Journal of Vegetation Science 16: 497-510, 2005

Multivariate exploratory analysis of ordinal data in ecology: Pitfalls, problems and solutions

Podani, János

Department of Plant Taxonomy and Ecology, Eötvös University, Pázmány P. s. 1/c, H-1117 Budapest, Hungary; Fax: +36 13812188; E-mail podani@ludens.elte.hu

Abstract

Questions: Are ordinal data appropriately treated by multivariate methods in numerical ecology? If not, what are the most common mistakes? Which dissimilarity coefficients, ordination and classification methods are best suited to ordinal data? Should we worry about such problems at all?

Methods: A new classification model family, OrdClAn (*Ord*inal *Cl*uster *An*alysis), is suggested for hierarchical and non-hierarchical classifications from ordinal ecological data, e.g. the abundance/dominance scores that are commonly recorded in relevés. During the clustering process, the objects are grouped so as to minimize a measure calculated from the ranks of within-cluster and between-cluster distances or dissimilarities.

Results and Conclusions: Evaluation of the various steps of exploratory data analysis of ordinal ecological data shows that consistency of methodology throughout the study is of primary importance. In an optimal situation, each methodological step is order invariant. This property ensures that the results are independent of changes not affecting ordinal relationships, and guarantees that no illusory precision is introduced into the analysis. However, the multivariate procedures that are most commonly applied in numerical ecology do not satisfy these requirements and are therefore not recommended. For example, it is inappropriate to analyse Braun-Blanquet abudance/dominance data by methods assuming that Euclidean distance is meaningful. The solution of all problems is that the dissimilarity coefficient should be compatible with ordinal variables and the subsequent ordination or clustering method should consider only the rank order of dissimilarities. A range of artificial data sets exemplifying different subtypes of ordinal variables, e.g. indicator values or species scores from relevés, illustrate the advocated approach. Detailed analyses of an actual phytosociological data set demonstrate the classification by OrdClAn of relevés and species and the subsequent tabular rearrangement, in a numerical study remaining within the ordinal domain from the first step to the last.

Keywords: Classification; Clustering; Dissimilarity; Multidimensional scaling; Non-metric; Ordinal measure; Partition.

Abbreviations: *AD* = Abundance/Dominance; *CL* = Complete Link; *DC* = Coefficient of Discordance; *ED* = Euclidean distance; *O* = Ordinal; *M* = Metric; *NMDS* = Non-metric Multidimensional Scaling; *OC* = Ordinal Clustering; *SL* = Single Link; *UPGMA* = Unweighted Pair Group Method or Group Average Clustering.

Introduction

Attention in numerical ecology is often focused on sequences or orderings. Sampling is the first phase of study in which an ecologist may be concerned with ordered information, for example, when phenological sequences, tolerance regimes, toxicity levels, and abundance/dominance relationships among species are assessed. Such observations can be summarized in terms of ordinal variables (Anderberg 1973). These are simpler than ratio scale ('quantitative') variables whose measurement requires more precision and effort. The possibility of defining a rank order arises again when expressing similarity and dissimilarity relationships among study objects or variables. One may decide that, no matter how detailed the raw data are, only rank values are to be used in calculating the similarity of objects (e.g. rank correlation, Legendre & Legendre 1998). This decision may be justified, for example, on the grounds that the original data are outcomes of imprecise measurement, so that only their rank order is reliable. In a later stage of the study, having computed a symmetric matrix of pairwise distances or similarities, one may decide to forget about the actual resemblance values and replace them by ranks before further processing (Clarke 1993, 1999). Given a matrix of Euclidean distances (*ED*) or other metric measures, the user may instruct the computer to perform a multivariate analysis of objects such that the metric properties are disregarded and the sequence of distance values is emphasized during the computations. Non-metric multidimensional scaling (NMDS, Kruskal 1964) is a case in point. The graphical illustration of results may also follow a similar logic. For example, in displaying a hierarchical classification by a ranked dendrogram (Lapointe & Legendre 1991; Podani 2000b) we forget about the values (weights, e.g. dissimilarities) assigned to the levels and interpretation is restricted to the fusion sequence of clusters.

The above examples sufficiently illustrate that ordinal properties are potentially present in any stage of a numerical ecological study. However, ecologists often fail to consider which analysis methods are appropriate for their ordinal data, notwithstanding Dale's (1989)

recommendations. There are some pitfalls that the ecologist must be aware of when making decisions along the complex pathways of exploratory data analysis. It is therefore worth examining in more detail the ordinal properties emerging in the different steps of ecological surveys. Particular attention must be paid to the validity of possible decisions made when passing information from one stage to the other. I put forward some general recommendations that may be considered by numerical ecologists in the preparation of study designs. First a short review of existing methodology is presented, with emphasis on the somewhat neglected methods of ordinal classification. A new ordinal clustering procedure is introduced that may serve as a simple alternative to existing multivariate techniques utilizing ordinal information in the data. For consistency, I shall use the term '*ordinal*' in the widest sense of the word, as a reference to any mathematical construct (a variable, a resemblance function, a classification etc.) or a method (scaling, clustering) that relies upon linear ordering of certain items (data values, coefficients, objects, clusters) and does not consider other types of relationships, especially products and ratios. Although 'ordinal' is often considered synonymous with '*ranked*' (Critchlow 1985; Dale 1989), this second term is applied here in a more restrictive manner to sequences whose n members are numbered by consecutive integers from 1 to n or by their mean values if some positions are tied. The term '*non-metric*' is also understood sometimes as an alternative to ordinal (e.g. Peay 1975), although it has a more general meaning: all operations and properties disregarding differences and ratios of data values are in fact non-metric. Thus, we have the following relations of inclusion: ranked data are a particular type of ordinal data, and ordinal data are a particular type of non-metric data (i.e. non-metric $\supset$ ordinal $\supset$ ranked), which govern the usage of these attributes throughout the paper.

Methodological pitfalls

In order to expand the topic outlined in the first paragraph of the Introduction, let us consider the conventional methodological sequence implied by exploratory analyses of ecological data, namely

$$\text{SAMPLING} \rightarrow \text{DATA} \rightarrow \text{RESEMBLANCE} \rightarrow \begin{matrix}\text{CLASSIFICATION} \\ \text{ORDINATION}\end{matrix}$$

Each move between these levels of abstraction can be characterized by whether ordinal (*O*, i.e., sequence- or rank-based) or metric (*M*, i.e., difference/ratio-based) information is considered. Associated to these sequences are some typical examples often encountered in the literature of numerical ecology:

Sequence 1

The sampling process yields abundance estimates such as percentage cover, counts or biomass; these quantitative data are used to calculate metric resemblance measures but then the investigator might decide to reduce the analysis to the ordinal level in order to exploit the advantages of non-metric multidimensional scaling (e.g. arch effect diminished, gradient recovery enhanced, solution forced into a user-specified number of dimensions, and linearity constraints released). In fact, the majority of NMDS applications to ecological data could be cited as examples.

This series, whose stages are denoted by *M-M-O*, is perhaps the least problematic, but all scientists must be aware of the fact that information is lost in the last step when actual differences between distances become neglected. Surely, data collection was too precise, compared to the precision implied by the latest stage of the analysis (Gill & Tipper 1978). There is always a possibility therefore that in such cases lower sampling effort by recording simpler types of data could have led to similar results and conclusions.

Sequence 2

Relevé data comprising Braun-Blanquet-type abundance/dominance (*AD*) scores commonly used in phytosociology (Mueller-Dombois & Ellenberg 1974) or other ordinal data formats mentioned in the Introduction are converted to metric resemblance measures by, e.g. the Euclidean distance function, and then the analysis becomes ordinal again by applying NMDS to achieve the same goals as above.

