.2	0.5	0.78	0.56
Inula viscosa	1.5	0	1.2	0.7	1.08	0.77
Dactylis glomerata	1.4	0	1.2	1.0	1.53	0.82
Linum corymbulosum	1.4	0	1.1	0.2	–	0.50
Erythraea centaurium	1.3	0	1.2	0.4	0.44	0.55
Lactuca saligna	1.3	0	1.1	0.2	–	0.57
Blackstonia perfoliata	1.3	0	1.1	0.5	0.58	0.66
Picris hieracioides	1.2	0	1.1	0.6	0.66	0.63
Hypochaeris aetnensis	1.2	0	1.1	0.3	0.33	0.48
Puccinellia disticha	4.1	0.01	2.0	0	–	0.96
Scorzonera laciniata	1.0	0	1.0	0.3	0.25	0.51

Table 9. **DPW analysis, second data set, with additional information and *Beta* values.**

Species Name	Elite	Resid	Diff.	Mixture	Beta
Rhus coriaria	2.77	0	2.86	1	2.86
Prunus spinosa	2.08	0	1.44	1	1.84
Crataegus oxycantha	1.92	0	1.39	1	1.66
Hedera helix	1.69	0	1.30	1	1.39
Ligustrum vulgare	1.46	0	1.21	1	1.15
Paliurus spina christi	1.44	0	1.20	.16	1.20
Brachypodium pinnatum	1.31	0	1.14	1	1.00
Cercis siliquastrum	1.19	0	1.09	.26	.20
Buglossoides purpurocaerulea	1.19	0	1.09	.26	.20
Rubus caesius	1.08	0	1.04	1	.80
Galium mollugo	1.03	0	1.02	.52	.36
Acer campestre	1.0	0	1.0	1	.73
Asparagus acutifolius	1.0	0	1.0	1	.73

Table 10. **Full data, Eident and DPW groups for comparison with original relevé groupings. Numbers in parenthesis indicate number of members in the group.**

Full	Groups from original publications							
Groups	(12)	(6)	(25)	(34)	(13)	(13)	(13)	(13)
1 (12)	12	–	–	–	–	–	–	–
2 (6)	–	6	–	–	–	–	–	–
3 (49)	–	–	23	1	11	16	3	1
4 (62)	–	–	2	33	2	3	10	12
Eident	**Groups from original publications**							
Groups	(12)	6)	(25)	(34)	(13)	(13)	(13)	(13)
1 (12)	12	–	–	–	–	–	–	–
2 (1)	1	–	–	–	–	–	–	–
3 (5)	–	6	–	–	–	–	–	–
4 (42)	–	–	22	4	5	7	3	1
5 (69)	–	–	3	30	8	6	10	12
DPW	**Groups from original publications**							
Groups	(12)	(6)	(25)	(34)	(13)	(13)	(13)	(13)
1 (12)	12	–	–	–	–	–	–	–
2 (6)	–	6	–	–	–	–	–	–
3 (57)	–	–	25	6	11	10	4	1
4 (54)	–	–	–	28	2	3	9	12

which shows the coordinate axes for the DPW reduced set together with the canonical axes for the DPW and full sets, does not suggest that much information was lost by the reduction.

The last stage of the evaluation is siply to examine the species lists and try to determine if they incude the same species as would be selected by an experienced ecologist. This we have attempted to do, with 2 of us, Ferrari and Venanzoni, appraising the species lists before seeing the numerical results.

For the first data set Ferrari comments that he does not regard the 4 transects of 13 sites as typical of the Association to which they are assigned. The possibility of there being 2 slightly modified subassociations is thereby decreased, though not to be entirely rejected. In any case it was not the intent of this paper to attempt definitive segregation of phytosociological entities. The Eident and DPW methods, and perhaps the TCOR method, have given results which are succesful in identifying interesting species with this data set. DPW selects 22 species, Eident 19, which 14 are shared. The lists for both seem to contain most of the species regarded as interesting in these environments. Given the low costs involved, both analyses could well be applied if a stronger evaluation seemed necessary. Some of the species selected clearly reflect subgroups within the data while others reflect wider affinities; *Dactylis glomerata* and *Daucus carota* are widely distributed but remain good indicators here, as constant species. The mixture values of the DPW method do reflect the constancy / fidelity of the species, but in this data set the distinctions are relatively sharp anyway. The species selected are salt-tolerant, and typical of old-field regrowth, and they are

Table 11. Ordination results, second data set.

