COMMUNITY ECOLOGY 2(1):81-92, 2001
1585-8553 © Akadémiai Kiadó, Budapest

Statistical evaluation of cross-classifications derived from rearranged community data matrices

S. Camiz[1] and J.-J. Denimal[2]

[1]*Dipartimento di Matematica "Guido Castelnuovo", Università di Roma "La Sapienza", Piazzale Aldo Moro,2, I-00185 Roma, Italy. E-mail: sergio.camiz@uniroma1.it*
[2] *Laboratoire de Probabilité et Statistique, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq, France. E-mail: jean-jacques.denimal@ univ-lille1.fr.*

Keywords: Cluster analysis, Contingency tables, Correspondence analysis, Cross-classification, Exploratory data analysis, Hierarchical clustering, Meadow vegetation.

Abstract: In order to enhance interpretation of two-way contingency tables (cross-classifications) derived from two hierarchical classifications, new indices are suggested to evaluate the relative contribution of nodes in either hierarchy to the nodes or to a partition of groups derived from the other hierarchy. Using these tools, cut-levels in both hierarchies can be found to define optimal partitions, and groups from both partitions can be associated in order to identify their mutual relationships. The method is illustrated with an actual example from vegetation ecology.

Abbreviations: AOC - Analysis of Concentration, CA - Correspondence Analysis, HC - Hierarchical clustering.

Introduction

An important area in exploratory data analysis (Benzécri et al. 1973-82, Orlóci 1978, Lebart et al. 1984, Lebart et al. 1995) is the rearrangement of the rows and columns of data matrices, in order to depict their underlying structures (Bertin 1977). An advantage of this approach is that the investigator directly and conveniently interprets the results, even if he is not aware of the technical details of how rearrangements were made. It is not surprising therefore that ecologists often call for these procedures. The rearrangement may involve reordering of columns and rows, the fundamental approach in *seriation* (Kendall 1971) and, in general, in ordination. For the purpose, *Correspondence Analysis* (CA, Benzécri et al. 1973-82, Hill 1973, 1974) has been used most extensively, although other scaling procedures capable of rearranging both the objects and the variables may apply. The other major approach intends to maximise separation among blocks of data values in the matrix, through simultaneous classification of the rows and the columns, and is often referred to as *block clustering*. There is a sharp division within this group of procedures: the first group optimises blocks directly (see e.g., Hartigan 1975, Podani 2000), whereas the second group of methods involves clustering the variables and the objects separately. In this second group, further distinction should be made between *partitioning* methods (such as the *k*-means procedure, MacQueen 1967, see also Diday 1971) and *hierarchical clustering* (HC, see e.g., Anderberg 1973). Partitioning strives for optimality more directly than hierarchical clustering, with the drawback that assumptions about the number of groups are to be made *a priori* (Orlóci 1967). On the other hand, HC methods generate *dendrograms* which provide more details on the association structure among objects than do non-hierarchical classifications. *Cutting* the hierarchy at any particular hierarchical level gives a partition, so that a hierarchy can be conceived as a series of partitions. Joint use of partitioning and HC has been considered by Lebart et al. (1995), who proposed to derive hierarchies from partitions or to optimise by rearrangement a partition obtained through HC. André (1988) criticises the classical *dichotomic* use of HC, suggesting *polythetic* hierarchies as more adequate representations of community data structures. In general, no *a priori* assumptions on the number of groups are available, therefore a HC is suggested to get a first impression on group structure. Regardless whether the classification of rows or the columns is completed via hierarchical or non-hierarchical methods, the data matrix is subsequently rearranged according to the new groupings and then the resulting blocks examined to mutually interpret row-wise and column-wise classifications.

Such rearrangements are termed *cross-classifications* in the literature and can be condensed into *contingency table* format. In such a contingency table, there are as many rows as the number of row groups, whereas the number of columns equals the number of column groups in the rearranged matrix. Each cell of this table is the sum of the original values in the corresponding block of the rearranged data matrix. For presence-absence data, for example, the cell frequency is the number of occurrences in the given block. An obvious requirement for an optimal matrix rearrangement that some blocks should contain many values, whilst other blocks should be as *empty* as possible, all depending on the actual problem, of course. A crucial question is then to measure *how sharp these blocks* are. A table may be called *well-structured* if the blocks are much (*significantly*) sharper than blocks formed from the data by chance. Based on this concept, we introduce new tests for the evaluation of group structure in hierarchy-based cross-classification tables. Given a hierarchical classification of the rows (r-hierarchy) and another of the columns (c-hierarchy), the idea is to investigate the interaction between the nodes of these hierarchies. Each node corresponds to the fusion of a pair of groups, so that given a node of the r-hierarchy and another from the c-hierarchy, a corresponding 2×2 contingency table may be constructed. This may be used to detect significant associations among the groups considered. In this way, the mutual influence of each node of the r-hierarchy on the nodes and groups of the c-hierarchy may be revealed, and vice versa. The generation of statistical distributions of the derived association and interaction indices provides a test for significance. In this paper, first the theoretical background of the problem is discussed and then the new method, incorporating indices based on particular χ^2 components, is introduced. The procedure is illustrated using Ellenberg's grassland data (Müller-Dombois and Ellenberg 1974, see also Gauch and Whittaker 1981), already used by Camiz (1994) for demonstrating the data rearrangement procedure itself.

Background theory

Analysis of concentration

Camiz (1988, 1991, 1993) reviewed methods currently available for vegetation scientists to detect inherent structure in data tables and proposed (Camiz 1994) a semiautomatic procedure available through the Mulva-4 package (Wildi and Orlóci 1990). This method relies upon *analysis of concentration* (AOC) suggested by Feoli and Orlóci (1979) to evaluate sharpness of blocks in rearranged data matrices. Although differently formulated by the authors, AOC is in fact a correspondence analysis of a contingency table in the sense of Benzécri (1973-82) and Hill (1973, 1974) with the only difference that each cell, for presence/absence data, is normalised to the number of its entries. AOC yields a joint ordination of row and column groups, and the resulting χ^2 value and the *mean square contingency coefficient* quantify the strength of correspondence between the two classifications. In addition, the square roots of eigenvalues (the *canonical correlations)* can be used to express the overall relation between the row and column groups of the table. The canonical correlations are informative on the number of background gradients responsible for the hidden data structure. When there is a non-random block structure, then the underlying gradients influence both the rows and the columns, so that their meaning should be comparable for both the columns and the rows. This may be true only for eigenvectors with high associated eigenvalues, i.e., highest canonical correlations. Thus, AOC can play a central role in decisions regarding the number of CA axes to display and the cut-levels of dendrograms to get optimum classifications.

