RAIRO. RECHERCHE OPÉRATIONNELLE

I. C. LERMAN

F. ROUXEL

Comparing classification tree structures : a special case of comparing q-ary relations II

RAIRO. Recherche opérationnelle, tome 34, n° 3 (2000), p. 251-281

<http://www.numdam.org/item?id=RO_2000__34_3_251_0>

RAIRO Operations Research
RAIRO Oper. Res. **34** (2000) 251-281

COMPARING CLASSIFICATION TREE STRUCTURES: A SPECIAL CASE OF COMPARING q-ARY RELATIONS II (*)

by I.C. LERMAN ([1]) and F. ROUXEL ([1])

Abstract. – *Comparing q-ary relations on a set $\mathcal{O}$ of elementary objects is one of the most fundamental problems of classification and combinatorial data analysis. In this paper the specific comparison task that involves classification tree structures (binary or not) is considered in this context. Two mathematical representations are proposed. One is defined in terms of a weighted binary relation; the second uses a 4-ary relation. The most classical approaches to tree comparison are discussed in the context of a set theoretic representation of these relations. Formal and combinatorial computing aspects of a construction method for a very general family of association coeficients between relations are presented. The main purpose of this article is to specify the components of this construction, based on a permutational procedure, when the structures to be compared are classification trees.*

Keywords: Classification tree, relations, mathematical representation, random permutational model.

PRELIMINARIES

This paper is the direct continuation of the one published with the same title in a preceding issue of this journal (see RAIRO Oper. Res. (1999) 339-365). The first article gives a general methodology for building association coefficients between classification trees, interpreted in terms of specific combinatorial relations on an object set $\mathcal{O}$. Known association coefficients for the concerned comparison, are situated with respect to this general methodology. Otherwise, we have shown in the context of this methodology the specific interest of the valuation given by the mean rank function defined on the ultrametric preordonnance associated with a classification tree (labelled and ranked dendrogram) (see Sects. 3.1 and 4 of the previous paper). In this context we have proposed a new correlation coefficient $\rho(\alpha, \beta)$ (see 52)).

(*) Received April, 1997.
([1]) Irisa - Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France.

Recherche opérationnelle/Operations Research, 0399-0559/00/03/$ 7.00

Instead of the expression (53) let us consider here a corrected and more accurate version of $\rho(\alpha, \beta)$:

$$\rho(\alpha, \beta) = \frac{1}{n(n-1)} \times \frac{s(\alpha, \beta) - \frac{1}{4} p(p+1)^2}{\sqrt{V_\alpha V_\beta + \frac{1}{2n}(W_\alpha - 2V_\alpha)(W_\beta - 2V_\beta)}},$$

where

$$p = n(n-1)/2, \quad V_\omega = A_\omega - \frac{1}{4}(p+1)^2 \quad \text{and} \quad W_\omega = B_\omega - \frac{1}{4}(p+1)^2,$$

where

$$A_\omega = \frac{1}{n(n-1)^2} \sum_i \left[\sum_{j \neq i} \lambda_\omega(i, j) \right]^2$$

and

$$B_\omega = \frac{1}{n(n-1)} \sum_{i \neq j} \lambda_\omega^2(i, j)$$

with

$$\omega = \alpha \text{ or } \beta.$$

As announced at the end of the first paper, we have now, to precisely establish a new coefficient between two classification trees according to the described general construction method, but in adopting new mathematical representation (coding) of a classification tree. Let us recall (see the end of Sect. 3.1 of the preceding paper) that this coding consists of representing a classification tree on an object set $\mathcal{O}$, by a specific subset of $\mathbb{P} \times \mathbb{P}$, where $\mathbb{P}$ designates the set of unordered object pairs from $\mathcal{O}$. One interest of this paper consists of clearly specifying the components of the combinatorial and algorithmic aspects for the exact calculation of the mean and variance of the random raw coefficient (see below). Another substantive point concerns the simulation of the probability distribution of the random standardized coefficient that we have denoted by $Q_4(\alpha, \beta^*)$ (see the end of Sect. 4 of the preceding paper). The paper concludes by evocating the most general case of comparing q-ary relations on an object set $\mathcal{O}$, where q is a given arbitrary integer.

For convenience reasons, I will denote below by I the first paper.

1. PERMUTATIONAL APPROACH FOR COMPARING CLASSIFICATION TREES; THE SECOND COMPARISON METHOD

1.1. Introduction

We adopt here the strict mathematical representation (coding) of a classification tree ω, described in the last part of Section 3.1 of I. Let us recall once again, that we consider the ultrametric preordonnance $\mathcal{UP}(\omega)$ associated with ω and which is a specific total preorder on the set $\mathbb{P}$ of unordered object pairs (see (15) and above of I). Reconsider here the notations given by (14) and (24) of I for $\mathbb{P}$ and for the set $R(\omega)$ defining the representation of ω. Namely,

$$R(\omega) = \{[(i,j),(i',j')] \mid [(i,j),(i',j')] \in \mathbb{P} \times \mathbb{P} \text{ and } l_\omega(i,j) < l_\omega(i',j')\} \quad (1)$$

where l_ω is the level function defined by the ω tree.