Relevé Number	DPW Coordinate		DPW Canonical		FULL Canonical	
	1	2	1	2	1	2
1	−20	11	−22	10	−22	−10
2	−23	5	−31	4	−35	−13
−17	3	− 9	12	− 2	−21	
4	−26	7	−21	5	−16	−11
5	− 9	6	− 5	6	− 7	12
6	−20	2	−19	11	−15	3
7	2	−18	0	−18	− 4	− 3
8	− 5	−39	− 8	−39	− 5	37
9	0	−24	6	−24	8	29
10	26	−18	32	−16	32	20
11	33	7	25	9	25	− 5
12	26	17	30	16	22	−24
13	33	19	23	21	17	−17

probably also related in terms of Noble and Slatyer's (1980) *vital attributes*.

For the second set, Venanzoni regards the selections as an adequate representation of important species. The vegetation is of anthropogenic orgin, and is placed in the Order Prunetalia spinosae, Class Rhamno-Prunetea. However it is not easily assigned to an Association or Alliance. This kind of vegetation is not well known in Central Italy (Pedrotti, 1982). Ten shrub species were accounted of special significance, but in the original study (Venanzoni unpublished) the herbaceous species had not been differentiated. If only shrub species are examined the Eident list is related significantly (X^2= 14.7, 1 degree of freedom, $p<0.001$) to the original list, but 2 new shrub species and 14 herbaceous species have been identified in addition. The DPW result is not significantly related ($X^2 = 2.26$, degrees of freedom=1, $p>0.5$) but adds only 5 herbaceous and the same 2 shrubs as the Eident analysis. Obviously such tests are dependent in part on the number of species retained, with the DPW retaining only 13 species, as against 25 for the Eident analysis. Of the 13, 11 are shared by both. The canonical correlation results show DPW to still be effective at recovering patterns.

It would seem that of the methods we have examined, the DPW serves both as an effective means of selecting a subset of species and also provides extra information which may aid in distinguishing constant and faithful species. The Eident method performs well, and sometimes better than the DPW, as a means of reducing the number of species. However it does not give so much extra information. The TCOR method uses standardised data and performs well, although more expensive to calculate. If weights were recalculated at each division of a classification then this would be a quite useful technique.

Our use of statistical methods to aid evaluation shows some of their weaknesses. Mantel's test was insensitive, the Fowlkes-Mallowes over-restrictive, and neither was strictly vital to the final evaluation. As soon as group comparison tables were available the measure for the degree of concordance became largely a matter of convenience. We could as easily have used information measures. The canonical correlation performed well enough but we attempted no substantial interpretation of the results. Such was not necessary here, but it is often a difficult task.

The propriety of using the results from the full data set as a target can be called in question, as we have noted earlier. We should not expect that the results obtained will be identical but until we can provide an unambiguous valorising procedure for interpretation we must rely on human judgement. By this standard both DPW and Eident performed well. We should also note the possible interest in site weights, and the environmental belief values associated with the extended DPW method. Both deserve more attention.

We can finally conclude that the Eident and DPW methods are both reasonably effective methods of aiding species reduction, with DPW perhaps adding sufficient extra information on constancy and fidelity to be preferred marginally. The statistical tests were of marginal use, except perhaps canonical correlation analysis, largely due to unrealistic null hypothesesand insensitivity. What is now needed is an examination of the usefulness of these methods especially in relation to robust and resistant analytic methods. With growing computer power we cannot acceptably rely on methods whose results may depend on a few outlying values.

Ackowledgements. This work was partially supported by Italian MPI (grant to C. Ferrari).

REFERENCES

BOOKSTEIN, A. and D. KRAFT, 1977. Operations research applied to document indexing and retrieval decisions. J. Assoc. Comput.Mach. 24: 418-427.

BOULTON, D.M. and C.S. WALLACE, 1973. An information measure for hierarchical classification. Comput. J. 16: 254-261.