The ability of AOC in partitioning the χ^2 of the contingency table is helpful for identifying the influence of environmental factors through *lattices* (Orlóci and Kenkel 1987), i.e., the contingency tables built by partial reconstruction of the original table, considering the χ^2 explained only by individual AOC axes. This may also be obtained through *Principal Components Analysis on Instrumental Variables* (Rao 1964, see also Non-symmetrical Correspondence Analysis, Lauro and D'Ambra 1984). Nevertheless, some limits have to be taken into account when its application is contemplated:

i) AOC has a meaning only when the table contains presence/absence data, though Podani and Feoli (1991) suggest procedures for other types of data;

ii) no distribution-based significance test of the indices is possible, apart from the χ^2 -test, so that their evaluation remains within the domain of rules of thumb or experience;

iii) the choice of the number of *meaningful* eigenvalues remains subjective. In fact, attempts to use eigenvalue distributions in Principal Components Analysis and CA (Lebart et al. 1977) or the theoretical Wishart's distribution (Hirotsu 1983, Greenacre 1988) never became practical;

iv) the overall picture on the association between groups may be visualized by CA joint plots. However, this approach is not sufficient to explain precisely all individual associations among pairs of groups from the rows and columns: in fact, the proximity between the positions on or-

dination planes of of a species and a relevé group does not always reflect true closeness or high association.

Optimal cut levels in dendrograms

Several methods are suggested in the literature for the identification of the optimal cut-levels of a hierarchy. Camiz (1994) proposes some rules of thumb, based on examining the fusion level sequence and its first and second discrete derivatives. The procedure looks for fusion levels followed by *significant* increases, since they potentially indicate that in the following step the two groups to be merged may be thought statistically different and should not be clustered together. In this paper, we shall use Mojena and Wishart's (1980) *moving average quality control rule,* based on the same principle but with a well-established statistical background. The method relies upon moving statistics: given a sequence of values, a moving average is the mean of *m* adjacent values. Let us consider the sequence of fusion levels corresponding to the ascending sequence of nodes in a hierarchy. Having *m* predetermined by the investigator, for every node the quantity

$$\mu_j + l_j + s_j + k_j d_j$$

is computed. In this, μ_j is the moving average of the *m* fusion levels to ν_j; l_j is the correction for trend lag at node *j*, so that $\mu_j + l_j$ is the expected value of ν_j; s_j is the moving least squares slope at node *j*, so that $\mu_j + l_j + s_j$ is the expected value of ν_{j+1}; d_j is the moving unbiased estimate of the population standard deviation at node *j*; and k_j is the standard variate $k_j=(\nu_{j+1} - \mu_j) / \sigma_i$, where σ_i is the standard deviation of ν_j. Beyond the threshold t_j, the fusion level increases are considered significant. Thus, all nodes *j* such that $\nu_{j+1} > t_j$ are candidates for a cut-point. They are then ordered according to their significance.

More recently, other methods have been suggested. Gordon (1998) selects the five best methods among the many reviewed by Milligan and Cooper (1985). Hardy and Deschamps (1999) compare them with their new technique which is based on the variation of the sum of Lebesgue measures of the hyper-volumes corresponding to each group. In addition to Feoli and Orlóci's (1979) AOC method, the cross-classifications have already been considered in the literature. Both Govaert (1984) and Podani and Feoli (1991) generalised *k*-means by rearranging both rows and columns to groups, thus optimising objective functions. Greenacre (1988) generates two hierarchies with Ward's (1963, see also Orlóci 1978), *minimum increase of sum of squares clustering* criterion. Then follows Hirotsu's (1983) suggestion to choose the cut-levels according to fusion levels considered significant based on the distribution of the largest eigenvalue of a Wishart matrix (Anderson 1984). Its use is very difficult and time-consuming, especially for many dimensions.

Interpretation of classifications

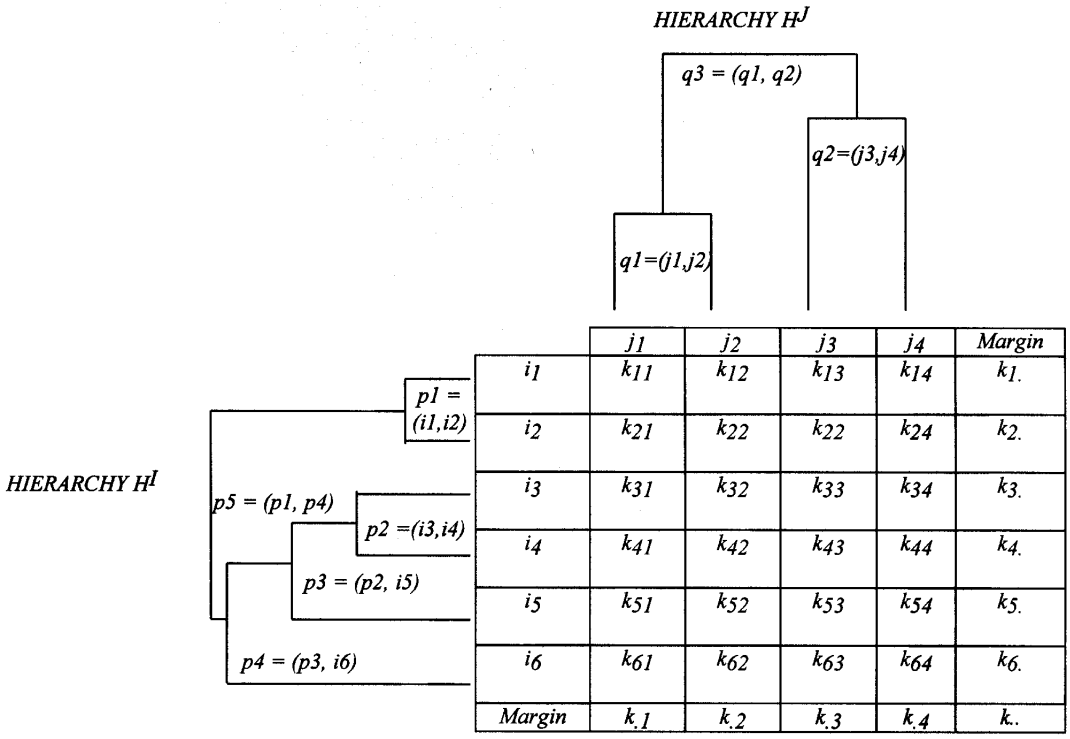
Many classification studies comprise four steps: i) hierarchical clustering of objects; ii) a partition of objects derived by cutting the dendrogram at some suitable level; iii) each group is characterised to reveal the differences among objects belonging to the different groups; and iv) the relations are summarised for a complete explanation. The latter could lead to the identification of a possible dependence structure in the data table: in case of vegetation, this may help to reveal ecological gradients.

For the interpretation of the groups of objects, their differences are depicted by considering the within-group distribution of characters. Each variable-group is explained by examining the behaviour of its members in the objects and vice versa. The vegetation tables may be seen in this context: groups of relevés may be attributed to the plant associations corresponding to the group of species present in the relevés.

For the explanation of the groups in a partition of units, Lebart et al. (1979) developed a complete set of tests to identify *typical characters* whose distributions are significantly different over groups. The tests are based on the statistical distributions of characters, in case of frequencies, the hypergeometric law is used. For each group of relevés, all species with frequency significantly higher or lower than a threshold for a given significance level (5%, 1%, or 1‰) may be listed, then sorted in order of significance. Although very accurate in the identification of the typical characters, especially in small tables, the application of these methods to very large data sets showed its limitations. Characters with very different behaviour may happen to be typical for some group, others may be typical of several groups but with a different average frequency, so that interpretation becomes difficult. In case of a cross-classification, Lebart et al.'s (1979) method is still applicable, but neither direct comparison is possible among groups of either partition, nor typical characters may be arranged according to a partition or a hierarchy. This requires further development of methods and tests, specifically tailored for the cross-classifications of contingency data tables.

