COENOSES 4 (3): 163-168, 1989

A TECHNIQUE FOR SPECIES WEIGHTING AND ITS UTILITY IN DATA REDUCTION AND MINIMIZATION OF MISCLASSIFICATION [1]

S.S. Shaukat, Department of Plant Sciences, University of Western Ontario, London, Ontario, Canada

Keywords: Species weighting, Within species sum of squares, Data reduction, Misclassification.

Abstract: This paper describes a technique for ordering species based on their contribution to the total within species sums of squares. The effect of data reduction, accomplished by the omission of lowly ranked species account for the information on the underlying trends or group structure in vegetation. The ranking procedure suggested here can be incorporated in an iterative ranking-clustering algorithm in order to minimize misclassification, by excluding species with larger random components in relevé sub-sets. The implications of within species sum of squares as a ranking criterion are discussed in relation to some multivariate methods of data analysis.

Introduction

An optimal sampling strategy is of utmost importance to phytosociologists whose interest lies in the recognition of pattern in vegetation. A successful disclosure of vegetational pattern depends upon an appropriate choice of number and position of sampling units (Borman 1953, Bourdeau 1953, Lindsey *et al.* 1958, McNeill *et al.* 1977), as well as the size of sampling units (van Dyne *et al.* 1963, Taaffee 1979, Zeide 1980, Shaukat and Orlóci 1984). Due to logistic considerations, the phytosociologist often has to partition the sampling effort without sacrificing the underlying pattern information. In practice, increased efficiency can be attained by restricting sampling to a subset of the total number of species present in the landscape.

Species differ in their sensitivity of response to environmental factors and factor complexes, and thereby in their ability to contribute to the vegetational pattern. Furthermore, the environmental response of many species is inherently similar, giving rise to redundancy of pattern information. Several ranking methods have been proposed that can provide weights based on species importance. Highly ranked species from among the total set of species obtained in a pilot survey can then be chosen to alleviate the species data recording effort in the main study, and also to economize the subsequent data analysis without undue loss in the resulting information (Orlóci and Mukkattu 1973, Orlóci 1978. Jancey 1979).

A wide variety of ranking techniques have been proposed; they differ in the criteria of species importance, as well as underlying ranking philosophy. Some involve computationally simple weighting criteria such as arithmetic mean and standard deviation (Grigal and Goldstein 1971, Grigal and Ohmann 1975) or mean square contingency (Williams *et al.* 1954, Macnaughton-Smith *et al.* 1964), while others are based on maximal unique variance or information (Orlóci 1973, 1975, Rohlf 1977), multiple correlation (Gower 1967) or mutual information and related quantities (Orlóci 1978).

Regardless of the weighting criterion, species ranking is a matter of local relevance and the reduced species set, comprising highly ranked species, represents the non-random component of pattern information in a specific tract of vegetation. Extension of this concept can be utilized for minimizing random events in order to achieve a more refined definition of the underlying group structure (Jancey 1980). In hierarchic agglomerative clustering, at each successive level of clustering only highly ranked species in the subset of relevés are included in the analysis to reduce the risk of misclassification. Inclusion of lowly ranked species, i.e., species with a larger random component, would only obscure the pattern. Details of the clustering algorithm based on the locality of species importance approach appear in Jancey (1980).

This contribution describes a method of species weighting based on the partition of total within-species sum of squares. The applicability of the present weighting method for data reduction prior to cluster analysis or ordination is investigated, and its utility in the minimization of noise in the recognition of group structure is examined.

The ranking algorithm

Many ranking procedures incorporate some measu-

1 Paper presented at the 2nd CETA International Workshop on Mathematical Community Ecology, Gorizia, Italy; 19-25 November 1988.

re of species interaction in their algorithm; in some this information actually determines the ranking of species. Co-variation of species is of fundamental importance when the eventual aim of the user is a formal ordination. On the other hand, if the final objective is typification or identification then a ranking algorithm based on independent pattern information of species is indispensable, as it permits increased discrimination. The algorithm described here ranks the species solely with respect to their unique component of pattern information.