DAHL, E. 1960. Some measures of uniformity in vegetation analysis. Ecology 41: 805-808.

DAHL, E., O. PRESTVIK and H. TOFTAKER, 1981. En kvantifiserung av karakerartbegrepet. Det kgl. Norske vidensk Abers Selskab Musei Botaniskes Ser. 1981-5, 215-233.

DALE, M.B. 1971. Information analysis of quantitative data. in: Patil, G.P., Pielou, E.C. and Waters, W.E. (eds.) *Statistical Ecology 3: Many species populations, ecosystems and systems analysis.* Penn. State Univ. Press. pps. 133-148.

DALE, M.B. and D.J. ANDERSON, 1973. Inosculate analysis of vegetation data. Austral. J. Bot. 21: 253-276.

DALE, M.B. and W.T. WILLIAMS, 1978. A new method of species reduction for ecological data. Austral. J. Ecol. 3: 1-5.

ESTABROOK, G.F., C.S. JOHNSON and F.R. MCMORRIS, 1976. A mathematical foundation for the analysis of cladistic character compatibility. Math. BioSci. 29: 181-187.

FEOLI, E. 1973. Un indice che stima il peso dei caratteri per classificazioni monotetiche. Gior. Bot. Ital. 107: 263-268.

FEOLI, E., M. LAGONEGRO and L. ORLÓCI 1984. Information analysis in vegetation research. Dr. W. Junk, den Haag. 143 pps.

FERRARI, C. 1971. La vegetazione dei calanchi nelle argille scagliose del Mt. Paderno. Not. Fitosoc. 6: 31-51.

FERRARI, C. and G. GALANTI, 1972. Specie indicatrici e struttura della vegetazione nei calanchi della Valle del Santerno (Bologna). Arch. Bot. Biogeogr. Ital. 48: 131-145.

FERRARI, C. and G. GRANDI, 1974. La vegetazione delle calanche nelle argille plioceniche della Valle del Santerno (Emilia-Romagna). Arch. Bot. Biogeogr. Ital. L 4-th series 20: 3-16.

FERRARI, C. and M. SPERANZA, 1976. La vegetazione delle salse di Nirano (Appennino Emiliano). Not. Fitosoc. 12: 1-18.

FISHER, R.A., A.S. CORBETT and C.B. WILLIAMS, 1943. The relation between number of species and the number of individuals in a random sample of an animal population. J. Ecol. 12: 42-58.

FOWLKES, E.B. and C.L. MALLOWES, 1983. A method for comparing two hierarchical clusterings. J. Amer. Statist. Assoc. 78: 553-569.

GABRIEL, K.R. and C.L. ODOROFF, 1983. Resistant lower rank approximation of matrices. in: J.E. Gentle (ed.) *Computer Science and Statistics; the interface.* North Holland, pps 304-307.

GOODALL, D.W. 1953. Objective methods for the classification of vegetation I. the use of positive interspecific correlation. Austral. J. Bot. 1: 39-63.

GOWER, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325-338.

GOWER, J.C. 1975. Generalized procrustes rotation. Psychometrika 40: 33-51.

GUTTMAN, L.L. 1953. Image theory for the structure of quantitative variates. Psychometrika 18: 277-296.

HARTER, S.P. 1975a. A probabilistic approach to automatic keyword indexing. 1. On the distribution of specialty words in a technical literature. J. ASIS 26: 280-289.

HARTER, S.P. 1975b. A probabilistic approach to automatic keyword indexing. 2. An algorithm for probabilistic indexing. J. ASIS 26: 290-299.

HILL, M.O. and H.G. GAUCH, 1980. Detrended correspondence analysis: an improved ordination technique. Vegetatio 42: 47-58.

HILL, M.O., R.G.H. BUNCE and M.W. SHAW, 1975. Indicator species analysis: a divisive polythetic method of classification and its application to a survey of native pine woods in Scotland. J. Ecol. 63: 597-613.

HOGEWEG, P. 1976. Iterative character weighting in numerical taxonomy. Comput. Biol. Med. 6: 199-211.

HUBERT, L.J. 1979. Generalized concordance. Psychometrika 44: 135-142.