The interest of clustering both rows and columns leads to *cross-classification* techniques, where some problems arise. In fact, the choice of optimal cut-levels and the interpretation of group structure should not be done separately for each hierarchy, since the search for a dependence structure between rows and columns be-

Table 1. A sample contingency table with two hierarchies.



	j1	j2	j3	j4	Margin
i1	k_{11}	k_{12}	k_{13}	k_{14}	$k_{1.}$
i2	k_{21}	k_{22}	k_{22}	k_{24}	$k_{2.}$
i3	k_{31}	k_{32}	k_{33}	k_{34}	$k_{3.}$
i4	k_{41}	k_{42}	k_{43}	k_{44}	$k_{4.}$
i5	k_{51}	k_{52}	k_{53}	k_{54}	$k_{5.}$
i6	k_{61}	k_{62}	k_{63}	k_{64}	$k_{6.}$
Margin	$k_{.1}$	$k_{.2}$	$k_{.3}$	$k_{.4}$	$k_{..}$

comes the primary goal. For this reason, an integrated approach calls for implementation.

In the following we shall deal with the last two steps of clustering procedures and we propose indices for the interpretation of cross-classifications. Such indices were first introduced by Denimal (1997) for the mutual interpretation of hierarchy nodes and partitions. Camiz and Denimal (1998ab) found the indices suitable for characterizing the reciprocal relationships between nodes and proposed a graphical representation of the association strength. Applying a grey scale to the indices themselves or to the corresponding statistics, the cells of the contingency table are shaded and the most evident associations between the corresponding row and column groups become clear. To complete this representation, the r- and c-dendrograms may be drawn beside the data table, as usual in quantitative ecology.

The new method

Notation and example

We consider a contingency table K for two sets of categories I and J (usually *species* and *relevés*). A general element of the table will be k_{ij}, $i \in I$, $j \in J$ and row, column, and grand totals will be $k_{i.} = \Sigma_{j \in J} k_{ij}$, $k_{.j} = \Sigma_{i \in I} k_{ij}$, and $k_{..} = \Sigma_{(i,j) \in I \times J} k_{ij}$, respectively. So, given two subsets $p \subseteq I$, $q \subseteq J$, we denote accordingly the partial sums $k_{pj} = \Sigma_{i \in p} k_{ij}$, $k_{iq} = \Sigma_{j \in q} k_{ij}$, $k_{p.} = \Sigma_{j \in J} k_{pj}$, $k_{.q} = \Sigma_{i \in I} k_{iq}$ and $k_{pq} = \Sigma_{(i,j) \in p \times q} k_{ij}$. Then, we suppose that two hierarchies H^I and H^J pertain to I and J, respectively, so that we can represent the data as in Table 1. In this example, the contingency table has 6 rows and 4 columns, and from the hierarchies we consider 5 and 3 nodes, respectively, denoted by $p1, ..., p5$, and $q1, ..., q3$. For node $p3 = (p2, i5)$ of H^I, we may want to look for the nodes of H^J whose association with $p3$ can be regarded as statistically significant (Denimal and Camiz 2001). Here, we considered every node as the couple of the component groups (in the example, $q3 = (q1, q2)$ or $p3 = (p2, i5)$), but it may be seen as well as a group where two subgroups merged (as in the example, $q3 = q1 \cup q2$ or $p3 = p2 \cup i5$). Two kinds of association should then be investigated. In the first case, the nodes are seen as pairs $p = (p1, p2)$ and in the second as groups $p = p1 \cup p2$ of H^I. In both cases, we look for the pairs $q = (q1, q2)$ from H^J, to explain them. Note that here and in the following we omit the reference to the notation of the elements in Table 1.

Exact conditional tests

In Cases 1-2 that follow, the absence of relations between elements of H^I and of H^J will be tested using the multiple hypergeometric law proposed for categorical data (Agresti 1990). In fact, the hypergeometric law is de-

Table 2. The contingency for Case 1 discussed in text.

	$q1$	$q2$	$\overline{q1 \cup q2}$	*Margin*
$p1$	k_{p1q1}	k_{p1q2}	$k_{p1.} - k_{p1q1} - k_{p1q2}$	$k_{p1.}$
$p2$	k_{p2q1}	k_{p2q2}	$k_{p2.} - k_{p2q1} - k_{p2q2}$	$k_{p2.}$
$\overline{p1 \cup p2}$	$k_{.q1} - k_{p1q1} - k_{p2q1}$	$k_{.q2} - k_{p1q2} - k_{p2q2}$	$k_{..} - k_{.q1} - k_{.q2} - k_{p1.} - k_{p2.} + k_{p2q1} + k_{p2q2} + k_{p1q1} + k_{p1q2}$	$k_{..} - k_{p1} - k_{p2.}$
Margin	$k_{.q1}$	$k_{.q2}$	$k_{..} - k_{.q1} - k_{.q2}$	$k_{..}$

Table 3. The contingency table for Case 2 discussed in text.

	$q1$	$q2$	$\overline{q1 \cup q2}$	*Margin*
p	k_{pq1}	k_{pq2}	$k_{p.} - k_{pq1} - k_{pq2}$	$k_{p.}$
$\overline{p}$	$k_{.q1} - k_{pq1}$	$k_{.q2} - k_{pq1}$	$k_{..} - k_{.q1} - k_{.q2} - k_{p.} + k_{pq1} + k_{pq2}$	$k_{..} - k_{p.}$
Margin	$k_{.q1}$	$k_{.q2}$	$k_{..} - k_{.q1} - k_{.q2}$	$k_{..}$

fined on the set of contingency tables having fixed dimensions and fixed margins.

Case 1. Here we consider the nodes as pairs: $p = (p1, p2)$. The 2×2 tables for groups $p1$ and $p2$ with $q1$ and $q2$ will be expanded to 3×3 tables. More precisely, writing the set complement of p in the total population, the 3×3 tables are obtained by crossing $p1$, $p2$ and $\overline{p1 \cup p2}$ with $q1$, $q2$ and $\overline{q1 \cup q2}$, as shown in Table 2. The table margins are supposed to be fixed and derive from the *a priori* known values $k_{..}$, $k_{p1.}$, $k_{p2.}$, $k_{.q1}$, and $k_{.q2}$. It is well known that the multiple hypergeometric law is defined on the set of these tables and only depends on the values k_{p1q1}, k_{p1q2}, k_{p2q1}, and k_{p2q2} (Agresti 1990).

We define the *association* of p and q, by the ratio

$$A_{pq} = k_{..} k_{pq} / (k_p k_q)$$

between observed and expected frequencies. It leads to the statistic $V = ((A_{p1q1} - A_{p2q1}) - (A_{p1q2} - A_{p2q2}))^2$ that can be used to explain the reciprocal influences among nodes $p = (p1,p2)$ and $q = (q1,q2)$. Given V_{obs}, the observed V, and a significance level α, a significant interaction between the nodes p and q will result once the probability of the event $\{V > V_{obs}\}$ is smaller than α under the multiple hypergeometric model. Note that in this case the interactions are symmetric according to both hierarchies.