Perhaps the worst-case scenario, *O-M-O*, is represented by this series. The distances, even if they are computed formally by a metric coefficient, are based on variables actually measured on the ordinal scale to which arithmetic operations do not apply (cf. Krauth 1986; Dale 1989). Furthermore, metrizing ordinal information introduces illusory precision to the analysis, and this new 'metric' information in the distances is largely ignored again in the final NMDS anyway.

Sequence 3

From the same starting point as in 'Sequence 2', the distance matrix is processed by Ward's (incremental sum of squares) clustering method or principal coordinates analysis, procedures implicitly assuming that the data space is Euclidean. Another, perhaps more common, example is that the raw data matrix is taken as a contingency table and then some form of correspondence analysis is used to obtain a simultaneous ordination

of species and sites.

This *O-M-M* sequence is also very problematic because of the same illogical distance conversion as in 'Sequence 2', so that the subsequent metric clustering or ordination procedure can be no more than a self-deceptive attempt to preserve a 'metric' structure that does not exist. Data collection is the weakest point in studies like this and the preceding one: the ecologist deliberately disregards (often very large) actual differences when recording ordinal scores. Then, it may not be justifiable that differences between distances obtained from such variables are taken as seriously as is implied in the use of sum of squares clustering or correspondence analysis.

Similar criticism applies to a closely related sequence, *O-O-M*, in which ordinal coefficients are subjected to metric analysis.

Sequence 4

Distances or dissimilarities calculated from abundance data are replaced by their ranks to obtain a new symmetric matrix. Then, the matrix of these converted measures is subjected to metric clustering or scaling methods, for example, those mentioned in 'Sequence 3', to obtain a classification or an ordination.

This methodological path implies an *M-O-M* series. The first shift from metric to the ordinal may be explained by the user's intention to reduce some noise, for example, but then it seems illogical to turn back to a metric space in which all arithmetic operations are meaningful. In other words, we wish to reach again an illusory increase of precision at the final stage.

It is seen immediately that these four methodological sequences have an important property in common: ordinal and metric properties are confounded along the analytical path. Switching from one mode to the other certainly has some risk. Stepping down from *M* to *O* is possible, but some information is always lost. A 'jump' from *O* up to *M* appears more controversial, because it implies increase of information which may not be mathematically correct. Ideally, if we are satisfied with a simplified sampling strategy that yields ordered data (such as *AD* scores, water quality categories), then in the resemblance space we should also be satisfied with the ordinal treatment, using an appropriate order-based statistic. Subsequent multivariate analysis should also be ordinal in nature: only the order of dissimilarities should be considered by the algorithm. That is, the sequence *O-O-O* is the admissible combination of choices for ordinal data. The manner in which one can achieve this by data collection, by the calculation of resemblance and by scaling and clustering is overviewed next in the discussion of the three main steps of multivariate exploratory analysis.

Step 1. Sampling: types of ordinal data

The ecologist makes up his/her mind at the outset about the data type to be recorded in field work. This is a most important decision because, as we have seen above, all subsequent steps of the analysis will depend on this initial stage. Of the four possible scale types (nominal, ordinal, interval and ratio, Anderberg 1973), the ordinal is the most problematic to deal with computationally, because differences, products and ratios of possible values are not interpretable. The choice of the multivariate method to be used is further complicated by the existence of subtypes within the class of ordinal variables (Dale 1989; Podani 1999).

Fully ordered versus partially ordered data

When a variable has as many states as the number of objects (for example, $\mathbf{x}_1 = \{2, 4, 7, 8, 11, 9\}$), we are concerned with a *fully* ordered variable. This allows unambiguous ordering of all objects. Complete ordering is rarely possible, however, because in most cases the number of realized data values is considerably fewer than the number of objects and the same value may occur more than once. We are thus concerned with *partially* ordered variables (Critchlow 1985; Dale 1989), for example, $\mathbf{x}_2 = \{1, 3, 4, 5, 4, 5\}$, and the rows of Table 1a. This means that in the sequence of values ties are

Table 1. Artificial data matrices illustrating various types of ordinal data. **a.** Partially ordered, non-commensurable variables (indicator values), **b.** Fully ranked variables; **c.** Partially ordered, commensurable variables (*AD* values). Variables are rows, objects are columns.

a	Plant species
	1 2 3 2 1 2 1 2 3 2
	4 2 3 2 3 2 3 2 3 2
Environmental	1 5 3 2 3 2 1 2 1 2
variables	2 1 3 2 2 5 3 2 4 2
	1 3 3 4 3 2 3 2 3 2
	1 2 2 2 3 2 0 2 3 2
	1 2 3 1 3 2 0 2 5 1
	1 3 1 2 2 2 0 1 3 2
b	Animal species
	1 1 2 4 8 6 7
	2 2 1 3 7 7 6
	3 3 3 2 6 8 8
Food	4 4 5 1 4 4 5
types	5 5 4 5 5 5 4
	6 6 7 6 2 2 3
	7 8 8 8 3 1 2
	8 7 6 7 1 3 1
c	Sites (plots, quadrats)
	0 + 0 4 0 1 1 0 + +
	1 0 0 4 0 1 1 0 + 0
Plant	0 1 0 3 3 1 2 0 + +
species	1 + 0 0 0 0 1 0 0 +
	0 + 0 0 0 1 2 4 0 +
	2 0 1 0 1 0 3 4 0 +
	0 1 0 2 0 1 4 0 + +
	1 + 0 0 0 1 1 0 2 1

unavoidable: objects having the same score will have the same position in the ordering. Many ecological variables are of this type, for example, those summarizing the humidity, light and acidity requirements of plants ('indicator values', Mueller-Dombois & Ellenberg 1974).

Ranks

Direct observations (e.g. food preference of target animals, phenological series or the sequence of individuals arriving at a light trap) or data conversions may lead to ranked data, a special subset within the ordinal category. A data vector is fully ranked if its entries correspond to consecutive integers from 1 to n (columns in Table 1b). For the example above, ranking implies the transformation of $\mathbf{x}_1$ into $\mathbf{x}_1' = \{1, 2, 3, 4, 6, 5\}$. Ties in the original sequence $\mathbf{x}_2$ lead to partially ranked data vectors, such as $\mathbf{x}_2' = \{1, 2, 3.5, 5.5, 3.5, 5.5\}$. That is, tied positions in the order are taken by the mean values of the corresponding ranks.

Commensurable versus non-commensurable variables

This distinction reflects whether a variable can be used to order the objects and at the same time a given object can also be used to rank all the variables (Podani 1999). This duality is true for *commensurable* variables (*sensu* Orlóci 1978). Fully ordered data satisfy these requirements: each row or a column of the data matrix represents an interpretable sequence – even though full ordering in one direction does not exclude the possibility of ties in the other direction. Certain partially ordered data types are also commensurable, for example the *AD* scale and many of its derivatives commonly used in vegetation ecology (Table 1c). In community data containing such *AD* values, the objects (sampling units, quadrats, relevés) can be ordered meaningfully for each species, starting from those with the lowest *AD* value up to those having the highest score. Also, the species can be ordered according to their importance in a particular sampling unit. Other types of partially ordered variables, such the indicator variables mentioned above, are *incommensurable*: ranking of objects for a given variable is meaningful, but no ordering of indicator variables for a sampling unit can be done: the comparison of an indicator value of 4 for humidity with another value of 3 for acidity would be invalid (Table 1a).