Recall that we have, without risk of ambiguty, denoted by ω the indicator function of ω (see (26) of I that we retake here):

$$\omega((i,j),(i',j')) = \begin{cases} 1 & \text{if } l_\omega(i,j) < l_\omega(i',j') \\ 0 & \text{if not} \end{cases} \quad (2)$$

for every $((i,j),(i'j')) \in \mathbb{P} \times \mathbb{P}$.

In these conditions, the raw similarity index associated with the comparison of two trees α and β, has the following expression

$$s'(\alpha,\beta) = \Sigma \{\alpha(\{i,j\},(\{i',j'\})\beta(\{i,j\},\{i',j'\}) \mid \{i,j\},\{i',j'\}) \in J \times J\} \quad (3)$$

where $J = \{\{i,j\} \mid 1 \leq i \neq j \leq n\}$ is the set of all unordered element pairs of $I = \{1,2,\ldots,i,\ldots,n\}$. J codes $\mathbb{P}$.

As previously and according to a general symmetry property, $s'(\alpha,\beta^*)$, $s'(\alpha^*,\beta)$ and $s'(\alpha^*,\beta^*)$ – where α^* and β^* are independent random trees – are equivalent versions of the same random raw index. Then, let us consider

$$s'(\alpha,\beta) = \sum \alpha(\{i,j\}\ ,\ \{i',j'\})\ \beta\ (\{\tau(i),\tau(j)\},\{\tau(i'),\tau(j')\}) \mid (\{i,j\}\ ,\ \{i',j'\}) \in J \times J\}, \quad (4)$$

where – as usual – τ is a random permutation in the set G_n all permutations on I, equally distributed.

In order to obtain the standardized index

$$Q_4(\alpha,\beta) = \frac{s'(\alpha,\beta) - E[s'(\alpha^*,\beta^*)]}{\sqrt{\text{var}[s'(\alpha^*,\beta^*)]}}, \tag{5}$$

we have to compute the exact values of $E[s'(\alpha,\beta^*)]$ and $E[(s'(\alpha,\beta^*))^2]$; where – as usually – E designates the mathematical expectation.

1.2. Mathematical computing of $E[s'(\alpha^*,\beta^*)]$

Equivalently, consider the computing of mathematical expectation of $s'(\alpha,\beta^*)$. For the latter, it is necessary to decompose $J \times J$ as follows:

$$J \times J = \triangle + G + H \quad \text{(set sum)} \tag{6}$$

where

$$\begin{cases} \triangle = \{(\{i,j\},\{i,j\})\}, \\ G = \{(\{i,j\},\{i,k\})\} \text{ and} \\ H = \{(\{i,j\},\{k,l\})\}. \end{cases} \tag{7}$$

In these expressions, distinct symbols indicate distinct elements of I and we have the following equations:

$$\begin{cases} \text{card}\,(\triangle) = p = n(n-1)/2 \\ \text{card}\,(G) = n(n-1)(n-2) \\ \text{card}\,(H) = n(n-1)(n-2)(n-3)/4. \end{cases} \tag{8}$$

Therefore, $E[s'(\alpha,\beta*)]$ can be written:

$$\begin{aligned} &\sum_G \alpha(\{i,j\},\{i,k\})\, E[\beta(\{\tau(i),\tau(j)\},\{\tau(i),\tau(k)\})] \\ &+ \sum_H \alpha(\{i,j\},\{k,l\})\, E[\beta(\{\tau(i),\tau(j)\},\{\tau(k),\tau(l)\})]; \end{aligned} \tag{9}$$

because, the sum over $\triangle$ vanishes.

By denoting $n^{[x]} = n(n-1) \times \ldots \times (n-x+1)$, for an integer x, the following result can be established (the detailed proof is left to the reader):

THEOREM 1:

$$E[s'(\alpha^*,\beta^*)] = n^{[3]}\, \pi_\alpha(G)\, \pi_\beta(G) \times \frac{n^{[4]}}{4}\, \pi_\alpha(H)\, \pi_\beta(H), \tag{10}$$

where $\pi_\omega(G)$ (resp. $\pi_\omega(H)$) is the proportion of $G-$elements $(\{i,j\},\{i,k\})$ (resp. $H-$element $(\{i,j\},\{k,l\})$) for which i and j are joined strictly before i and k (resp. k and l) in the $\omega\, tree; \omega = \alpha\, or\, \beta$.

We may qualify a G (resp. H) – element, as an "attested" ω G (resp. H) – element; if the latter is counted in the numerator of the above $\pi_\omega(G)$ (resp. $\pi_\omega(H)$) proportion. Hence, the problem arises to have a method for determining the number of attested ω G (resp. H) elements. These numbers depend strongly on the ω tree shape ($\omega = \alpha$ or β). They can be denoted by $n_\omega(G)$ and $n_\omega(H)$; and then, obviously, we have:

$$\pi_\omega(G) = \frac{n_\omega(G)}{n(n-1)(n-2)} \quad \text{and}$$
$$\pi_\omega(H) = \frac{4\, n_\omega(H)}{n(n-1)(n-2)(n-3)}. \tag{11}$$

Clearly, each subtree of ω as (a) (see figure below) does increment $n_\omega(G)$ two unities; one for $(\{i,j\},\{i,k\})$ and one for $(\{i,j\},\{j,k\})$. Then twice the number of such ω subtrees gives $n_\omega(G)$.