Case 2. Here we consider the nodes as groups: $p = p1 \cup p2$. In this case, by merging $p1$ and $p2$, the table reduces to the size 2×3, as shown in Table 3, whose fixed margins derive from the values $k_{..}$, $k_{p.}$, $k_{.q1}$, and $k_{.q2}$. Therefore, the hypergeometric law depends on the values k_{pq1} and k_{pq2}. Here we get the statistic $U = (A_{pq1} - A_{pq2})^2$ that can be used to explain the influence of the pair $(q1,q2)$ on p. Given U_{obs}, the observed U, and α, if the probability of the event $\{U > U_{obs}\}$ is smaller than α under the multiple hypergeometric model, one may attribute a significant impact to $(q1,q2)$ in the explanation of p. A symmetric approach leads to the quest of nodes $(p1,p2)$ with significant impact in the explanation of nodes q of the other hierarchy H^J.

Test-values

In order to understand easily and quickly the differences between the considered associations, we refer every A_{pq} to its distribution under the hypergeometric model. Once the probability p to get a value lower than or equal to the observed value is found, we transform it according to the inverse cumulative normal distribution F (Morineau and Alevizos 1992). The obtained value, called *test-value*, measures the deviation from the expectation expressed in standard deviation units, much easier for the user to compare to the usual bounds, namely ± 1.96 for α

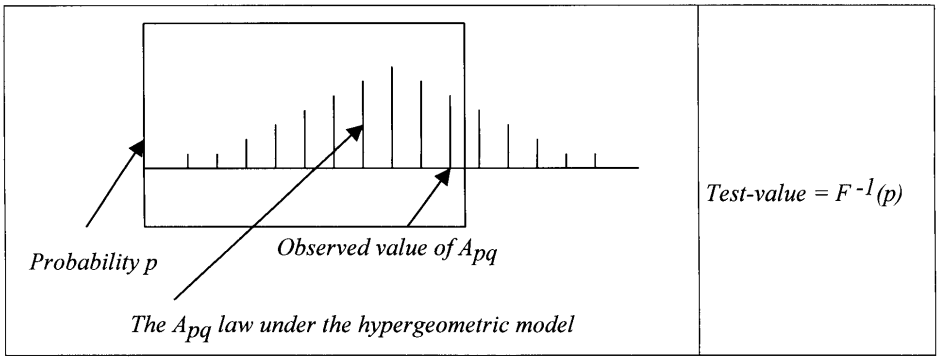


Figure 1. The test-value associated with $A_{pq,}$ the observed value.

= 5%, or ±2.57 for α=1%. In Figure 1, the probability *p* is outlined by the box enclosing all the values of A_{pq} lower than or equal to the observed one.

Geometrical interpretation

There is a straightforward geometrical interpretation of the statistics *U* and *V* used in the tests. For *V*, after setting

$$\delta_{(p1,p2),(q1,q2)} = (A_{p1q1} - A_{p2q1}) - (A_{p1q2} - A_{p2q2}) =$$

$$k_{..}\left(\frac{k_{p1q1}}{k_{p1.}k_{.q1}} - \frac{k_{p2q1}}{k_{p2,}k_{.q1}}\right) - k_{..}\left(\frac{k_{p2q1}}{k_{p1.}k_{.q2}} - \frac{k_{p2q2}}{k_{p2.}k_{.q2}}\right)$$

it becomes clear that $V = \delta^2_{(p1,p2),(q1,q2)}$. The expectation and the variance of $\delta_{(p1,p2),(q1,q2)}$ can be calculated under the multiple hypergeometric model with

$$E(\delta_{(p1,p2),(q1,q2)}) = 0$$

$$VAR(\delta_{(p1,p2),(q1,q2)}) = k_2/(k_{..}-1) \times (k_{p1.}+k_{p2.})/(k_{p1.}k_{p2.}) \times (k_{.q1}+k_{.q2})/k_{.q1}k_{.q2}\ .$$

Now, if we use Ward's (1963) agglomerative method for the construction of the hierarchies in the frame of χ^2 metrics, the aggregation indexes ν(*n*) of the nodes *n = (p1,p2)* of H^I can be decomposed, up to the constant (*k*-1), into the sum of squares of standardized variables $\delta^2_{(p1,p2),(q1,q2)} / VAR(\delta_{(p1,p2),(q1,q2)})$ (Denimal 1997), namely

$$\nu(n) = \frac{1}{(k-1)} \cdot \sum_{(q1,q2)\in H_j} \frac{\delta^2_{(p1,p2),(q1,q2)}}{VAR(\delta_{(p1,p2),(q1,q2)})}\ .$$

As a consequence of decomposition, one may use the previous tests to identify the siginificant cut-levels in both hierarchies. In fact, if at least one node of the c-hierarchy has a significant interaction with the nodes of the r-hierarchy at a given α, a cut level may be chosen so that all the nodes under that have no interaction. This rule should be used with care in the case of hierarchy built on distances computed on reduced dimensional factor spaces, since interactions are estimated within the contingency table, and significant low-level associations among items may occur due to the loss of information in the considered factors space. For this reason, in this case one may decide to use the rule only in pre-defined upper parts of the hierarchies, chosen with other criteria.

For *U* we introduce $\delta_{p,(q1,q2)} = (Apq1 - Apq1) = k_{..}[k_{pq1}/k_pk_{q1}) - k_{pq2}/(k_pk_{q2})]$ so that $U=\delta^2_{p,(q1,q2)}$ and similar comments may be made concerning its distribution. Note that in the table reporting the results of the application, $\delta^2_{p,(q1,q2)}$ will be displayed as the share of (*q*1, *q*2) to the sum of δ^2-s.

Application to Ellenberg's grassland data

As an example, we consider Ellenberg's grassland data table (Müller-Dombois and Ellenberg 1974, see also Gauch and Whittaker 1981) used also in Camiz (1994). The table represents 25 relevés of meadows from Germany, with 76 species present in more than one relevé (the singleton species were omitted). They comprise three community types: *Bromus-Arrhenatherum*, *Geum-Arrhenatherum*, and *Cirsium-Arrhenatherum*. We do not discuss here the analysis already made by Camiz (1994) for structuring the table and accept our previous choice to keep only the first two CA axes. We compare here the rule of thumb for the cut-levels, validated by the AOC, with the cut-levels obtained by the second rule of Mojena and Wishart (1980) and we discuss the information obtained by the use of the tests described here.

In Figures 2 and 3, relevés and species are represented respectively on the plane of the first two CA axes. An arch-effect is evident, due particularly to relevés 25 and 19 (corresponding to group 1 of 7) and species 76, 91, and 27 (groups 3 and 2 of 13), which only partially hides a possible second gradient.