Step 2: Ordinal measures of resemblance

A fundamental requirement if resemblance measures are to be used with ordinal data is order invariance. This means that the coefficient cannot change as long as the ordering of data values remains the same. Measures for expressing pairwise dissimilarities between objects based on ordered data are widely available, yet they are less commonly used in numerical ecological studies than regular coefficients suitable to interval or ratio-scale variables. The choice among them is governed primarily by the subtypes of the ordinal variables present in the data. Rank statistics (such as Spearman's ρ or Kendall's τ, see Legendre & Legendre 1998; Podani 2000a) could be mentioned first. However, these are appropriate only for comparing fully ordered variables with equal intervals between the neighbouring values on the ordinal scale or, more commonly, in cases where the original data scores are replaced by their ranks. This is an important condition because differences between ranks are interpreted (and squared by Spearman's ρ). These coefficients are more (ρ) or less (τ) sensitive to ties, and require correction terms if ties are present. There is a third coefficient, Goodman-Kruskal's (1954) γ which simply disregards the ties. For objects j and k, this measure is the number of variable pairs similarly ordered for j and k divided by the number of variable pairs that are ordered at all (App. 1 provides an example of calculation). Since the ties are excluded from the comparison, the similarity values can be based on a very different number of variable pairs. This imbalance can be removed through the hybrid coefficient of discordance suggested by Podani (1997) which is primarily an ordinal measure, but it does consider presence/absence relationships for species pairs that are not ordered unambiguously for the object pair being compared (App. 1). This coefficient becomes identical to Kendall's τ and Goodman-Kruskal's γ for fully ordered variables (Podani 1997). For more information on the mathematical properties of measures of rank correlation, including those not mentioned here, the reader is referred to Siegel & Castellan (1988). For partially ordered data, Dale (1989) recommends the use of other coefficients, such as the Levenshtein measure which finds the minimum number of moves necessary to transform one series into the other. This problem can be solved by combinatorial optimization, not applied routinely in numerical ecology.

Mixed data containing ordinal variables

In ecological data matrices, nominal, ordinal and 'quantitative' variables often appear simultaneously. In these cases, Gower's general formula, extended to accept ordinal data (Podani 1999), may offer a solution for expressing similarity between objects. For any two objects j and k, the contribution of a given ordinal variable to the similarity value depends on the number of objects between j and k in the rank order. In a sense, this is a

nearest neighbour measure of similarity in partial rank orders; it may violate the triangle inequality principle and is therefore non-metric – which is not a problem in ordinal data analysis. The formula may also be extended using the logic of rank correlations: the absolute difference between the ranks is considered in the comparison, a quantity understood as the number of elementary changes needed to move an element into a position taken by the other element in the rank order. Of course, Gower's coefficient may also be applied to a data matrix comprising ordinal variables only. In that case, the coefficient differs from all those discussed in the previous paragraph, because it will correspond to the average of number of steps required to move one object to the position of the other in the rank order for one variable.

Step 3: Ordinal exploratory analysis

Ordination

In statistical ecology, the most widely accepted and routinely used ordination technique which relies upon ordinal information is undoubtedly non-metric multidimensional scaling. Its relative merits and potential disadvantages have been discussed by many authors, sometimes with contrasting conclusions (Kenkel & Orlóci 1986; Gauch et al. 1981; Gordon 1999; Digby & Kempton 1987; Clarke 1993; Legendre & Legendre 1998; Podani 2000a). Nevertheless, most authors agree that NMDS and its variants (e.g. local NMDS, Sibson 1972; Prentice 1977) represent a good alternative to the metric procedures such as principal components analysis and correspondence analysis. Release of the strict metric criteria allows, for example, that the entire ordination may be restricted to two dimensions.

In NMDS, the dissimilarity matrix of objects is not involved in the calculations directly. The dissimilarities are used to constrain a set of ordination distances to fit their rank order as closely as possible. Consequently, any change in the dissimilarities which does not influence the ordering relationships will have no impact on the final ordination; that is, the method is order-preserving. Differences among results may arise only from the iterative nature of the scaling algorithm. However, the result is certainly a Euclidean representation of points. Therefore, Gordon (1999) suggests that the name *ordinal scaling* is perhaps more appropriate than non-metric scaling.

Clustering

Given the success and popularity of NMDS (or ordinal scaling), the natural question arises: is there a possibility to develop analogous methods in the other large family of exploratory data analysis techniques, namely cluster analysis? This is answered by examining the problem in detail and by exploring the literature for existing proposals and approaches. Although mathematical details are kept to the minimum, it is hoped that the following discussion illuminates the complexity of this matter satisfactorily. Readers interested only in the newly proposed classification models may skip this part, and turn immediately to the description of the clustering criterion and algorithm at the end of this subsection.

Order invariance is satisfied for those ordinal clustering strategies for which the classification topology changes only if the rank order of input dissimilarities is modified, while any other change should not influence the results. On the analogy of ordinal scaling, the term 'ordinal clustering' (OC) will be coined with these methods. This emphasizes contrast with metric clustering, which includes widely used hierarchical and non-hierarchical procedures of numerical classification, such as group average sorting (UPGMA), incremental sum of squares agglomeration (Ward method), median and the centroid strategy (plus many others that require direct access to data during the analysis, such as *k*-means clustering, fuzzy *c*-means clustering; see major reviews of numerical classification, e.g. Anderberg 1973; Everitt 1980; Gordon 1999). The fact that metric methods are not order invariant suggests that these cannot be recommended for analysing matrices derived by ordinal coefficients.

There are two widely-known hierarchical techniques, namely single link (SL, nearest-neighbour) and complete link (CL, furthest neighbour) sorting, which possess the property of order invariance (Hubert 1973; Boberg & Salakoski 1993). This is true even though both have completely meaningful metric interpretation as well: nearest neighbour distance for SL and cluster diameter for CL. SL clustering has the further advantage of being insensitive to tied minimum distances (Jardine & Sibson 1971). However, the use of nearest neighbours and cluster diameters as clustering criteria has the dramatic consequence that only a small portion of the ordered coefficients is used. The dissimilarity values that fall into the interval between fusion levels g and $g+1$ may be rearranged into any sequence within that interval without any effect upon the resulting classification. In fact, for a set S of m objects, if the rank order of the $m-1$ values that give the fusion levels remains fixed, we have complete freedom to change arbitrarily the other $(m^2-3m)/2+1$ values in the semi-matrix of dissimilarities! These values therefore play only a passive role in clustering. It means that order invariance must be satisfied for a small subset of dissimilarities and that a large amount of information present in the matrix remains

unexploited. I do not enter into details regarding potential advantages and disadvantages of these two strategies, because this topic has been thoroughly examined and discussed in the classification literature (cf. Jardine & Sibson 1971; Gordon 1999). Nevertheless, ecological use of SL and CL, especially in the past decades, appears relatively limited.

Several extensions of the CL and SL method have been proposed, all maintaining the order invariance property. Peay's (1975) approach, for example, leads to overlapping clusters, a solution mostly of theoretical, rather than practical significance, owing to the relative complexity of the result even for small problem sizes. Krauth (1986) did not use any dissimilarity coefficient, and defined nearest neighbours directly based on partially ordered raw data. His concept of neighbourhood, however, appears less applicable to cases when objects take values that are far from being neighbours.

In order to allow the use of different types of ecological variables, including ordinal ones, Matthews & Hearne (1991, see also Matthews et al. 1991a, b; Landis et al. 1997) suggested a non-metric classification method that is also free from the use of any dissimilarity coefficient. In it, the clusters are directly identified from the examination of each variable separately after the data values are discretized into a few categories. Cluster validity is measured by the proportional reduction in error, as expressed by Goodman & Kruskal's (1954) λ coefficient. Clusters are examined from random starts and iterative relocations are used to produce heuristic approximations to the optima. The partition supported by the majority of variables is accepted as the best solution, thus reflecting some optimal 'consensus' among the variables. The method is truly non-metric, rather than just ordinal, because the λ-coefficient is unaffected by permutations of categories. That is, sequential information in ordinal or interval characters is not utilized during the optimization search; the method reduces ordinal information to nominal, and the loss of information remains uncontrolled in this analysis.