On the other hand, each ω subtree of the following forms (b), (c), (d) or (e) (see figure below) intervenes in counting $n_\omega(H)$: once for (b) or (c); but three times for (d) or (e). More explicitly, the contribution of (b) or (c) to $n_\omega(H)$ is given by $(\{i,j\},\{k,l\})$ and the contribution of (d) or (e), by $(\{i,j\},\{k,l\}),(\{i,k\},\{j,l\})$ and $(\{j,k\},\{i,l\})$. Therefore, $n_\omega(H)$ is the total number of subtrees having the forms (b) or (c) with addition of three times the number of trees having the forms (d) or (e).

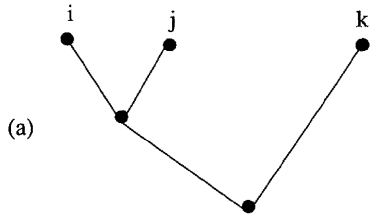

Figure 1.

In Section 3.1 of I we have given a characterization of the tree shape of a classification tree ω by means of what we have called the indexed type of ω and that we have denoted by $t(\omega)$. Clearly $n_\omega(G)$ and $n_\omega(H)$ depend only on $t(\omega)$.

However, it seems very complicated to derive mathematical formula for $n_\omega(G)$ and $n_\omega(H)$. An appropriate solution for this problem is an algorithmic

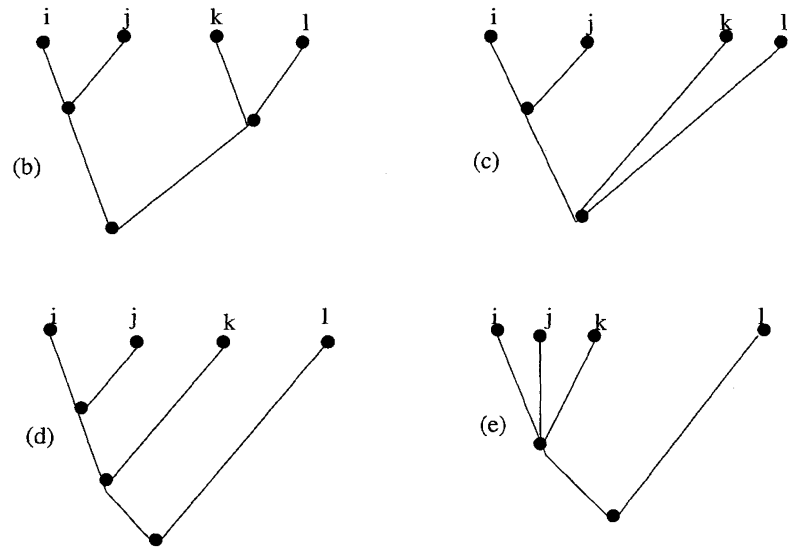

Figure 2.

one. The specified algorithm has to enumerate all the ω subtrees of the above forms (a), (b), (c) and (d). The same type of problem will arise, in much more complicated version, in case of computing the variance of $s'(\alpha^*, \beta^*)$.

1.3. Algorithmic computing of $E[s'(\alpha^*, \beta^*)]$

Consider a classification ω tree (ω has to be instanciated by α and next by β to perform the above calculations). To make clear the general principle of the computing we begin by treating the easiest case of the exact evaluation of $n_\omega(G)$.

In this case, we have to enumerate how many times the above tree form (a) (see Fig. 1) can be retrieved in the whole tree ω. As an example, the following figure (Fig. 3) gives some of the fashions for retreiving the tree form (a) in the following ω tree (Fig. 3). As expressed above (see below) each way of finding (a) in ω is counted two unities in $n_\omega(G)$.

A direct method for the exact evaluation of $n_\omega(G)$ consists of:

(i) generating the set $\mathbb{P}_3(O)$ of all subsets with 3 elements of the object set $\mathcal{O}$;

(ii) considering for each subset $\{x, y, z\}$, the restriction of the ω tree on this subset with 3 elements;

(iii) adding 2 or 0 to an integer variable $N_\omega(G)$, whether the ω subtree on $\{x, y, z\}$ has more than on level or not.

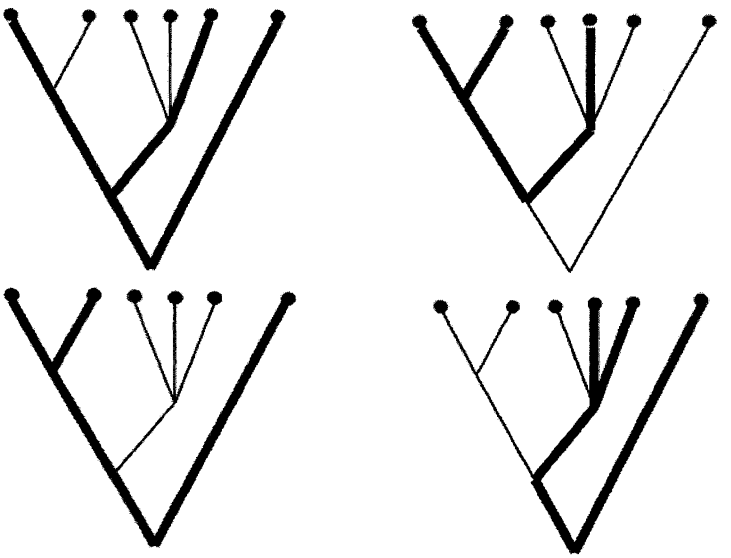

Figure 3. – Some ways for finding the tree form (a) in the below ω tree.