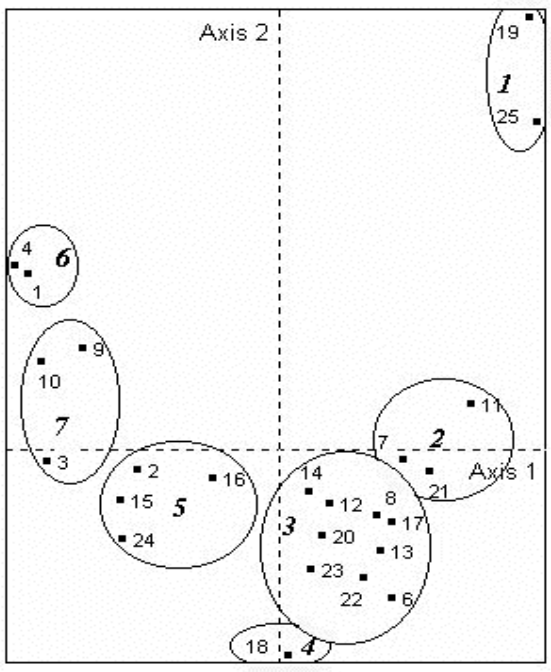


Figure 2. Representation of relevés along correspondence analysis axes 1 and 2. The labels of groups of relevés according to the chosen partition are in bold italics

Two HCs were performed through Ward's (1963) method on the Euclidean distances between objects on the plane of the first two CA axes. Camiz's (1994) rule, based on the inspection of fusion levels and derivative sequences, suggests 3, 5, and 7 groups for the relevés, and 5, 8, and 13 groups for species. The results of the AOC, in particular the inspection of canonical correlations, suggest no more than two gradients, with 0.4 as a threshold. Camiz (1994) chose the 7×13 cross-classification, in order to rearrange the vegetation table, according to the position of the groups along the first AOC axis. It must be pointed out also that all partitions with 8 groups of species had less significant results than the partition with only 5 groups. The application of Mojena and Wishart's (1980) second stopping rule suggests 3, 5, or 4 groups of relevés and 3, 2, 7, and 13 groups of species (or 3, 2, 7, 5, 8, and 13, according to different window sizes used in the computation of moving statistics). According to this rule, without access to the results of AOC, one could select first the 3×3 cross-classification followed by the 5×7 table. Considering in detail the tables constructed using interaction indices and tests results, the first table (Table 4) allows to detect significant interactions between nodes ($p1$,$p2$) and ($q1$,$q2$). Here, each row corresponds to a node of the species hierarchy with its two branches and corresponding relative weights; each column corresponds to a node of the relevé hierarchy with its branches and relative weights. In each cell, the interactions between pairs of branches of each node are reported in terms of four test-values with the p-value of the interaction test. Recall that the p–value represents the probability of the critical region in the corresponding test.

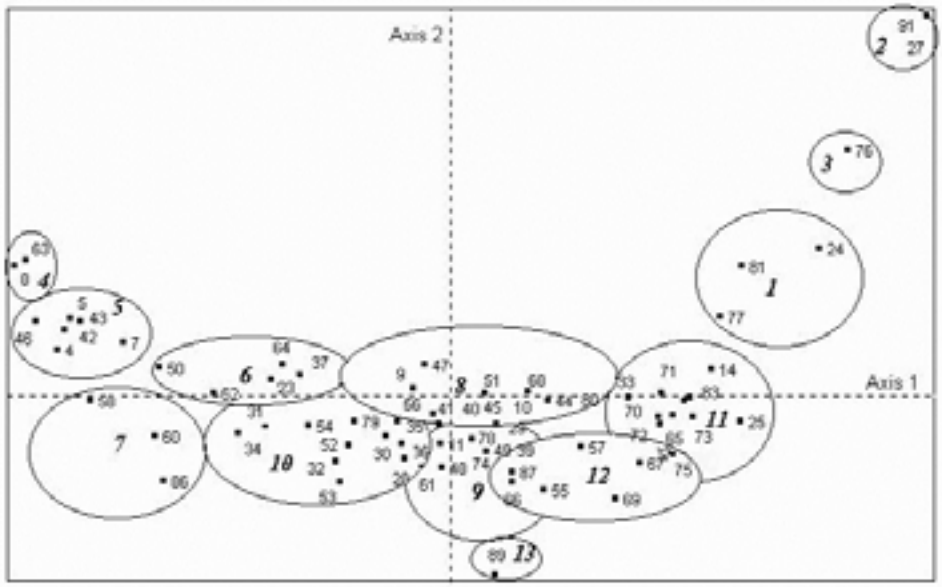


Figure 3. Correspondence analysis ordination of species for axes 1 and 2. The labels of groups of species according to the chosen partitioning are in bold italics. The horizontal scale is doubled for clarity.

Table 4. Interactions between nodes of the relevè hierarchy (columns) with nodes of the species hierarchy (rows). The four test values in each cell show the associations between row branches a_j and b_j and column branches a_i and b_i. In both headings and label cells, the two branches of each node are given, with relative weight. In each cell, test values and significance level are provided. In the table, weights and *p*-values are multiplied by 1000 and test values by 10. *p*-values must be compared to 5% and test values must be compared to ±1.96, corresponding to the 5% significance level.

	ni	49		48		47		46		45		44	
nj aj waj	ai bi	40	48	46	47	39	45	43	44	27	41	42	18
bj wbj	wai wbi	71	928	576	351	160	191	215	361	72	119	325	35
151 149 27	Test value	999	-55	-11	-30	-11	-23	20	-31	-8	-14	-27	0
150 972	Test value	-55	999	16	37	19	999	-15	38	999	999	35	999
	P-value		0		186		683		5		1000		902
150 147 111	Test value	-22	30	-88	999	29	999	-47	-57	999	999	-53	-9
148 861	Test value	-15	19	999	-76	-18	-69	33	999	-45	-44	999	21
	P-value		78		0		0		854		188		898
149 136 18	Test value	35	-28	2	-20	-4	-16	26	-21	-4	-9	-18	2
144 9	Test value	999	-48	-19	-16	-5	-7	0	-17	2	-2	-15	7
	P-value		0		585		775		247		1000		932
148 145 660	Test value	-23	26	2	12	20	-2	-18	19	-5	2	16	11
146 201	Test value	15	-12	999	-90	-42	-73	48	32	-45	-51	32	5
	P-value		61		0		683		35		932		571
147 140 48	Test value	-15	999	-77	999	15	999	-37	-53	52	48	-49	-6
142 62	Test value	-11	19	-47	999	26	43	-26	-27	26	32	-26	0
	P-value		707		16		8		769		173		910
146 132 115	Test value	29	-25	51	-74	-41	-52	41	14	-31	-36	17	-3
143 85	Test value	-10	17	52	-45	-14	-42	22	31	-25	-28	27	14
	P-value		4		712		457		159		966		352
145 135 354	Test value	9	-7	0	-2	1	-3	0	1	3	-5	6	-12
141 305	Test value	-36	40	3	16	20	2	-18	19	-6	9	10	25
	P-value		8		508		555		190		330		38
144 77 5	Test value	999	-40	-18	-9	0	-1	-3	-9	6	2	-8	11
76 3	Test value	33	-21	-2	-6	2	0	11	-6	8	4	-5	12
	P-value		90		723		1000		457		1000		1000
143 139 83	Test value	-10	16	51	-43	-13	-42	24	28	-24	-28	27	9
89 2	Test value	10	999	999	-2	5	3	2	999	10	7	12	30
	P-value		863		713		833		143		968		1
142 130 44	Test value	-5	14	-32	41	14	36	-17	-17	33	17	-13	-5
137 17	Test value	-3	999	-34	47	29	25	-17	-19	-3	34	-25	14
	P-value		726		122		236		953		0		125
141 131 175	Test value	-19	24	17	-4	8	-12	-4	22	-15	0	14	22
138 129	Test value	-26	33	-13	29	19	18	-17	2	9	15	0	11
	P-value		562		23		548		927		608		703
140 86 7	Test value	3	999	-25	999	-3	999	-7	-14	33	28	-13	8
126 41	Test value	-13	999	-70	999	21	999	-33	-48	45	41	-44	-4
	P-value		1000		1000		27		1000		477		1000

Examining the second table (Table 5) identifies the significant differences of association between species nodes n=($p1$,$p2$) (considered in this case as unions $p1 \cup p2$) and each of the sub-clusters ($q1$, $q2$) defining the relevé nodes (the reciprocal table was calculated but is not shown here). Each row corresponds to a node of the species hierarchy with its two branches and corresponding relative weights; each column corresponds to a node of the relevé hierarchy with its branches and relative weights. In each cell, shares of indices $\delta_{p(q1,q2)}$ are given, together with the corresponding test-values and the interaction *p*-value. The two tables may be used jointly for a mutual interpretation of both hierarchies.