Further possibilities of partitioning arise from the use of linear programming based on the optimization of an objective function. Marcotorchino & Michaud (1979) suggested using the sum of rank differences counted for each variable as partitioning criterion. Owsinski & Zadrozny (1986) proposed a more complicated formulation. It involves calculating an objective function which also considers the rank order of objects separately for each variable. Then, for each pair *jk* of objects the numbers of object pairs that fall between and outside *j* and *k* in the order are counted, and the differences summed to provide a statistic for that object pair. This statistic, computed for all pairs of objects, is subjected to linear programming with various parameters to provide an array of different solutions. None of the two approaches produces a nested hierarchy of objects if the optimization is performed for increasing numbers of clusters.

Divisive methods described by Hubert (1973) consider dissimilarities of objects to members of a particular pair of objects in building a hierarchical classification. Three clustering criteria are suggested, all based on finding in each iteration step the most dissimilar pair of objects, *jk*, that still belong to the same cluster, say C. Objects *j* and *k* form the nodes of the resulting two subgroups, C' and C'', and their relationships to the remaining objects of C will decide on assignment into either C' or C''. The three alternatives differ from each other in this assignment procedure, two of those analogous to the SL and the CL criterion, respectively. The third one proceeds with finding the next object which falls farthest from either *j* or *k*. Hutchinson & Mungale (1997) have suggested a procedure, pairwise partitioning, based on the ordering of all similarities. A partition of objects into two groups is obtained by examining each pair, *jk*, of objects. One group will contain objects that are more similar to *j* than to *k*, whereas the objects of the other group will have the opposite relationship to this pair. This partition can be coded as a binary feature vector, and the vectors determined for all possible pairs of objects are summarized in a feature matrix, the starting point of a hierarchical classification. A comparison with SL, CL and Hubert's strategies reveals, however, that pairwise partitioning produces conflicting results even for a small number of objects. Therefore, this method is less promising in exploratory data analysis, and has no more than theoretical relevance. Hubert's divisive strategies, to my knowledge, have never been applied to ecological data either.

A potential reason for neglect is that Hubert himself (1973) showed superiority of an agglomerative procedure which optimizes an objective function, called α index, throughout the analysis. The function is a goodness-of-fit measure for partitions. The denominator is the number of object pairs already in the same cluster for which there is at least one pair of objects that are in different clusters, yet their dissimilarity is lower. This sum is divided by the possible maximum to yield a range of [0,1]. The value of this index is minimized in each algorithmic step, the resulting hierarchy is nested but the change of the function is not necessarily monotonic.

As a true alternative to NMDS, Faith (pers. comm. in Clarke 1993) raised the possibility of developing a classificatory method which minimizes the difference (stress) between the original distances and those implied by a dendrogram so as to preserve the rank order of distances as faithfully as possible. According to Shah & Farach-Colton (in press), there is no guarantee that any particular distance matrix can be converted to a tree

such that the ordering relationships among distances are completely maintained by path lengths. Furthermore, even if such a tree does exist for a given matrix, its finding poses computational difficulties that are as yet unsolved (i.e. the problem is NP-hard, Lewis & Papadimitriou 1978). The situation does not change if we are satisfied by partial orders in such a way that the ordering relationships for any three objects in the tree are consistent with their original distances. Although Shah & Farach-Colton coined the term 'total ordinal clustering' and 'triangle ordinal clustering' for the determination of these two types of trees, respectively, these become classifications only if the trees are rooted (to provide a dendrogram) or broken into isolated subtrees (to provide a partition). Ordinal tree fitting is perhaps a more appropriate name for this method. Shah & Farach-Colton emphasize the importance of finding good approximations to the optimal trees, without concrete suggestions, and no recommendations are given on the derivation of dendrograms and partitions from these trees.

This discussion shows that whereas ordinal properties have been examined for clustering and ordination, the studies are scattered among a diverse literature and are confined to the theoretical aspects. This partly explains why currently available procedures satisfying order invariance are almost completely neglected in ecology. Further reasons for the general ignorance are the complexity of the problem and its solutions, and the lack of an easy-to-use software. I describe below a conceptually simple, heuristic procedure which adopts well-known clustering algorithms and utilizes the entire set of ranked dissimilarities in every cycle of the computations. Furthermore, it is made available through a multivariate data analysis package for general use.

A new clustering criterion. As a starting point, let us consider a measure of the explanatory power of variables in a partition of m objects (Podani 1998). This is originally suggested as an *a posteriori* measure of the goodness of partition which can then be used to determine the optimum number of clusters. However, the underlying idea of using within-cluster and between-cluster ranks of distances can also be fruitful as a clustering criterion to build up both hierarchical and non-hierarchical classifications. App. 2 describes the derivation of this measure in detail. Here, I present only the formula and a brief explanation necessary for understanding the definition of clustering algorithms.

Both algorithms of ordinal clustering suggested below, first order the $m(m-1)/2$ distance values so that each d_{jk} is replaced by its rank, r_{jk}. Then, the same clustering criterion is considered:

$$U = (R_w - R_{\min}) / (R_{\max} - R_{\min}) \qquad (1)$$

where R_w is the sum of ranks of within-cluster dissimilarities, $R_{\min}$ is the possible minimum sum of such ranks for the given number of clusters and for the given numbers of objects in each cluster, and $R_{\max}$ is the possible maximum of such sums. The value of U has a range of 0-1, 0 indicating that all within-cluster dissimilarities are smaller than the between-cluster dissimilarities, and 1 indicating the opposite situation. App. 2 discusses a related criterion in which the expectation, rather than the maximum is used to standardize the coefficient. Note that a similar approach has been taken by Clarke (1993) in developing an *a posteriori* statistical test of similarities between and among groups of sample sites. He used the averages of ranks in the procedure ANOSIM, most widely used in environmental impact assessment studies. There is a certain similarity to Hubert's (1973) goodness-of-fit measure as well; both his α index and U reflect a global property of the classification being constructed, rather than a local pairwise relationship of two objects.

Algorithms. The *non-hierarchical* clustering strategy, OrdClAn-N, advocated here is essentially an iterative relocation procedure, similar to k-means clustering. Since the underlying algorithm is well-known from the literature (Anderberg 1973; Gordon 1999), there is no need to give a more formal description. The classification problem is to find a partition P of a set S of m objects into κ clusters which minimizes U for a given dissimilarity matrix $\mathbf{D}$, where κ is chosen by the investigator. The analysis starts from a random or a user-specified partition. The algorithm examines $m(\kappa-1)$ relocations in each step and moves object i from group a to b if this move leads to the highest decrease in the value of U in that step. The iterations stop if no relocation of a single object would improve the clustering criterion any further. No move is taken if it would result in the decrease of κ. Being iterative, the analysis may be trapped in different sub-optimal solutions, depending on data structure and on the initial configuration. Therefore, several runs are necessary to select the best of all results. There is no guarantee, however, that the iterations will always find the optimum classification.

The *hierarchical* version, OrdClAn-H, proceeds in the same manner as all well-known agglomerative clustering methods, except that in each cycle only one fusion is allowed. The pair of objects or clusters is amalgamated into a single cluster for which U calculated for all clusters that have been created up to this point is the minimum. In the dendrogram resulting from this agglomerative clustering, the ranks of fusions (values from 1 to m–1) are used, rather than the U values themselves, because this criterion does not change monotonically. That is, the result is a *ranked dendrogram*, rather than a weighted

dendrogram (cf. Lapointe & Legendre 1991; Podani 2000b), a name which is consistent with the ordinal properties considered throughout the analysis. (Compare with NMDS, whose result is in fact a metric construct.) Although the result is an ultrametric tree like any conventional dendrogram, its topology may serve as a starting point for finding an approximation to the Shah & Farach-Colton tree, a possibility that could be explored in the future. The result of agglomerative clustering for a given number of clusters, κ, can be subjected to non-hierarchical clustering with the same value of κ to see whether improvement is possible. The method belongs to the double matrix approach described by Podani (2000a): the ranks of similarities are stored in a semi-matrix (upper half of a square matrix), and these values are used in each cycle to compute a second matrix **U** containing the clustering criteria which are used to select object or cluster pairs to fuse. Therefore, once the ranks have been determined, computing time will be proportional to the 5th power of m – i.e. the clustering algorithm has a time complexity of $O(m^5)$ – which may seem at first glance too much. On a personal computer with Pentium IV, the clustering of 80 objects in one of the examples below took less than 10 seconds, so the relatively complex algorithm poses no practical difficulties.