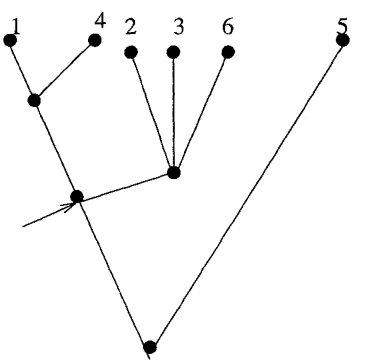

Figure 4. – Tree ω.

Thus, the final value of $N_\omega(G)$ is $n_\omega(G)$. The computational complexity of this procedure is clearly $n(n-1)(n-2)/6$ and then, it order is $\mathcal{O}(n^3)$. This method is qualified as enumerative.

The set up method (Rouxel 1997) is a recursive one. The recursion consists in

(i) positionning the root of the tree form (a) at a given node of the whole tree ω;

(ii) selecting two descendant branches from the concerned node (there might exist only two branches of which this node is the origin);

(iii) developing the left and the right sides of the tree (a) on, respectively, these two branches taken in a certain order; and interverting the respective roles of these two branches for the development.

This development leads to the enumeration of the set of the "attested" ωG elements. For example, the result of the development in placing the root of (a) at the node designated by an arrow (see Fig. 4) is

$$2 \times (1 \times 3 + 3 \times 2) = 18,$$

the first order taken for the two branches being (left, right).

In case of a completely balanced m-ary tree, the order of the computational complexity of this procedure is shown to be $\mathcal{O}(n \log_m n)$. Even more, by preserving some informations at each node of the ω tree, the complexity order can decrease until $\mathcal{O}(n)$ (see the above reference for this point and for all that concern the algorithmic aspects). More completely we have the following table established in case of a completely balanced m-ary tree with n leaves ($n = m^k$ where k is an integer):

	(a)	(e)	(d)	(b)	(c)
direct version	$O(n\log_m(n))$	$O(n\log_m(n))$	$O(n\log_m^2(n))$	$O(n^2\log_m(n))$	$O(n\log_m(n))$
accelerated version	$O(n)$	$O(n)$	$O(n)$	$O(n^2)$	$O(n)$

Figure 5. – Table of the computational complexities.

This table gives the computational complexities of the preceding algorithm for determining how many times a given tree form $((a), (b), (c), (d)$ or $(e))$ can be retrieved in the whole ω tree. Level constraints are included in the tree forms. Thus for (b) the junction between k and l occurs later than that between i and j. Consequently the algorithmic search is more extensive in this latter case than for (e) where the whole tree form results from the fusion of two component forms without any level considerations.

The algorithmic procedure and the above equation (10) have been verified by generating all the $n!$ permutations for moderate sizes of n (*e.g.* $n = 8$).

1.3.1. *Computing of* var$[s'(\alpha^*, \beta^*)]$

We have to evaluate exactly $E[(s'(\alpha^*, \beta^*))^2]$. The direct expression of this mathematical expectation is given by:

$$\begin{aligned}&\sum\{\alpha(\{i,j\},\{i',j'\})\ \alpha(\{i'',j''\},\{i''',j'''\})\\&\times E[\beta(\{\tau(i),\tau(j)\},\{\tau(i'),\tau(j')\})\\&\beta(\{\tau(i''),\tau(j'')\},\{\tau(i'''),\tau(j''')\})]\\&|((\{i,j\},\{i',j'\}),(\{i'',j''\},\{i''',j'''\})\in(J\times J)\times(J\times J)\}.\end{aligned}\tag{12}$$

In order to detect invariance properties, we have to decompose the set $(J \times J) \times (J \times J)$, over which the sum is, according to the structure of $((\{i,j\},\{i',j'\}),(\{i'',j''\},\{i''',j'''\}))$.

This structure is defined from repetitions of I elements in the couple of couples of unordered element pairs of I. Each structure determines a "configuration". As an example consider the following one, where distinct symbols indicate different elements of I:

$$((\{i,j\},\{k,l\}),(\{i,m\},\{j,m\})),\tag{13}$$

it belongs to $H \times G$ (see (51) of I). This configuration defines a class of $H \times G$ which comprises $n(n-1)(n-2)(n-3)(n-4)/2$ elements.

Therefore and first, decompose $(J \times J - \triangle) \times (J \times J - \triangle)$ according to the bipartition of $J \times J - \triangle$ into the two classes G and H:

$$\begin{aligned}&(J\times J-\triangle)\times(J\times J-\triangle)\\&\quad=(G+H)\times(G+H)\\&\quad=G\times G+G\times H+H\times G+H\times H\quad(\text{set sum})\end{aligned}\tag{14}$$

and split each of the four classes into subclasses respectively associated with the different configurations. The detail of all the configurations and the number of represented elements for each of them is explicitly given in Section 3. The table of Figure 6 gives the number of configurations included in each of the above subsets (see (14)).

set	$G \times G$	$G \times H$	$H \times G$	$H \times H$
number of configurations	34	25	25	26

Figure 6. – Table of the configuration numbers.