Let X_{ij} be an element in the raw data matrix **X** with p species (rows) and n relevés (columns). First, partition the total sum of squares in the data matrix into two independent components, a within species sum of squares and a between species sum of squares:

$$\sum_{i=1}^{p}\sum_{j=1}^{n}(X_{ij}-\bar{X}..)^2 = \sum_{i=1}^{p}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 + \sum_{i=1}^{p} n(\bar{X}_i-\bar{X}..)^2$$

where $\bar{X}..$ is the grand mean and $\bar{X}_i$ equals the mean of ith species. The first term to the right is the total within species sum of squares. Species are ranked in accordance with their contribution to the total within species sum of squares, represented here as Q_i. For the ith species this would be:

$$Q_i = \sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 \;/\; \sum_{i=1}^{p}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2$$

Next, the values of Q_i are ranked (Q_j) and a cut-off level C is specified to obtain the reduced species list consisting of m species that stand above the cut-off point:

$$C \geqslant \sum_{j=1}^{m} Q_i \;/\; \sum_{i=1}^{p} Q_i$$

The greatest advantage of using the above ranking approach is that it is directly compatible with some important numerical techniques of data analysis, such as minimum variance (sum of squares) method of hierarchical agglomerative clustering (Ward 1963, Orlóci 1967) and Jancey's k-means method of non-hierarchical clustering (Jancey 1966) which are based on total within group sum of squares. In addition, the present ranking method also appears to be compatible with distance-based informal ordination methods such as Kruskal's nonmetric multidimensional scaling (Kruskal 1964a, b).

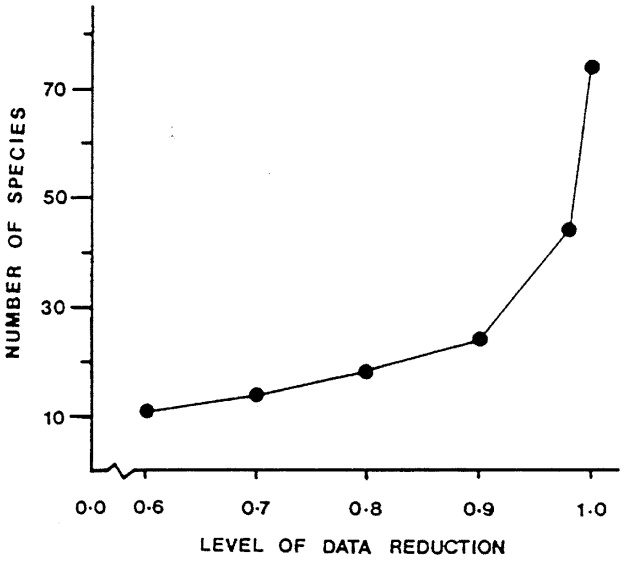

Fig. 1. The number species retained at various cut-off levels using species sum of squares ranking criterion. CD = complete data.

Applications

The data set

The vegetation data set that will be used to test the applicability of the present ranking method comes from pilot sampling of urban weed communities of London, Ontario. A stratified multistage sampling was performed using a plot size of 2.5x2.5 m. The data consists of 30 relevés and 74 species for which relative abundance was estimated by a modified Braun-Blanquet cover-abundance scale (van der Maarel 1979).

Effect of species reduction on group structure

The 74 species data set was reduced using the within species sum of squares criterion at cut-off levels of 0.6, 0.7, 0.8, 0.9 and 0.98. Fig. 1 shows the number of species retained at different cut-off levels. Clustering was performed on complete and reduced data sets using a minimum variance technique (Orlóci 1967) and Euclidean distance as the resemblance function. Topology matrices (Phipps, 1975) and cophenetic matrices (Sneath and Sokal 1973, Rohlf 1974) for each of the resulting dendrogram were computed. The dissimilarity matrices, topology matrices and cophenetic matrices derived from reduced data sets were compared with those based on complete data set using the correlation coefficient, r (Sokal and Rohlf 1962) and $\Delta_\mu(D_1, D_2)$, of Jardine and Sibson (1968).

The dendrograms resulting from minimum variance clustering of complete and reduced data sets are presented in Fig. 2. A marked degree of similarity between dendrograms based on the complete data set and those based on the reduced data sets is readily appa-