Both OrdClAn clustering algorithms have been implemented in the SYNTAX 2000 program package (Podani 2001) developed for exploratory data analysis in the biological sciences. The programs run on personal computers equipped with WINDOWS operation systems.

Illustrative examples

Admissible methodological schemes and the proposed clustering algorithms are illustrated by three sets of artificial data already referred to in the previous sections (Table 1). The objective is to show the combinations of analytical options best used under different circumstances. A more elaborate example relies upon actual vegetation data, and many details of this analysis are given in App. 3.

Artificial data

The data in Table 1a exemplify non-commensurable, partially ordered variables. The objects (columns) are plant species while each hypothetical variable corresponds to an environmental factor, such as humidity, temperature requirement, light, nutrient availability and similar characters of the ordinal type. Thus, the scores are indicator values reflecting species preferences or optima. Calculation of extended Gower similarity among species and subsequent application of OrdClAn-H and NMDS are straightforward (Fig. 1). However, similar analyses of variables are impossible, because the values pertaining to different variables are not comparable in any way. In fact, the present data format does not allow any analysis of relationships among the variables, which would only be possible with quantitative data.

Table 1b contains a fabricated set of data reflecting food preferences of different animal species (columns) obtained from feeding experiments. The data set is fully ordered, character states are equidistant, and therefore NMDS and OrdClAn-H using the *U* function can be

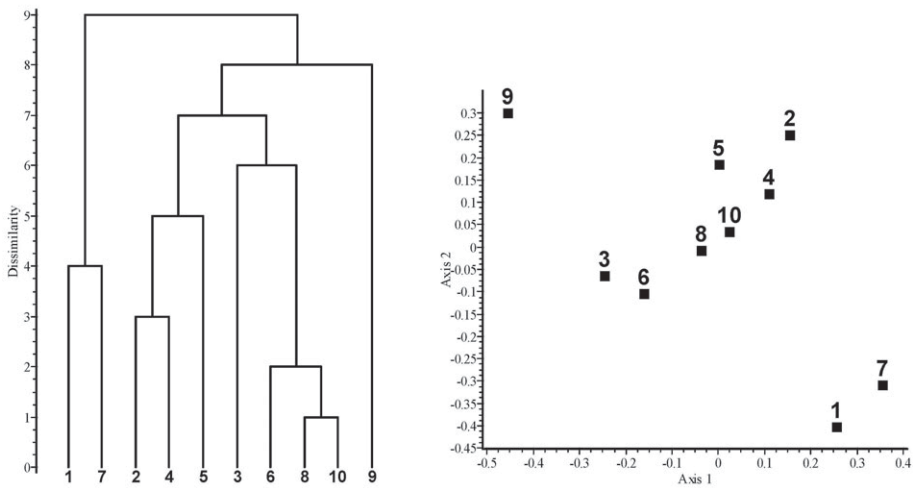

Fig. 1. Ordinal analyses for species (columns) of Table 1a. Ranked tree obtained by OrdClAn-H and two-dimensional NMDS ordination. Both analyses are based on the expanded Gower coefficient with the nearest neighbour interchange option. Note that the 'dissimilarity' shown on the axis pertaining to the dendrograms in this figure and in Figs. 2-3 is merely the rank of the fusion leading to the given cluster.

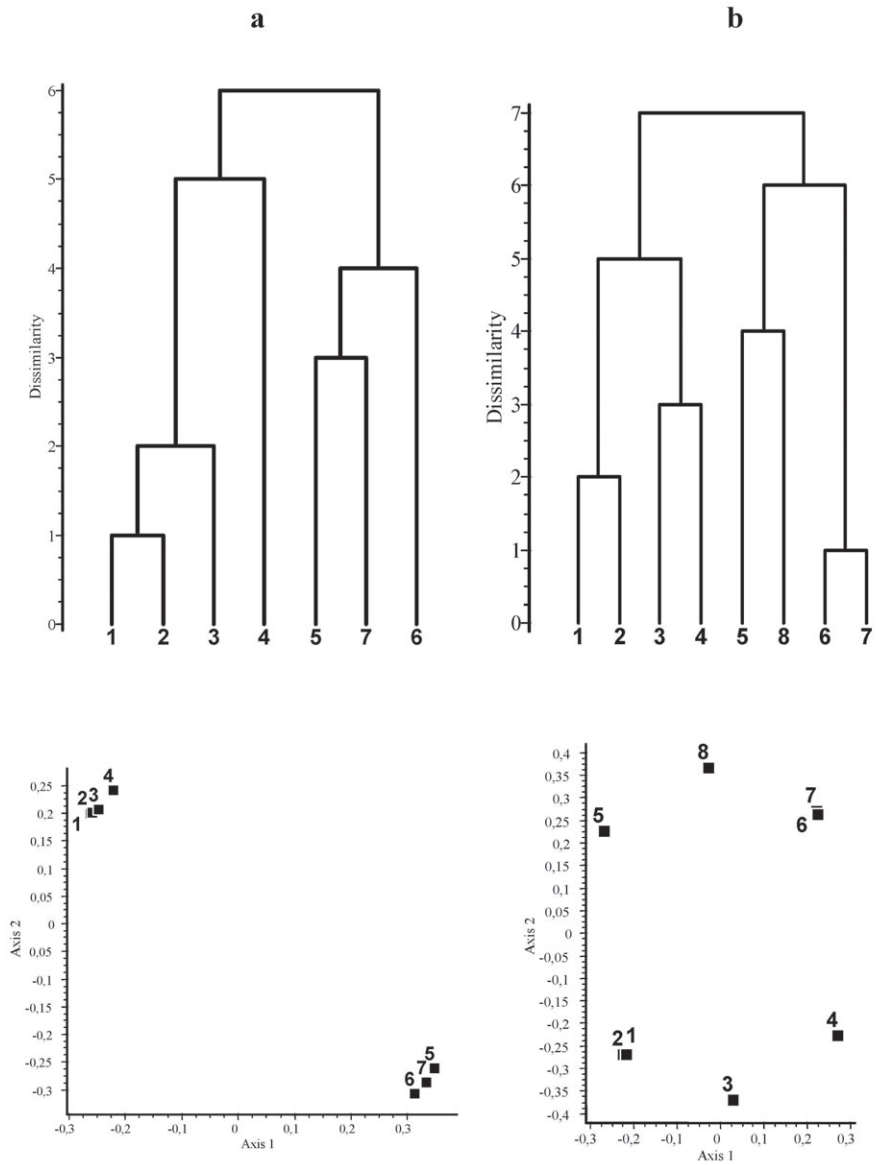

Fig. 2. Ordinal analyses of data in Table 1b. **a.** Ranked tree obtained by OrdClAn-H and a 2D NMDS ordination of objects (columns), **b.** Ranked tree and 2D NMDS ordination of variables (rows). All analyses are based on Goodman-Kruskal's γ (App. 1).

performed from a matrix of rank correlations between species. The property of commensurability holds, therefore the food types (i.e., the rows), which serve here as descriptors of species, can also be subjected to similar analyses to identify potential groupings. In the rows, however, there are many ties so that Spearman's ρ is a good choice only if interest lies in column-wise clustering. Therefore, Goodman-Kruskal's γ is applied to both the columns and the rows to facilitate comparable R and Q mode analyses of this data set (Fig. 2).