Now, let us designate by $\mathcal{C}$, the set of all configurations. According to the above table, $\mathcal{C}$ comprises 110 elements; and $\mathcal{C}$ can be generated according

to the above decomposition into four classes. If c is an element of C and $C = C(c)$ the associated subset of $(J \times J - \triangle)^2$; by denoting $m(C)$ the cardinality of C, we may express the following property:

THEOREM 2:

$$E[(s'(\alpha^*, \beta^*)^2] = \sum_{c \in C} m(C)\, \pi_\alpha(c)\, \pi_\beta(c), \tag{15}$$

where $\pi_\omega(C)$ is the proportion of C-elements $[(\{i,j\},\{i',j'\}), (\{i'',j''\},\{i''',j'''\})]$ belonging to $(\mathbb{P} \times \mathbb{P}) \times (\mathbb{P} \times \mathbb{P})$ for which the first and the third pairs ($\{i,j\}$ and $\{i'',j''\}$) are joined strictly before the second and the fourth pairs $\{i',j'\}$ and $\{i''',j'''\}$, in the ω tree, $\omega = \alpha$ or β.

Basically the proof of this property has the same nature as that of Theorem 1. However much more complicated structures intervene for grouping the terms of the above sum (see Sect. 3).

Consequently, we have to enumerate the set of C-elements for which the stated condition of the above theorem, holds. For this purpose, we have to introduce the notion of a c-compatible type of an ω subtree. The number of leaves of the latter is the number of distinct elements which intervene in the c configuration, it is comprised between 3 and 8; 3 in case of $[(\{i,j\},\{i,k\}),(\{i,j\},\{i,k\})]$ type and 8 in case of $[(\{i,j\},\{k,l\}),(\{p,q\},\{r,s\})]$ type. In the latter and as previously, distinct symbol letters indicate distinct elements of I.

On the other hand, it is required for the compatibility condition that there exists at least one element of $C = C(c)$ which can be built from the leaves of the ω subtree and such as the theorem above condition holds.

As an example, consider the following c-configuration which belongs to $H \times G$:

$$[(\{i,j\},\{k,l\}),(\{i,m\},\{j,m\})].$$

First, in order to illustrate the calculation of $m(C)$, notice that we have here

$$m(C) = n(n-1)(n-2)(n-3)(n-4)/2,$$

since there are $\binom{n}{2} \times \binom{n-2}{2}$ fashions for instanciating the first ordered pair of unordered object pairs $(\{i,j\},\{k,l\})$. With respect to this instanciation that we denote by $(\{i_o,j_o\},\{k_o,l_o\})$, there are two possibilities for placing the components of the first pair in, respectively, the two following

pairs; namely $(\{i_o, m\}, \{j_o, m\})$ or $(\{j_o, m\}, \{i_o, m\})$. And finally, we have $(n-4)$ choices for specifying m.

We are going now to illustrate two cases (among others) of compatible trees. For each of them we will give the number of times where the above configuration c is instanciated.

The first compatible tree which is defined on the set $\{x, y, z, u, v\}$, is the following (Fig. 7).

It is easy to see that the subset $\{i, j, m\}$ must be instanciated by $\{x, y, z\}$. On the other hand, the repeated element m, that we can call a pivotal element, is necessarily x or y. And then we have the two following instanciations of c:

$$((\{y, z\}, \{u, v\}), (\{x, y\}, \{x, z\}))$$

and

$$((\{x, z\}, \{u, v\}), (\{x, y\}, \{y, z\})).$$

The second compatible tree which is also represented on the set $\{x, y, z, u, v\}$, has the following form (Fig. 8). It gives rise to eight instanciations of the above c-configuration. To realize that, begin by constituting the right ordered pair of unordered element pairs $(\{i, m\}, \{j, m\})$, where m indicates the pivotal element. For this purpose, we have to choose a subset of size three in the set $\{x, y, z, u\}$. Afterwards we have to choose on among two possible elements. As an example, consider the 3-subset $\{x, z, u\}$, the pivotal elements can be x or z. Therefore, the eight instanciations of the configuration c can

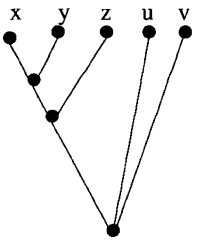

Figure 7.

be expressed as follows:

$$(\{y,z\},\{u,v\}),(\{x,y\},\{x,z\}))$$
$$(\{x,z\},\{u,v\}),(\{x,y\},\{y,z\}))$$
$$(\{y,u\},\{z,v\}),(\{x,y\},\{x,u\}))$$
$$(\{x,u\},\{z,v\}),(\{x,y\},\{y,u\}))$$
$$(\{z,u\},\{y,v\}),(\{x,z\},\{x,u\}))$$
$$(\{x,u\},\{y,v\}),(\{x,z\},\{z,u\}))$$
$$(\{z,u\},\{x,v\}),(\{y,z\},\{y,u\}))$$
$$(\{y,u\},\{x,v\}),(\{y,z\},\{z,u\})).$$

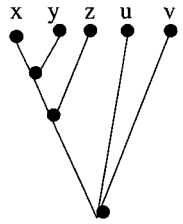

Figure 8.

Therefore, for a given configuration c and an ω tree, the general enumeration method can be decomposed as follows:

- derive all types of c-compatible subtrees;
- for a given type, determine how many subtrees of this type there are, in the whole ω tree;
- for a subtree of a given type, determine how many countable elements of $C(c)$, it doe give rise.