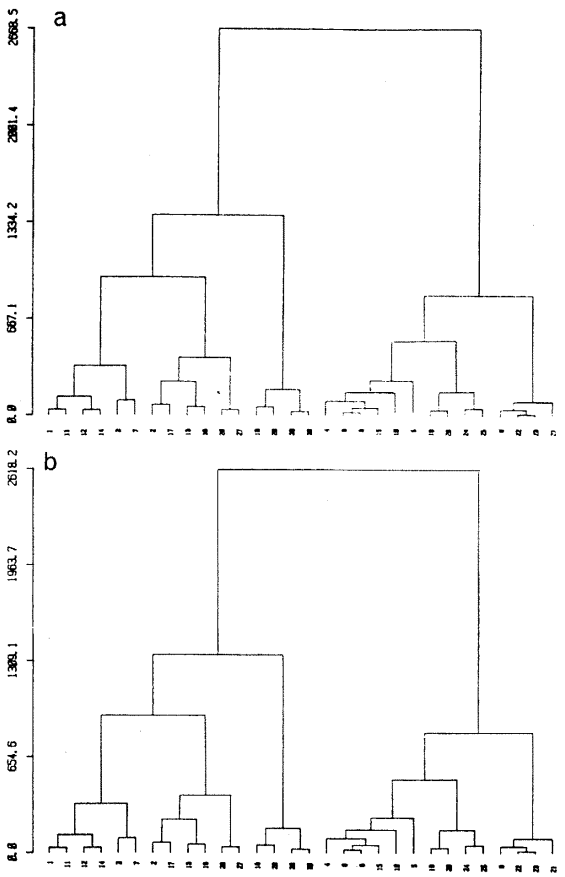

Fig. 2. Dendrogram resulting from minimum variance clustering of complete and reduced data sets obtained at various ranking cut-off points. a) Complete data, b-f) reduced data sets obtained at various cut-off levels: b, 0.98; c, 0.9; d, 0.8; e, 0.7; f, 0.6. Numbers at the base identify releve numbers, while vertical axis represents total within group sum of squares.

rent. Fig. 3 provides an objective comparison of Euclidean distance matrices of reduced data sets with that derived from the complete data set. The correlation between the distance matrix of the complete data set and that of the reduced data sets declines slightly with decreasinc cut-off levels. However, even at a cut-off level of 0.6, corresponding to a subset of 11 species, the correlation coefficient equals 0.858. Jardine and Sibson's Δ_{μ} (D_1, D_2), which measures the degree of disagreement between the matrices rather than similarity, shows progressively reduced levels of disagreement between dissimilarity matrix of complete data and that of reduced data sets with the increasingly higher cut-off levels of ranking.

Fig. 4 shows the comparison of the cophenetic matrices of complete data set to those of reduced data sets derived from minimum variance clustering. High levels of correlation indicate a great degree of similarity between the dendrograms derived from clustering of reduced data sets with that developed from the comple-

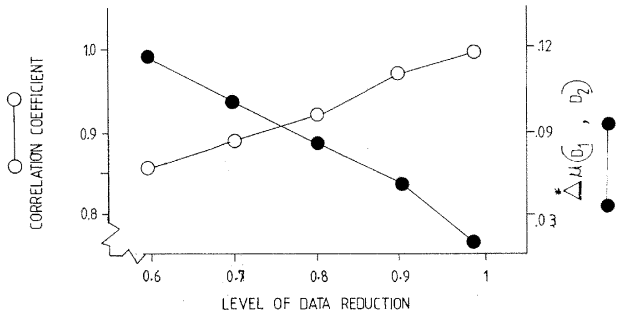

Fig. 3. Degree of similarity (correlation coefficient) and disagreement (Δ_μ (D 1, D 2)) between the Euclidean distance matrix of complete data set and that of reduced data sets obtained at various ranking cut-off points.

te data set. Though the level of r declines slightly and steadily with the reduction in the ranking cut-off point, even at the species reduction level of 0.6 the value of r is 0.9775. The disagreement function shows a declining trend (Fig. 4) with increasing cut-off point. However, consistent low values of the disagreement function indicate a great degree of correspondence between the cophenetic matrices derived from reduced data sets and that corresponding to complete data set.