Commensurability holds for the partially ordered data of Table 1c. The entries of the data matrix are Braun-Blanquet's *AD* scores for plant species as observed in sample sites or quadrats (columns). The possible values on this scale are 0, +, 1, 2, 3, 4, and 5. The appearance of the non-number + in the data, indicating species presence with very low abundance, excludes the possibility of any analysis more sophisticated than ordinal. There is an obvious symmetry in this case, so that sample sites can be analysed by the same methods as the species. To incorporate presence/absence information for tied pairs, I selected the hybrid coefficient of discordance (App. A) to calculate the dissimilarity matrices, subjected in turn to NMDS and OrdClAn-H based on the *U* function (Fig. 3).

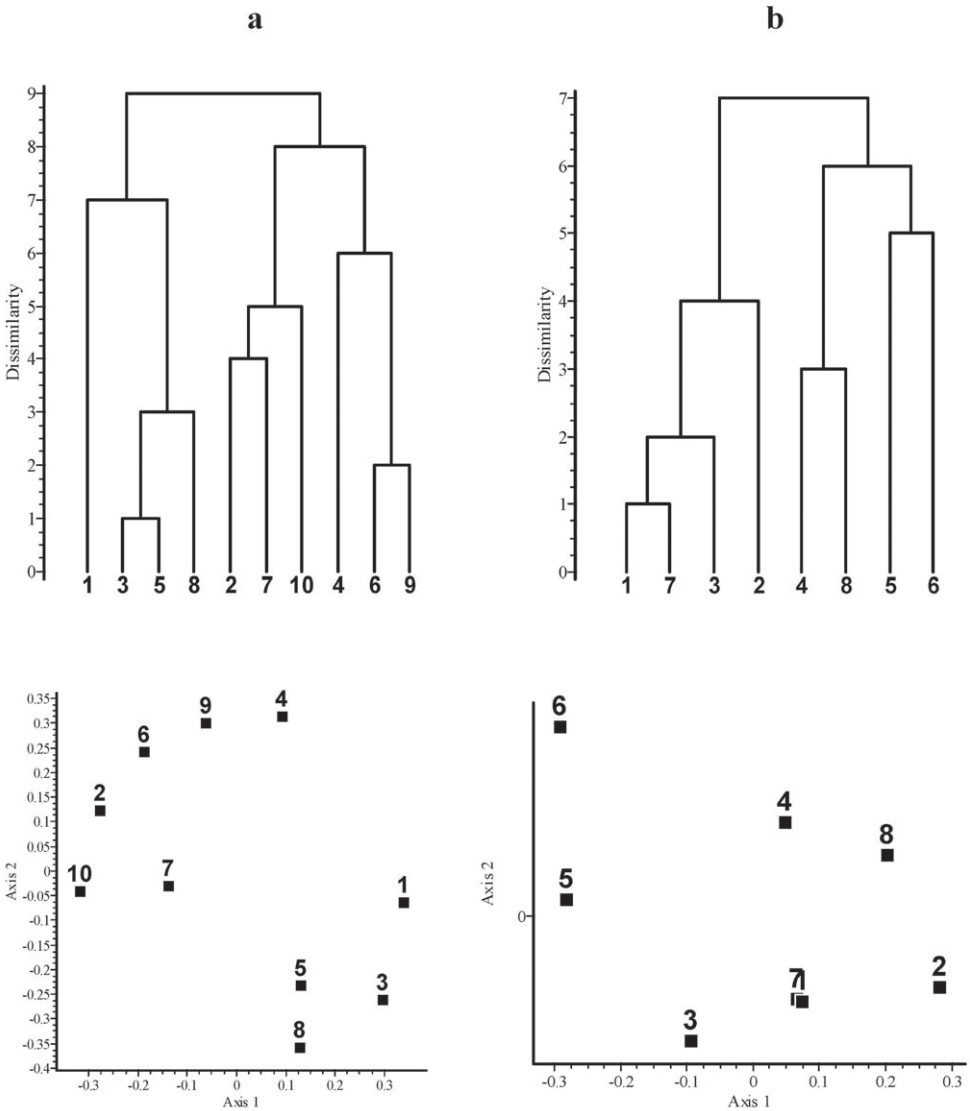

Fig. 3. Ordinal analyses of data in Table 1c. **a.** Ranked tree obtained by OrdClAn-H and 2D NMDS ordination of objects (columns); **b.** Ranked tree and 2D NMDS ordination of variables (rows). All analyses are based on the coefficient of discordance (App. 1).

Actual data

The raw data matrix comprises phytosociological *AD* scores for 123 plant species in 80 sampling units placed in a rock grassland community (Sas-hegy Nature Reserve, Budapest, Hungary). The data were recorded by the author in 1979 as percentage cover values, which were then converted 'down' to Braun-Blanquetian *AD* scores to create the sample data file. The 80 quadrats were classified by seven combinations of dissimilarity function and clustering strategy representing different methodological sequences. OrdClAn-H and CL from the matrix of the coefficients of discordance (App. A) were chosen as *O-O-O* sequences, whereas UPGMA clustering from the same matrix is an *O-O-M* series. Euclidean distances were formally computed from the ordinal data and then input to UPGMA and the Ward method (*O-M-M* sequence), as well as to OrdClAn-H and CL (*O-M-O*). The results are not reproduced here (but see the detailed comparison of two selected analyses in App. 3). Instead, the seven dendrograms were evaluated by ordination to reveal their overall similarity relationships. Each classification was described by cluster membership divergences (Podani 2000b), and then the dendrograms were compared in all possible pairs using *ED*. The 7×7 distance matrix was then subjected to principal coordinates ordination (for more details and possibilities of such meta-analysis, see Podani 2000a). The first three axes account for 27, 20 and 19% of the total variance, so that these are useful

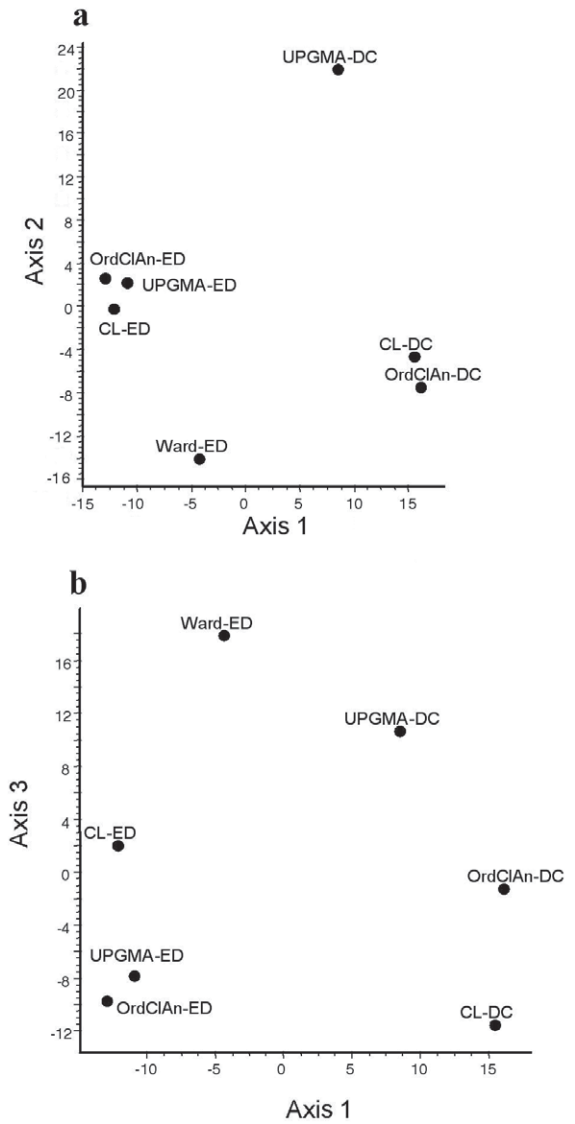

Fig. 4. Principal coordinates ordination of seven dendrograms, each representing a different combination of resemblance coefficient and clustering strategy applied to the Sashegy phytosociological data. **a**. Axis 1 vs 2; **b**. Axis 1 vs 3.

for depicting the dissimilarity structure of the dendrograms quite faithfully (Fig. 4). Axis 1 deserves particular attention, because it corresponds to the contrast between the metric and the ordinal coefficients. The relative closeness of OrdClAn-H and CL dendrograms obtained from the *DC* matrix indicates that these two ordinal procedures are similar in performance in this case, even though there are substantial differences between their algorithms. Dendrograms obtained through options that are not recommended at all are arranged relatively far from these two results.