Let us consider one more example for which the number of elements of $C(c)$ associated with a c-compatible ω subtree, is rather big.
Relative to the following c-configuration, belonging to $H \times H$:

$$[(\{i,j\},\{k,l\}),(\{i,p\},\{q,r\})],$$

the following subtree is c-compatible (Fig. 9).

We focus here on the pairs $\{i,j\}$ and $\{i,p\}$ which respectively are the first components of both ordered pairs of object pairs $(\{i,j\},\{k,l\})$ and $(\{i,p\},\{q,r\})$. The subset $\{i,j,p\}$ has necessarily an empty intersection with the subset $\{u,v,w,t\}$. Because if not, it would be impossible to constitute

$\{k,l\}$ or $\{q,r\}$ with the conditions $\{i,j\} < \{k,l\}$ and $\{i,p\} < \{q,r\}$, according to the tree structure. Therefore $\{i,j,p\}$ is identical to $\{x,y,z\}$. In these conditions, there are six possibilities for forming $(\{i,j\},\{i,p\})$; namely:

$(\{x,y\},\{x,z\}), (\{x,y\},\{y,z\}), (\{x,z\},\{x,y\}), (\{x,z\},\{y,z\}), (\{y,z\},\{x,y\})$ and $(\{y,z\},\{x,z\})$.

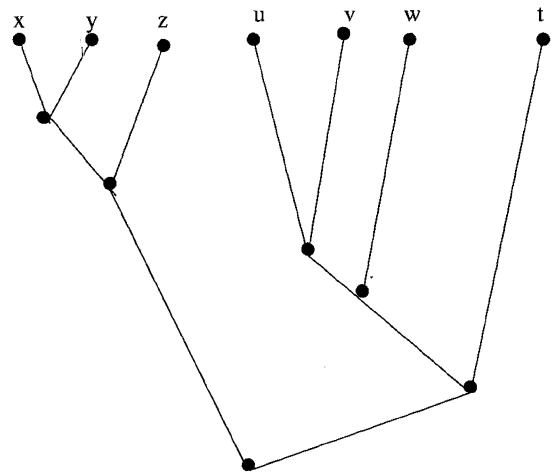

Figure 9.

For each possibility, there are $2 \times \binom{4}{2} = 12$ choices for forming $(\{k,l\},\{q,r\})$, where necessarily $\{k,l,q,r\} = \{u,v,w,t\}$.

Then in all, there are 72 instanciations of the above configuration, from the above tree.

Now, let us denote by $T_\omega(c)$ the set of all ω subtrees types compatible with the c configuration. If $t_\omega(c)$ is a given element of $T_\omega(c)$, we may designate by $n[t_\omega(c)]$ the number of times for which the type $t_\omega(c)$ is instanciated in the whole ω tree. For a given instanciation, $l[t_\omega(c)]$ indicates the number of distinct replications of the c-configuration, which can be obtained in a compatible way, from a given $t_\omega(c)$ subtree. In these conditions, the cardinal – that we denote by $m(\omega, C)$ – which defines the numerator of the ratio $\pi_\omega(c)$, can be put in the following form:

$$\sum_{t_\omega(c) \in T_\omega(c)} n[t_\omega(c)] \times l[t_\omega(c)]. \tag{16}$$

Hence, we may state the subsequent property:

PROPERTY 1: *Relative to a given configuration c, the proportion of c-elements compatible with an ω–tree can be expressed by*

$$\pi_\omega(C) = \frac{m(\omega, C)}{m(C)} = \frac{\sum\{n[t_\omega(c)] \times l[t_\omega(c)] \mid t_\omega(c) \in T_\omega(c)\}}{m(C)} \qquad (17)$$

where the different components of this equation are specified above.

Mathematical expression for $m(C)$ can be provided without great difficulty (see above and Sect. 3). As for the determination problem of $n_\omega(G)$ and $n_\omega(H)$ (see Sect. 1.2), tractable analytical formula for $m(\omega, C)$ depending on the ω tree shape, seems to be very hypothetical to obtain. And that, even characterization is provided in order to capture formally the ω tree shape. Recall that such a characerization has been proposed in the first part of this article (see Sect. 3.1 of I).

1.4. Algorithmic computing of var$[s'(\alpha^*, \beta^*)]$

The problem is reduced to the determination of the number of occurences of a given tree structure (ranked but not labelled dendrogram) on few elements, that we denote here by o (see Fig. 9 for an example), in a whole tree (labelled and ranked dengrogram) denoted by ω.

As for the computing of the mathematical expectation (see Sect. 1.3) a recursive procedure is retained with two main steps. The former consists of assigning the root of the small tree o on a given node of the ω tree (starting by its root) and the later step consists in developing the substrees of o along the respective branches of the whole tree ω.

The purpose of this part of the algorithm is to determine how many times the form of the o tree is encountered in the ω tree, independently of level conditions. For example, by considering the assignment pictured in Figure 10 we obtain $2 \times 4 \times 2 \times 4 \times 2 = 128$ occurences of the o tree form. More formally, we obtain this occurrence number by multiplying the respective numbers of the leaves of the o tree which underly the nodes where the leaves of the o tree are positionned.

However, a given assignment of the o tree shape in the o tree has not to be retained if the level conditions of the o tree are not ordinaly respected in the ω tree. To fix ideas assume the increasing sequence of the levels of the o tree numbered by the first integers 1, 2, 3, ... Thus, the fusion level of two given leaves defines a version of their ultrametric ordinal proximity.