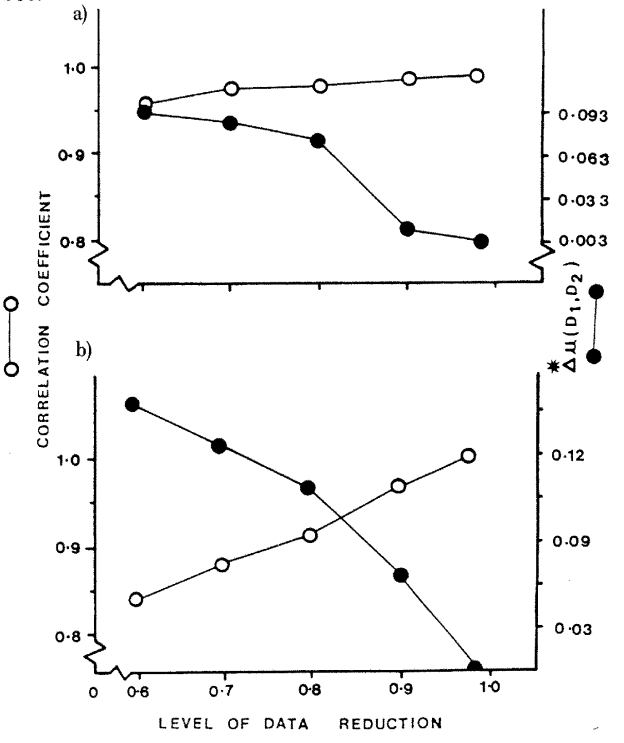

Fig. 4a. Degree of similarity (correlation coefficient) and disagreement (Δ_μ (D 1, D 2)) between the cophenetic matrix of complete data set and that of reduced data sets obtained at various ranking cut-off points.

Fig. 4b. Degree of similarity (correlation coefficient) and disagreement (Δ_μ (D 1, D 2)) between the topological matrix of complete data set and that of reduced data sets obtained at various ranking cut-off points.

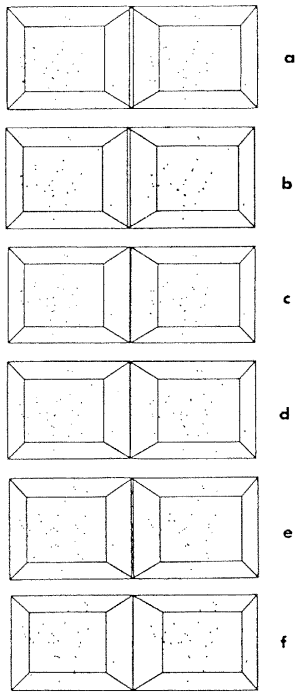

Fig. 5. Stereograms based on NMDS ordinations of complete and reduced data sets obtained at various ranking cut-off points. a) complete data, b-f) reduced data at various cut-off levels: b, 0.98; c, 0.9; d, 0.8; e, 0.7; f, 0.6.

Fig. 4b shows a steady decline in the level of correlation coefficients between the topology matrices derived from clustering of reduced data sets and that derived from complete data set, but consistently high values of r indicate a high order of similarity between the dendrograms. Jardine and Sibson's measure of disagreement shows a mirror image trend to that shown by the correlation coefficient.

Effects of species reduction on relevé ordination

Reduced data sets were obtained at cut-off levels of 0.98, 0.9, 0.7 and 0.6 using the within species sum of squares ranking criterion. Ordinations were performed on complete and reduced data sets using Kruskal's nonmetric multi-dimensional scaling (NMDS). Relevé configurations were obtained for 3-dimensions and represented as stereograms (Fig. 6). In general, the distribution of relevés in the stereogram for complete data set matches remarkably well with the configuration based on reduced data sets. The objective comparisons are made in Table 1 that gives product-moment correlation coefficients (r) between the pairwise relevé distances in the ordination of complete data set with those for reduced data sets in X, XY and XYZ ordination spaces. Note that the correlation coefficient is used here as an index of similarity rather than as a statistic and hence no significance level is attached. The values of r are generally high, indicating a high level of similarity bet-

Table 1. Degree of similarity (r) between MDSCAL ordination of complete data set and that of reduced data sets. Data reduction was performed by total within species sum of squares ranking.

Date sets compared	Number of dimensions used		
	1	2	3
CD* and 11 spp. (0.6)+	0.88645	0.870805	0.939265
CD and 14 spp. (0.7)	0.94616	0.934553	0.964844
CD and 18 spp. (0.8)	0.978202	0.960259	0.974728
CD and 24 spp. (0.9)	0.985994	0.974302	0.985698
CD and 44 spp. (0.98)	0.991923	0.982578	0.991579

*CD = complete data set with 74 species.
+Numbers in brackets indicate the ranking cut-off points (proportion of the sum of within species sum of square for the species above cut-off point to the total within species sum of squares of all species).

ween the ordination based on complete data set and those based on reduced data sets. Degree of similarity appear to be higher when the ordinations are considered in 3-dimensions (XYZ space) rather than one or two-dimensions. Degree of similarity between ordinations slightly decreased as more and more species were eliminated from the analyses.