Discussion

The significance of the subject matter of the present paper may be underlined by noting two historical facts:

• In the past 60 years, enormous amounts of ecological data have accumulated in publications and databases, a great proportion in ordinal format. Lepš & Hadincová (1992) suggested that more than 100 000 relevés had been made by that time following the principles of the Braun-Blanquet approach. This number is certainly a strong underestimate, because for The Netherlands alone there are ca. 400 000 digitized relevés available in databases (Ewald 2001, 2003), and only for the province of Mecklenburg-Vorpommern in Germany there are 50 000 stored (Berg et al. 2001). In fact, there may be millions of relevés potentially available for data analysis.

• In the past decades, multivariate analysis of ecological data have received considerable attention in hundreds of papers and dozens of textbooks. In striking contrast, very few authors emphasize the importance of mathematically correct treatment of ordinal information in exploratory data analysis.

It is therefore imperative to show the possibilities for ordinal analysis and to see under what circumstances these methods should be applied to ecological studies. Although there is a wide range of methodologies available for appropriate processing of ordinal data, these are not considered seriously in multivariate contexts in most publications. For example, Guisan & Harrell (2000) review the application of regression models to ordinal ecological data, while Agresti (1999) expands the topic to several modelling problems. This paper attempts to fill the gap on the multivariate side by reviewing available methodologies, thus illuminating the unexpectedly high conceptual diversity of this subject. As a summary, arguments in favour of using ordinal methods are grouped below.

Consistency

The first issue is *consistency* of the analysis. It means that once the ecologist decides to rely upon ordinal information, then changes the values leaving this order intact, there should be no influence on the steps of the study which follow. I referred to this as *order invariance* for a study that implements a given combination of options for resemblance and scaling or clustering. In an absolutely optimal situation, the result is also of the ordinal type, such as a ranked dendrogram, but this is not so with a partition (which is at the nominal level) and an NMDS ordination (which is in fact a metric construct).

However, a data matrix whose columns and rows are reordered to follow their separate one dimensional NMDS ordinations or separate OC classifications (App. 3) is an ordinal result.

Order invariance is easily violated, unfortunately. Euclidean distances or other metric coefficients computed formally from ordinal data can change considerably if the data are modified without influencing the ordering of values. As a consequence, none of the subsequent analyses based on these metrics will be order invariant. UPGMA or a principal coordinates analysis from a matrix of ordinal dissimilarities will also change if the coefficients are changed, even though their rank order remains the same. It is possible to get very different metric results from two dissimilarity matrices in which the rank order of entries is identical. This is not surprising, of course, if the dissimilarity was obtained from metric data, but it is a problem if ordinal data were used in deriving the dissimilarities. If we deal uniformly with the data throughout the full study, then order invariance is guaranteed.

Precision

Overall statistical *precision* in the analysis is a closely related matter. Ordinal data are often collected in ecology for simple logistic reasons: to reduce costs and effort in field work. That is, data quality is deliberately low. Of course, there are other situations in which ordinal values are the only carriers of information, and there is no possibility for conversion to quantitative data. In any case, the whole analysis cannot be any better than the start (Gill & Tipper 1978), so that the use of metric methods from an ordinal starting point can only introduce a self-deceptive, illusory precision into the analysis. The explanation of this statement is straightforward: division and multiplication are common operations during computations in metric procedures, and these are incorrect manipulations on orderings, while subtraction is meaningful only for ranks.

Nevertheless, attempts to embed ordinal information into metric space may often prove successful. In other words, if the eigenvalues of a matrix of ordinal dissimilarities are all non-negative, then the inter-point relationships can be faithfully represented by Euclidean distances. I demonstrated this for mixed taxonomic data (Podani 1999), using principal coordinates analysis. This possibility, however, is not a justification for using or preferring metric methods for ordination and classification from ordinal measures. *Rather, it is a reflection of our inability to convert ordinal information into ordinal measures in an ordinal way!* Even though we try to minimize arithmetic operations on ordinal scores, we still count numbers of certain changes in the sequences or calculate differences between ranks. It is therefore understandable that the information flow from data to resemblance is associated with a virtual increase of precision and with the emergence of metric properties. This point becomes clearer if we consider that ordered dissimilarities may directly come from observations without recording raw data. Psychological experiments often provide subjective judgments like 'objects 1 and 2 are more dissimilar than objects 3 and 4, but less dissimilar than object pairs 2 and 5' etc. This is purely ordinal information, but no one should be surprised to see that if judgments are converted to ranks, then we create a dissimilarity matrix possibly with Euclidean properties. However, if the same statements are expressed by arbitrary ordinal numbers keeping only the monotonicity of pairwise dissimilarities, there is little chance for the matrix to satisfy the metric conditions.

Tolerance

Several dissimilarity coefficients, irrespective of the nature of raw data, are known to violate the metric axioms (e.g. the Sørensen index for presence/absence scores, cf. Podani 2000a). Similar problems result from the imbalance of dissimilarities caused by missing values, affecting, e.g. values of the Gower index (cf. Legendre & Legendre 1998). The dissimilarity matrix can have negative eigenvalues in this case, which may be apparent for the investigator only if the complete output list of a principal coordinates ordination is scrutinized with sufficient attention. A consequence is that perfect Euclidean representation of the dissimilarity structure is not possible. Legendre & Legendre (1998) provide several solutions for eliminating the negative eigenvalues from the input matrix, whose effect is negligible anyway if their magnitude is small compared to large positive eigenvalues. Nevertheless, it does not matter if violation of 'Euclideanarity' is strong or weak; it is illogical to base the analysis on the absolute values, differences or averages of such dissimilarities. Ordinal exploratory analysis may provide alternative tools in such cases.

Ordinal analysis can tolerate even more than the violation of metric axioms. Certain coefficients are not suitable directly for arithmetic operations (e.g. averaging, division) implied by the analytical procedure to be used. Typical cases of this incompatibility arise when linear correlations or measures related to angles between vectors (angular separation, geodesic metric, chord distance) are subjected to certain types of clustering (Ward's method, for example). Needless to say, ordinal agglomerative clustering has no restrictions in this regard.

Why worry?

A potential counter-argument against the recommendations outlined above may be raised from experience accumulated in earlier studies. Upon examining the literature, we find many papers reporting on 'successful' applications of metric procedures to ordinal data. Often, without being aware of the incompatibility problem, the authors find the results highly interpretable, sometimes superior to simultaneous analyses of the same objects described in terms of other types of data. The authors may even realize that ordinal values cannot be used in calculating conventional diversity measures or linear correlations, but this does not prevent their subjecting the data to correspondence analysis, perhaps giving 'good' results. Ordinal values, of course, 'do not cry' when treated incorrectly during the highly automated, black-box type processing of data by 'user friendly' computer software. Robustness in the data may only explain interpretability of results obtained by an illogical combination of methods. As the phytosociological example (App. 3) demonstrated, obvious contrasts (closed *vs* open grasslands, rare *vs* more common species) may be detected in both ways – but this cannot support the application of Euclidean methods to ordinal data. Therefore, ecologists dealing with ordinal data should be encouraged to revise the results of previous analyses to see how strong the discrepancies are that are caused by the misuse of metric procedures.