JOHNSON, R.W. and D.W. GOODALL, 1979. A maximum likelihood approach to nonlinear ordination. Vegetatio 41: 133-142.

KRUSKAL, J.B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29: 1-27.

LAMBERT, J.M., S.E. MEACOCK, J. BARRS and P.F. M. SMARTT 1973. AXOR and MONIT: two new polythetic divisive strategies for hierarchical classification. Taxon 22: 173-176.

LANCE, G.N. And W.T. WILLIAMS, 1966. A generalized sorting strategy for computer classifications. Nature 212: 218.

LANCE, G.N. and W.T. WILLIAMS, 1971. Attribute contributions to a classification. Austral. Comput. J. 9: 128-129.

LU, S-Y. and K.S. FU 1978. A sentence-to-sentence clustering procedure for pattern analysis. IEEE Trans. Systems, Man and Cybernet. SMC-8: 381-389.

MACNAUGHTON-SMITH, P., W.T. WILLIAMS, M.B. DALE, and L. G. MOCKETT, 1964. Dissimilarity analysis: a new technique of hierarchical subdivision. Nature 202: 1034-1035.

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

MOJENA, R. 1977. Hierarchical grouping methods and stopping rules: an evaluation. Comput. J. 20: 359-363.

NOBLE, I.R. and R.L. SLATYER 1980. The use of vital attributes to predict successive changes in plant communities subject to recurrent disturbance. Vegetatio 43: 5-21.

ORLÓCI, L. 1973. Ranking characters by a dispersion criterion. Nature 244: 371-373.

ORLÓCI, L. 1976. Ranking species by an information criterion. J. Ecol. 64: 417-419.

ORLÓCI, L. 1978. *Multivariate Analysis in Vegetation Science* 2nd ed., Dr. W. Junk, den Haag.

PANKHURST R.J. 1978. Biological identification. The principles and practice of identification methods in biology. E. Arnold, London.

PEDROTTI, F. 1982. Les haies du Mont Fiegni (Camerino). Guide-Itineraire de l'excursion Internationale de phytosociologie en Italie centrale, Camerino 2-11 July. pps 316-319.

RAND, W.M. 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 66: 846-850.

RATKOWSKY, D.A. and G.N. LANCE, 1978. A criterion for determining the number of groups in a classification. Austral. Comput. J. 10: 115-117.

ROHLF, F.J. 1977. A note on the measurement of redundancy. Vegetatio 34: 63-64.

RIJSBERGEN, C.J. VAN, S.E. ROBERTSON and M.F. PORTER, 1980 New models in probabilstic information retrieval. Brit. Libraries R and D Rep. 5587, Computer Laboratories, Univ. Cambridge.

ROSS, D. 1983. TAXON Users manual, ed. P3B. CSIRO Div. Computing Res., Brisbane.

SATTATH, S. and A. TVERSKY, 1977. Additive similarity trees. Psychometrika 42: 319-345.

TRUNK, G.V. 1968. Statistical estimation of the intrinsic dimensionality of data collections. Inform. Control 12: 508-525.

WALLACE, D.L. 1983. Comments on a paper by Fowlkes and Mallowes. J. Amer. Statist. Assoc. 78: 569-576.

WATANABE. S. 1969. *Knowing and Guessing: a quantitative study of inference and information.* Wiley.

WEBB, L.J., J.G. TRACEY, W.T. WILLIAMS, and G.N. LANCE, 1967. Studies in the numerical analysis of complex rainforest communities II. the problem of species sampling. J. Ecol. 55: 525-538.

WILLIAMS, W.T. and M.B. DALE, 1962. Partitioned correlation matrices for heterogeneous data. Nature 196: 202.

WILLIAMS, W.T., M.B. DALE and P. MACNAUGHTON-SMITH, 1964. An objective method of weighting in similarity analysis. Nature 201: 426.

WILLIAMS, W.T. and J.M. LAMBERT, 1959. Multivariate methods in plant ecology. 1 Association analysis in plant communities. J. Ecol. 47: 83-101.

WONG, A.K.C. and T.S. LIU, 1975. Typicality, diversity and feature pattern of an ensemble. IEEE Trans. Comput. C-24: 158-181.

Manuscript received: May 85