Minimization of misclassification

As pointed out earlier, in order to minimize misclassification in a hierarchical clustering, species displaying locally a greater random component should be excluded from the clustering procedure at the corresponding clustering level. Thus, at each clustering pass, the total species set should be subjected to the ranking procedure since species that are unimportant in the complete relevé set may attain significance in a sub-set of relevés being used for sub-clustering.

An iterative species ranking and agglomerative clustering technique developed by Jancey (1980) was used to reduce the risk of misclassification. The technique incorporated the within group sums of squares clustering (minimum variance method) and the within species sum of squares ranking method which is directly compatible with the former algorithm.

Fig. 7 shows a dendrogram derived from the iterative agglomerative clustering algorithm. Comparison of this dendrogram with that based on the complete data set (Fig. 1a) reveals that the group structure in the former is much more compact, with lower total within group sum of squares, than in the latter. Furthermore, relevés 2 and 13, which were members of group 2 in the hierarchical clustering based on total species set, fall in group 1 in the iterative clustering. Examination of various ordinations, average linkage clustering and environmental data corresponding to relevés all indicated that relevés 2 and 13 were misclassified in the agglomerative classification based on all species. In order to test the affiliation of relevés 2 and 13 with either group 1 or 2, Mahalanobis' generalized distance (D^2) (Rao, 1952) was computed between the relevés and the groups 1 and 2. The D^2 values between group 1 and relevés 2 and 13 were 2.98 and 4.12 respectively as opposed to 63.05 and 36.65 between group 2 and relevés 2 and 13 respectively, indicating that the two relevés were better placed in group 1 as in the iterative clustering procedure.

Discussion and conclusions

This paper describes a computationally simple technique of species ranking based on the contribution of species to the total within species sum of squares. The technique differs from many other ranking techniques (e.g. Orlòci 1973, 1975, Gower 1976, Rohlf 1977) in that the present method is exclusively based on the independent pattern information contained in the species and disregards the shared information. In a recent review, Jancey (1979) concluded: "The decision whether to rank on maximal shared variance as opposed to specific variance has not been resolved". He pointed out that: "Perhaps more important than the question of which is fundamentally correct is the need to match the ranking philosophy to that of subsequent data analysis". In this regard the ranking method based on the partition of total within species sum of squares is exactly compatible with one of the most popular clustering method, namely the minimum variance method (Orlóci 1967), which minimizes the total within group sum of squares. Besides, the ranking method proposed here is also congrous with the k-means method of non-hierarchical clustering, in which total within species sum of squares provides the basis for cluster detection (Jancey 1966).

When minimization of misclassification is the objective, Jancey's iterative clustering using minimum variance clustering and the within species sum of squares ranking criterion will not only minimize "noise" in unfolding group structure but will also provide internal consistency and compatibility in the iterative clustering.

One advantage of such a consistency in the algorithm is that at each clustering level the ratio of the total within species sum of squares of the reduced species set to that of complete species set is a constant ratio which is equal to the ranking cut-off level used. Thus, despite the changing role of species in pattern formation in the different sub-sets of the study area, together the ranked species represent a correspondent non-random component of pattern information as a consistent proportion of the total information in the sample sub-set. Such a consitency and compatibility is unique to the presently described ranking method when used in conjunction with the minimum variance method for itera-

tive clustering.

An examination of the effect of species reduction on the disclosure of group structure strongly indicated that a relatively small proportion of the total species present in a study area accounts for the information on the underlying group structure in vegetation. There was little change in the classification hierarchies derived from successively reduced sub-set of species, obtained at various ranking cut-off points based on within species sum of squares ranking criterion. This corroborates and extends the findings of Orlóci and Mukkattu (1973) who found little distortion in the resemblance structure of the relevé sample even when the number of species were drastically reduced, though their species sub-sets were derived from covariance ranking.

Relevé ordinations based on sub-sets of species obtained at successively lower cut-off levels also showed close correspondence with that resulting from the full species set. The ordination method chosen was Kruskal's non-metric multi-dimentional scaling (NMDS) to achieve some degree of compatibility between the present ranking method and the ordination algorithm. The fact that NMDS works on proximities (or distances) between objects, and that the sum of squares divergence between two objects is a function of distances, provides the basis for compatibility between the ranking and the ordination method. In the case of eigenvector methods of ordination (viz., principal component analysis, reciprocal ordering), however, co-variation of species is of paramount importance and in that situation species ranking on the basis of covariance or shared information would be more desirable.