'Interpretability' of results produced by invalid or doubtful combinations of analytical options, and similarity of results obtained in correct and incorrect ways, may suggest that the whole argumentation in favour of ordinal methods has very little practical relevance. Should we worry about the incompatibility between data type and analytical method at all? May be we insist on mathematical rules that are too rigorous so that the entire discussion above is just a perfectionist commentary? We should not forget, however, that ecology is often considered as a soft science by representatives of other disciplines, and it is our task to modify this attitude often accepted faint heartedly by ecologists as well. Treating ordinal information in an ordinal way throughout the analysis may be a small step forward.

Acknowledgements. I am grateful to J. Garay and I. Miklós for discussions and comments and to the libraries of Collegium Budapest and Eötvös University for the successful search for some hard to access publications. I thank J. Rapson, J.B. Wilson and two anonymous referees for their comments made on earlier versions of the manuscript. Financial support came from the Hungarian National Scientific Grant (OTKA) no. 43732.

References

Agresti, A. 1999. Modelling ordered categorical data: recent advances and future challenges. *Statistics Medicine* 18: 2191-2207.

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

Berg, C., Dengler, J. & Abdank, A. (eds.) 2001. *Die Pflanzengesellschaften Mecklenburg-Vorpommerns und ihre Gefährdung - Tabellenband*. Weissdorn-Verlag, Jena, DE.

Boberg, J. & Salakoski, T. 1993. General formulation and evaluation of agglomerative clustering methods with metric and non-metric distances. *Pattern Recogn*. 26: 1395-1406.

Clarke, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol*. 18: 117-143.

Clarke, K.R. 1999. Non-metric multivariate analysis in community-level ecotoxicology. *Environ. Toxicol. Chem*. 18: 118-127.

Critchlow, D.E. 1985. *Metric methods for analyzing partially ranked data*. Lecture Notes in Statistics 34. Springer, Berlin, DE.

Dale, M.B. 1989. Dissimilarity for partially ranked data and its application to cover-abundance data. *Vegetatio* 82: 1-12.

Digby, P.G.N. & Kempton, R.A. 1987. *Multivariate analysis of ecological communities*. Chapman and Hall, London, UK.

Everitt, B.S. 1980. *Cluster analysis*. 2nd. ed. Heinemann, London, UK.

Ewald, J. 2001. Der Beitrag pflanzensoziologischer Datenbanken zur vegetationsökologischen Forschung. *Ber. Reinhold-Tüxen-Ges*. 13: 53-69.

Ewald J. 2003. A critique for phytosociology. *J. Veg. Sci*. 14: 291-296.

Gauch, H.G., Whittaker, R.H. & Singer, S.B. 1981. A comparative study of non-metric ordinations. *J. Ecol*. 69: 135-152.

Gill, D. & Tipper, J.C. 1978. The adequacy of non-metric data in geology: tests using a divisive omnithetic clustering technique. *J. Geol*. 86: 241-259.

Goodman, L.A. & Kruskal, W.H. 1954. Measures of association for cross classifications. *J. Am. Stat. Ass*. 49: 732-764.

Gordon, A.D. 1999. *Classification*. 2nd. ed. Chapman and Hall, London, UK.

Guisan, A. & Harrell, F.E. 2000. Ordinal response regression models in ecology. *J. Veg. Sci*. 11: 617-626.

Hubert, L. J. 1973. Monotone invariant clustering procedures. *Psychometrika* 38: 47-62.

Hutchinson, J. W. & Mungale, A. 1997. Pairwise partitioning: a non-metric algorithm for identifying feature-based similarity structures. *Psychometrika* 62: 85-117.

Jardine, N. & Sibson, R. 1971. *Mathematical taxonomy*. Wiley, London, UK.

Kenkel, N.C. & Orlóci, L. 1986. Applying metric and non-metric multidimensional scaling to ecological studies: some new results. *Ecology* 67: 919-928.

Krauth, J. 1986. Classification procedures for ordered categorical data. In: Gaul, W. & Schader, M. (eds.) *Classification as a tool of research*, pp. 249-255. Elsevier, Amsterdam, NL.

Kruskal, J.B. 1964. Non-metric multidimensional scaling: a numerical method. *Psychometrika* 29: 115-129.

Landis, W.G., Matthews, R.A. & Matthews, G.B. 1997. Design and analysis of multispecies toxicity tests for pesticide registration. *Ecol. Appl.* 7: 1111-1116.

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

Legendre, P. & Legendre, L. 1998. *Numerical ecology*. 2nd. ed. Elsevier, Amsterdam, NL.

Lepš, J. & Hadincová, V. 1992. How reliable are our vegetation analyses? *J. Veg. Sci.* 3: 119-124.

Lewis, H.R. & Papadimitriou, C.H. 1978. The efficiency of algorithms. *Sci. Am.* 238: 96-109.

Marcotorchino, F. & Michaud, P. 1979. *Optimisation en analyse ordinale des données*. Masson, Paris, FR.

Matthews, G. & Hearne, J. 1991. *Clustering without a metric*. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, pp. 175-184.

Matthews, G.B., Matthews, R.A. & Hachmöller, B. 1991a. Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. *Can. J. Fish. Aquat. Sci.* 48: 2184-2190.

Matthews, R.A., Matthews, G.B. & Ehinger, W.J. 1991b. Classification and ordination of limnological data: a comparison of analytical tools. *Ecol. Model.* 53: 167-187.

Mueller-Dombois, D. & Ellenberg, H. 1974. *Aims and methods of vegetation ecology*. Wiley, New York, NY, US.

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

Owsinski, J.W. & Zadrozny, S. 1986. Clustering for ordinal data: a linear programming formulation. *Control Cybern.* 15: 183-193.

Peay, E.R. 1975. Non-metric grouping: clusters and cliques. *Psychometrika* 40: 297-313.

Podani, J. 1997. A measure of discordance for partially ranked data when presence/absence is also meaningful. *Coenoses* 12: 127-130.

Podani, J. 1998. Explanatory variables in classifications and the detection of the optimum number of clusters. In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H. & Baba, Y. (eds.) *Data science, classification, and related methods,* pp. 125-132. Springer, Tokyo, JP.

Podani J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon* 48: 331-340.

Podani, J. 2000a. *Introduction to the exploration of multivariate biological data*. Backhuys, Leiden, NL.

Podani, J. 2000b. Simulation of random dendrograms and comparison tests: some comments. *J. Classif.* 17: 123-142.

Podani J. 2001. SYN-TAX 2000. *Computer programs for data analysis in ecology and systematics. User's manual*. Scientia, Budapest, HU.

Prentice, I.C. 1977. Non-metric ordination methods in ecology. *J. Ecol.* 65: 85-94.

Shah, R. & Farach-Colton, M. In press. On the complexity of ordinal clustering. *J. Classif.*

Sibson, R. 1972. Order invariant methods for data analysis. *J. R. Stat. Soc. B* 34: 311-349.

Siegel, S. & Castellan, N.J. 1988. *Nonparametric statistics for the behavioral sciences,* McGraw-Hill, New York, NY, US.

Received 15 September 2004;
Accepted 3 September 2005.
Co-ordinating Editor: J.B. Wilson.

For Apps. 1-3, see JVS/AVS Electronic Archives: www.opuluspress.se