The inspection of Table 4 reveals several significant interactions between the nodes of the two hierarchies, up to the 8 × 16 cross-classification. Lower level significant interactions were considered uninteresting, since they seemed too isolated or limited to too small groups. Based on our former observations, we limited the study to cross-classifications up to 7 × 13. In the following discussion, their number, derived from the HC procedure, will indicate the nodes. The association with the number of partition groups, represented in Figures 2 and 3 will be given in the structured table (Table 6).

Table 4 suggests that the highest node in the relevé hierarchy, node 49 = (40,48) has opposite highly significant effects on the branches of the highest node of the species hierarchy, node 151 = (149,150). From Table 5 it is seen that the influence of node 49 = (40,48) can be noticed at the 5% level. In fact, the species belonging to branches 150, 147, 145, 141, and 140 are rare or entirely absent from group 1 (40) of relevés, whereas the species of branches 149 and 144 are abundant in the same relevés.

Considering relevé node 48 = (46,47) in Table 4, we observe significant interactions with species nodes 150 = (147,148), 148 = (145,146), 147 = (140,142), and 141 = (131,138). In Table 5, one may notice that the relevés of

Table 5. Influences of the nodes of relevè hierarchy (columns) on the nodes of the species hierarchy (rows). The two test values in each cell show the associations between row nodes n_j and column branches a_i and b_i. In both headings and label cells, the two branches of each node are given, with relative weights. In each cell, (called delta in the table) shares are listed, together with test values and significance level. In the table, weights, deltas, and p-values are multiplied by 1000 and test values by 10. p-values must be compared to 5% and test values must be compared to ±1.96, corresponding to the 5% significance level.

nj	aj waj bj wbj		ni ai bi wai wbi	49 40 48 71 928	48 46 47 576 351	47 39 45 160 191	46 43 44 215 361	45 27 41 72 119	44 42 18 325 35
151	149 27 150 972			1000 1000	1000 1000	1000 1000	1000 1000	1000 1000	1000 1000
150	147 111 148 861	Delta Test values P-value		796 985 -55 999 0	979 996 16 37 203	991 1000 19 999 703	950 996 -15 38 4	1000 1000 999 999 1000	995 1000 35 999 941
149	136 18 144 9	Delta Test values P-value		203 141 999 -55 0	20 3 -11 -30 186	8 0 -11 -23 703	49 3 20 -31 4	0 0 -8 -14 1000	4 0 -27 0 941
148	145 660 146 201	Delta Test values P-value		777 867 -15 19 38	954 725 999 -76 0	801 662 -18 -69 0	932 967 33 999 317	618 688 -45 -44 238	967 962 999 21 954
147	140 48 142 62	Delta Test values P-value		18 118 -22 30 12	25 270 -88 999 0	190 337 29 999 0	18 29 -47 -57 723	381 311 999 999 187	28 37 -53 -9 894
146	132 115 143 85	Delta Test values P-value		277 195 15 -12 158	291 37 999 -90 0	66 13 -42 -73 289	337 263 48 32 63	0 22 -45 -51 757	268 222 32 5 502
145	135 354 141 305	Delta Test values P-value		500 672 -23 26 11	662 687 2 12 498	735 648 20 -2 135	595 703 -18 19 21	618 666 -5 2 556	699 740 16 11 661
144	77 5 76 3	Delta Test values P-value		111 1 999 -48 0	2 0 -19 -16 701	0 0 -5 -7 1000	6 0 0 -17 499	0 0 2 -2 1000	0 0 -15 7 1000
143	139 83 89 2	Delta Test values P-value		37 89 -10 17 217	128 26 52 -45 0	49 6 -14 -42 218	128 128 22 31 982	0 11 -25 -28 820	126 148 27 14 697
142	130 44 137 17	Delta Test values P-value		18 65 -11 19 336	25 131 -47 999 0	115 144 26 43 314	18 29 -26 -27 645	145 144 26 32 980	28 37 -26 0 890
141	131 175 138 129	Delta Test values P-value		92 321 -36 40 0	309 342 3 16 364	380 310 20 2 220	245 347 -18 19 24	254 344 -6 9 260	329 518 10 25 42
140	86 7 126 41	Delta Test values P-value		0 52 -15 999 41	0 139 -77 999 0	74 193 15 999 0	0 0 -37 -53 999	236 166 52 48 57	0 0 -49 -6 999

branch 47 contain species of nodes 147, 142 and 140, whereas in branch 46 these species are nearly absent (142) or absent (140). In the same way, it appears that the relevés of branch 46 (contrary to those of branch 47) contain species of branch 146, which explains the significant difference of associations observed between branch 146 and each of the two branches (46,47).