Strictly, the leaves of the o tree have not to be labelled. However for clarity reasons we consider a canonical labelling depending on their mutual ultrametric proximities distributions. More precisely, the first leave labelled by 1 is that for which the variance of its ordinal ultrametric proximities to the other leaves is maximal. The other leaves are then labelled increasingly, according to their ultrametric ordinal proximities to this first leave. The process is recursively repeated for all subsets of equally distant elements.

As a matter of fact the o trees are derived from their ordinal ultrametric matrices. These are generated thanks to a theorem and the associated algorithm given in (Lerman 1981, Chap. 0, Sect. IV.2). This theorem characterizes a reduced form of an ultrametric distance matrix which can be obtained by the described algorithm. The algorithm we propose here, giving to the ordinal ultrametric matrix its canonical form, is in fact a specification of the previous one.

In Section 3 we have specified all the configurations of an ordered pair $[(\{x,y\},\{x',y'\}),(\{x'',y''\},\{x''',y'''\})]$ belonging to $(\mathbb{P}\times\mathbb{P})\times(\mathbb{P}\times\mathbb{P})$ of which each component is by itself an ordered pair of unordered object pairs. We have also given the respective cardinalities of the different configurations. These configurations are grouped into four general categories $G\times G$, $G\times H$, $H\times G$ and $H\times H$ expressed in the above table (see Fig. 6).

As above, for a given configuration $c, C = C(c)$ designates the subset of $(\mathbb{P}\times\mathbb{P})\times(\mathbb{P}\times\mathbb{P})$ corresponding to c. Let us now denote by $N_\omega(C)$ an integer variable representing the number of distinct elements of C met during the algorithm process. The final value of $N_\omega(C)$ is the number $n_\omega(C)$ of $(\mathbb{P}\times\mathbb{P})\times(\mathbb{P}\times\mathbb{P})$ elements having the c configuration and for which the ordinal conditions of Theorem 2, hold.

Consider now an o tree on e elements, associated with an ordinal ultrametric matrix having its canonical form. *A priori* e is comprised between 3 and 8. However as it will be seen just below, it is possible to avoid the treatment of the case $e = 8$. This omission will not have any effect on the exact calculation of the cardinalities $n_\omega(C)$ and will notably reduce the computational complexity. Consider also the set of the configurations c for which exactly e distinct objects intervene. We denote by c_e a current configuration of this set and by C_e the subset of $(\mathbb{P}\times\mathbb{P})\times(\mathbb{P}\times\mathbb{P})$ elements for which the configuration is c_e.

For each assignment of the o tree in the ω tree respecting simultaneously the form and the ordinal level conditions of the o tree, the respective values of the different variables $N_\omega(C_e)$ are incremented. More precisely, each $N_\omega(C_e)$ is incremented by the number of instanciations which can

be obtained from a given assignement of the o tree. Notice that for each such assignement, the leaves of the o tree are identified in the ω tree by objects of $\mathcal{O}$.

As mentionned above we have not to consider directly the contribution of the o trees with 8 leaves. Indeed this contribution can be deduced by difference according to the following argument and the subsequent equation (see (20)).

According to the above notation $n_\omega(H \times H)$ designates the number of $(H \times H)$ elements for which the stated condition of Theorem 2 holds. We clearly have

$$n_\omega[(H \times H) \times (H \times H)] = [n_\omega[(H)]^2 \tag{18}$$

where $n_\omega(H)$ has been defined in Section 5.2.

On the other hand, we have the following decomposition

$$n_\omega[(H \times H)] = \sum\{n_\omega[\mathcal{C}(c)] \mid c \in \mathcal{D}\} \tag{19}$$

by denoting $\mathcal{D}$ the set of all configurations which intervene in $(H \times H)$.

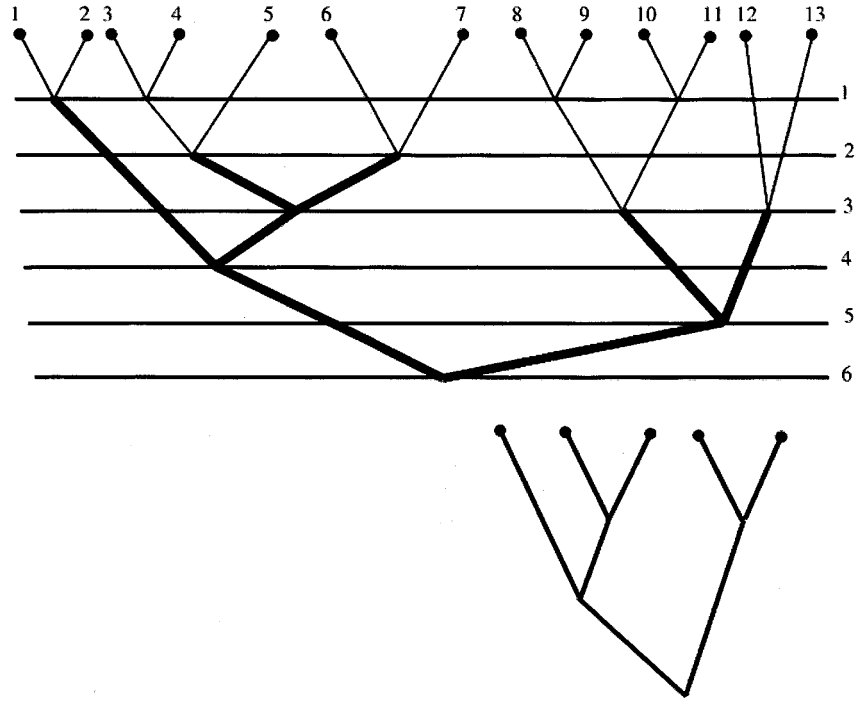

Figure 10. Example of grouped solutions for the assignement.