Apart from the utility of the within species sum of squares ranking for data reduction prior to both classification and ordination and in the minimization of misclassification, the ranking method requires minimal computational time. The ranking program (RANKO5) required only 1.85 c.p.u. seconds for the data set used in the numerical examples above on a P.D.P. 10 computer with a KL10D central processor. Listings of the computer program written in BASIC are available from the author.

REFERENCES

BORMAN, F.H. 1953. The statistical efficiency of sample plot size and shape in forest ecology. Ecology 34: 474-487.

BOURDEAU, P.F. 1953. A test of random versus systematic ecological sampling. Ecology 34: 499-512.

DYNE, G.M. VAN, W.G. VOGEL and H.G. FISSER. 1963. Influence of small plot size and shape on range herbage production estimates. Ecology 44: 746-759.

GOWER, J.C. 1967. A comparison of some methods of cluster analysis. Biometrics 23: 623-637.

GRIGAL, D.F. and R.A. GOLDSTEIN. 1971. An integrated ordination classification analysis of an intensively sampled oak-hickory forest. J. Ecol. 59: 481-492.

GRIGAL, D.F. and L.F. OHMANN. 1975. Classification, description and dynamics of upland plant communities within a Minnesota wilderness area. Ecol. Monogr. 45: 389-407.

JANCEY, R.C. 1966. Multidimensional group analysis. Aust. J. Bot. 14: 127-130.

JANCEY, R.C. 1979. Species weighting: a heuristic approach. *In:* L. Orlòci, C.R. Rao, and W.M. Stiteler (eds.). *Multivariate methods in Ecological Work,* pp.87-100. International Co-operative Publishing House Fairland.

JANCEY, R.C. 1980. The minimisation of random events in the search for group structure. Vegetatio 42: 99-101.

JARDINE, N. and R. SIBSON. 1968. The construction of hierachic and non-hierarchic classifications. Comput. J. 11: 177-184.

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

KRUSKAL, J.B. 1964b. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29: 115-129.

LINDSEY, A.A., J.D. BARTON, Jr. and S.R. MILES. 1958. Field efficiencies of forest sampling methods. Ecology 39: 428-444.

MAAREL, E. VAN DER. 1979. Transformation of cover abundance values in phytosociology and its effect on community similarity. Vegetatio 39: 97-114.

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

MCNEILL, L., R.D. KELLY and D.L. BARNES. 1977. The use of quadrat and plotless methods in the analysis of the tree and shrub component of woodland vegetation. Proc. Grassld. Sco. 5th. Afr. 12: 109-113.

ORLÓCI, L. 1967. An agglomerative method for classification of plant communities. J. Ecol. 55: 193-206.

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

ORLÓCI, L. 1975. Measurement of redundancy in species collection. Vegetatio 31: 65-67.

ORLÓCI, L. 1978. Ranking species based on the components of equivocation information. Vegetatio 37: 123-125.

ORLÓCI, L. and M.M. MUKKATTU. 1973. The effect of species number and type of data on the resemblance structure of a phytosociological collection. J. Econ. 61: 37-46.

PHIPPS, J.B. 1975. Dendrogram topology: nomenclature. Can. J. Bot. 53: 2047-2049.

RAO, C.R. 1952. *Advanced Statistical Methods in Biometric Research.* Wiley, New York.

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

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

SHAUKAT, S.S. and L. ORLÓCI. 1984. Plot size influence on trend seeking and group recognition in phytosociological samples. Abst. Ontario Ecol. and Ethol. Colloq., Waterloo, Ont., April, 1984.

SNEATH, P.H.A. and R.R. SOKAL. 1973. *Numerical Taxonomy.* Freeman, San Francisco.

SOKAL, R.R. and F.J. ROHLF. 1962. The comparison of dendrograms by objective methods. Taxon 11: 33-40.

TAAFFE, K.E. 1979. Computing optimum plot size for wildland inventories. Resource Inventory Notes (Denver) USDI BLM 23: 8-15.

WARD, J.H. 1963. Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assoc. 58: 236-144.

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

ZEIDE, B. 1980. Plot size optimization. Forest Sci. 26: 251-257.

Manuscript received: March 1989