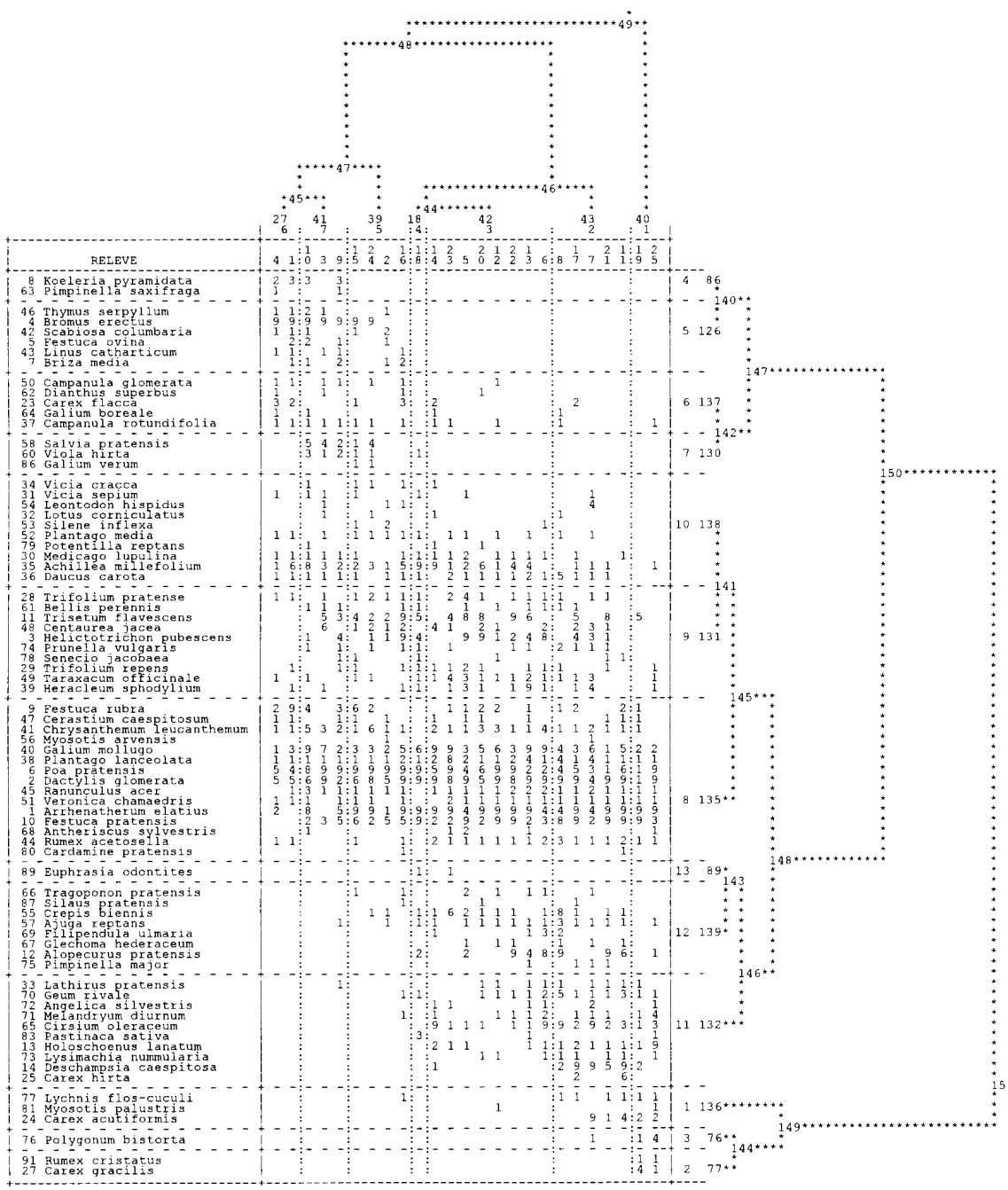
Considering now relevé node 47 = (39,45), strong interactions are found with the species nodes 150 = (147,148) and 147 = (140,142). The explanation comes from the more abundant number of species of branches 147 and 140 observed on branch 45 of relevès.

As regards relevé node 46 = (43,44), significant interactions are noticed with nodes 151 = (149,150) and 148 = (145,146). The first interaction was already explained and the second can be interpreted from the more important number of species of branch 145, present in the branch 44 of relevés.

A study of some remaining interactions completes the discussion. Node 45 = (27,41) has an interaction with the node 142 = (130,137) coming from the absence of species of branch 137 in relevé branch 27. The node 44 = (42,18) has two significant interactions with species nodes 145 = (135,141) and 143 = (139,89). This is explained by the poor number of species of group 8 (135) in relevé number 18 and by the presence of the species *Euphrasia odontites* (89) only in relevé 18, but in none of the branches of 42.

In summary, the partition of seven groups of relevés reveals significant interactions with all the 13 groups of species. A deeper investigation did not seem of higher interest and as well one may reduce the number of groups, according to both the discussed choice of cut-levels and the phytosociological interpretation.

Table 6. Structured data table according to the 7 × 13 partition based on correspondence analysis coordinates. Groups are rearranged according to AOC axis 1 coordinates. Species and relevès are arranged within groups according to CA axis 1 coordinates. In the dendrograms, nodes are indicated by the corresponding node numbers and near to both row and column groups are the group numbers according to the chosen partition.



The previous observations suggest now which interactions must be taken into account. Besides, the inspection of the structured table (Table 6), based on the information resulting from the two tables with the indices of interactions, becomes easier. The first evident interaction concerns node 49 = (40,48) of relevés and node 151 = (149, 150) of species. The relevés of branch 48 are very poor of species of branch 149 (among others, *Carex acutiformis, Polygonum bistorta,* and *Carex gracilis*), whereas they are present in the two relevés of group 1 (branch 40), poor of the species of the branch 150. In fact,

only species of group 8 (135) are also present (but these are ubiquitous, as *Arrhenatherum elatius*, *Poa pratensis, Dactylis glomerata*) and those of group 11 (132): *Cirsium oleraceum, Geum rivale, Melandrium diurnum, Deschampsia caespitosa*, etc. Relevé node 49 interacts with species node 149 = (136,144), since species of branch 144 (*Polygonum bistorta, Rumex cristatus* and *gracilis*) are present only in relevés of branch 40, whereas some presence of those of branch 136 may be found in groups at node 48 (*Lychnis flos-cuculi, Myosotis palustris, Carex acutiformis*).

The second interaction of interest concerns the node pairs 48 = (46,47) and 150 = (147,148). Species of branch 147, such as *Bromus erectus, Koeleria pyramidata, Carex flacca*, etc., are present only in the relevés of branch 47. Most of the species of the branch 148 are present in both relevé branches, but another interaction should be considered now, between nodes 48 = (46,47) and 148 = (145,146). In fact, only species of branch 145 are actually present in node 48, whereas those of branch 146 (*Cirsium oleraceum, Euphrasia odontites, Ajuga reptans, Alopecurus pratensis, Holoschoenus lanatum,* etc.) are present nearly only in the branch 46. Considering the described interactions, a 3 × 5 cross-classification makes sense. Continuing the discussion of the results, one may find interactions concerning all nodes of both hierarchies up to the described 7 × 13 cross-classification. This describes in finer detail the structure of the studied table and justifies the choice to maintain Camiz's (1994) partitions in our description: it will be the ecologist's decision to choose the appropriate level of detail.

Discussion

Interpreting the results of classical exploratory data analysis, based on ordination and classification, requires sophisticated tools. The attempt of Lebart et al. (1979) in proposing a complete set of such tools, however, is not useful in the case of cross-classifications since only one partition of individuals is taken into account at a time and interpretation is based on the behaviour of single variables in groups. The exact conditional tests proposed here are based on the multiple hypergeometric law and take into account hierarchies of variables and objects simultaneously. They are applicable to cross-classifications of any contingency data table. In addition, the associated statistical tests allow selecting the interaction level of the highest significance, thus revealing the mutual relations among rows and columns of the table.

Feoli and Orlóci's (1979) AOC is a useful tool for the quick inspection of the sharpness of a cross-classification, for measuring the quality of a restructured vegetation table, and for showing correspondence between the groups of both partitions. The ability to represent the results with a classical ordination scatter diagram is helpful, together with its use to identify *lattices* of background environmental factors. Nevertheless, as all exploratory ordination methods, it is limited to *suggesting* the possible relations, rather than to *hypothesis testing*. The exact conditional tests associated to the proposed indices allow a more precise and reliable estimation of influences and interactions, thus enabling the scholar to be more certain of his results.

As proposed here, the joint application of AOC and the exact conditional tests allowed a very clear description of Ellenberg's data table. AOC keeps its place as a tool for the analysis of vegetation tables, since the decisions concerning the number of factors, the evaluation of the sharpness of the results, and the overall quality of the cross-classification remain its advantages. It is interesting to observe that the inclusion of the new tests to clustering in classical data analysis procedures (Camiz 1994) adds elements of *hypothesis testing* in the otherwise exploratory frame. Concerning the choice of optimal cut-levels in particular, it is evident that any exploratory method based on the fusion level sequence may provide only a first guess. Tests by the examination of the interactions are required. In addition to interactions detected by the data analyst, the ecological meaning of the interactions should always be considered.

We kept here the structured table rearranged according to the sequence of groups along the first AOC axis observing the constraints of contiguity established by the dendrograms, and in each group rearranging items according to their position along the first axis of CA. This may be refined by the information obtained through the proposed statistics, since one may choose to exchange the branches of each node according to the influence of nodes of the other hierarchy. Such a method of organisation should be made automatic, a goal of our future research, together with the development of a better method of presentation of the results. In fact, up to now the program printout is rather hard to read and calls for a better graphical presentation of results. This should be done automatically, by introducing a grey scale of cell patterns tied to the level of significance of the interaction as usual in trellis diagrams. Both graphical improvements should be helpful for an average user.

Extensions of these tools to other kinds of data tables are currently under study, concerning in particular the three- and multi-way tables, as well as classical individual × variables tables. The aim is a complete set of interpretation aids, suitable for different data structures.

Acknowledgements. This paper was written with the grant of Facoltà d'Architettura dell'Università di Roma "La Sapienza" and of Socrates program of Université des Sciences et Technologies de Lille. Thanks are due to the two referees for their careful reviewing and useful suggestions and to J. Podani for his fruitful suggestions and patience.

References

Agresti, A. 1990. *Categorical Data Analysis*. Wiley, New York.

Anderberg, M.R. 1973. *Cluster Analysis for Applications*. Academic Press, New York.

Anderson, T.W. 1984. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

André, H.M. 1988. Variable centered methods and community classification. *Coenoses* 3(2): 69 78.

Benzécri, J.P. et coll. 1973-82. *L'Analyse des Données*. 2 vol., Dunod, Paris.

Bertin, J. 1977. *La graphique et le traitement graphique de l'information.* Flammarion, Paris.

Camiz, S. 1988. Expert systems: utility in community studies and examples. *Coenoses* 3(1): 33-39.

Camiz, S. 1991. Reflections on spaces relationships in ecological data analysis: effects, problems, possible solutions. *Coenoses* 6(1): 3-13.

Camiz, S. 1993. Computer assisted procedures for structuring community data. *Coenoses* 8(2): 97-104.

Camiz, S. 1994. A procedure for structuring vegetation tables. *Abstracta Botanica* 18(2): 57-70.

Camiz, S. and J.J. Denimal. 1998a. Interpretation of a cross-classification: a new method and an application. In: A. Rizzi, M. Vichi and H.-H. Bock (eds), *Advances in Data Science and Classification*. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin. pp. 555-560.

Camiz, S. and J.J. Denimal.1998b. A new method for cross-classification analysis of contingency data tables. In: R. Payne and P. Green (eds.), *Compstat 98 - Proceedings in Computational Statistics,* Physica-Verlag, Heidelberg: pp. 209-214.

Denimal, J.J. 1997. Aides à l'interpretation mutuelle de deux hiérarchies construites sur les lignes et les colonnes d'un tableau de contingence. *Revue de Statistique Appliquée* 45(4): 93-110.

Denimal, J.J. and S. Camiz. 2001. Exact conditional tests for a reciprocal interpretation of hierarchical classifications built on a two-way table. *Metron* (in press).

Diday, E. 1971. La méthode des nuées dynamiques. *Revue de Statistique Appliquée* 19(2): 19-34.

Feoli, E. and L. Orlóci. 1979. Analysis of concentration and detection of underlying factors in structured tables. *Vegetatio* 40:49-54.

Gauch, H.G. and R.H. Whittaker. 1981. Hierarchical classification of community data. *J. Ecol.*69: 537-557.

Gordon, A.D. 1998. How many clusters? An investigation on five procedures for detecting nested cluster structure. *Proceedings of the IFCS-96 Conference,* Kobe. pp. 109-116.

Govaert, G. 1984. *Classification croisée*. Université de Paris VI, Thèse d'état.

Greenacre, M. 1988. Clustering the rows and columns of a contingency table. *J. Classif.* 5: 39-51.

Hardy, A. and J.F. Deschamps. 1999. Apport du critère des hypervolumes à la détermination du nombre de classes en classification automatique. *XXXI[e] Journées de Statistique - Résumés*. Société Française de Statistique, Grenoble. pp. 103-106.

Hartigan, A. 1975. *Clustering Algorithms*. Wiley, New York.

Hill, M.O. 1973. Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.* 61: 237-249.

Hill, M.O. 1974. Correspondence analysis: a neglected multivariate method. *Appl. Stat.* 23: 340-354.

Hirotsu, C. 1983. Defining the pattern of association in two-way contingency tables. *Biometrika* 70: 579-589.

Kendall, D.G. 1971. Seriation from abundance matrices. In F.R. Hodson, D.G. Kendall, and P. Tautu (eds.), *Mathematics in the Archaeological and Historical Sciences.* Edinburgh University Press, Edinburgh. pp. 215-252.

Lauro, N. and L. D'Ambra. 1984. L'analyse non symmétrique des correspondances. In: E. Diday et al. (eds.), *Data Analysis and Informatics*, Elsevier, Amsterdam, vol. III. pp. 433-446.

Lebart, L., A. Morineau and J.P. Fénelon. 1979. *Traitement des données statistiques*. Dunod, Paris.

Lebart, L., A. Morineau and M. Piron. 1995. *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

Lebart, L., A. Morineau and N. Tabard. 1977. *Techniques de la Description Statistique*. Dunod, Paris.

Lebart, L., A. Morineau and K. Warwick. 1984. *Multivariate Descriptive Statistical Analysis*. Wiley, New York.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the V. Berkeley Symposium 1965. pp. 281-297.*

Milligan, G.W. and M.C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159-179.

Mojena, R. and D. Wishart. 1980. Stopping rule for ward's clustering method. *Compstat '80 - Proceedings in Computational Statistics*. Physica Verlag, Wien. pp. 426-432.

Morineau, A. and P. Alevizos. 1992. Tests et Valeurs tests. Application à l'étude des mastics dans la fabrication de vitraux. *Revue de Statistique Appliquée* 40(4): 27-43.

Müller-Dombois, D. and E. Ellenberg. 1974. *Aims and Methods of Vegetation Ecology*. Wiley, New York.

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

Orlóci, L. 1978. *Multivariate Analysis in Vegetation Research*. 2nd ed. Junk. The Hague.

Orlóci, L. and N. Kenkel. 1987. *Data Analysis in Population and Community Ecology*. Department of Plant Sciences, the University of Western Ontario, London, Canada.

Podani, J. and E. Feoli. 1991. A general strategy for the simultaneous classification of variables and objects in ecological data tables. *J. Veg. Sci.* 2: 435-444.

Podani, J. 2000. *Introduction to the Exploration of Multivariate Biological Data*. Backhuys, Leiden.

Rao, C.R. 1964. The use and interpretation of principal components analysis in applied research. *Sankya* 26: 329-357.

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

Wildi, O. and L. Orlóci. 1990. *Numerical Exploration of Community Patterns*. SPB, The Hague.