Otherwise, an o tree with 8 leaves does only contribute with respect to the set C_8 corresponding to the configuration $[(\{i,j\},\{k,l\},\{lm,n\},\{p,q\})]$ which includes 8 distinct objects. Therefore

$$n_\omega(C) = [n_\omega(H)]^2 - \sum \{n_\omega[C_8(c)] \mid c \in \mathcal{D} - \{c_8\}\}. \qquad (20)$$

From the general above considerations we may outline the sequence of the main stages of the computing algorithm:

1. generation of all the ultrametric matrices reduced to their canonical forms on e elements, $3 \leq e \leq 7$. These matrices are grouped by classes according to the number e;
2. association with each matrix an o tree (ranked dendrogram);
3. for a given o tree with e leaves, determine the instanciations of its form in the ω tree;
   - 3.1. for each potential instanciation, check the ordinal condition levels of the o tree in the ω tree. If this condition is not satisfied, delete all the instanciations of this form (see Fig. 10);
   - 3.2. if the preceding condition holds in the ω tree, consider for each configuration c_e the number of its instanciations for one instanciation of the o tree in the ω tree. By denoting $n_o(c_e)$ this number, increment the variable $N_\omega(C_e)$ (see above) by:

$$n_\omega(o) \times n_o(c_e)$$

   where $n_\omega(o)$ is the instanciation number of the o tree in the whole ω tree.

We have analyzed (see Rouxel 1997) the computational complexity of this algorithm by evaluating the number of solutions of what can be considered as the worst case. For this, the shape of the o tree is that of a likecomb tree and the ω tree is a binary completely balanced tree.

As usually we denote by e the number of leaves of the o tree and by n that of the ω tree. If k is the number of levels (or the depht) of the ω tree, then $n = 2^k$; and the j^{th} level includes exactly 2^{k-j} nodes, $0 \leq j \leq k$. On the other hand, the depht of a comblike tree with e leaves, is $e-1$. It is shown (see the above reference) that the number of compatible assignments of the o tree on the ω tree can be approximated by $n(\log_2 n)^{e-2}$. This number vanishes if k is strictly lower than $e-1$.

The following diagram compares the behaviours of the computational complexities of the above proposed solution with the enumerative one of

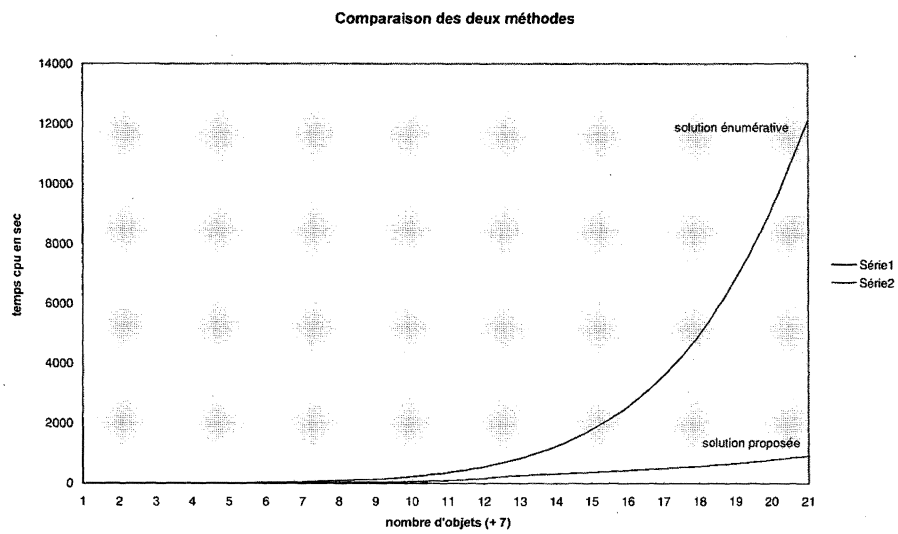

Figure 11. – Comparison between the two algorithmic solutions, the horizontal axis represents the number of objects (+7) and the vertical axis represents the cpu seconds of time.

which the general principle is given in Section 1.3. For this latter solution we have to generate all the sets enumerated in Section 3, except the last one (Sect. 3.3.5).

Nevertheless, because of the large number of the o trees, the computational complexity becomes too high when n increases substancially; for example, for n about 100. Anyway, for our problem of classification trees comparison, one may limit this comparison to the most interesting parts of these trees by only retaining the last levels of both trees. This truncation may consist of deleting the first levels, starting with a significant partition for each classification tree, with more or less the same number of classes (Lerman and Ghazzali 1991). This provides a major simplification in determining $m(\omega, C)$ (see Eq. (17)) by an algorithmic manner. Roughly speaking, the number n is replaced by the number of leaves of the retained tree, where each leave represents an object class.

1.5. Simulations of the probability distributions of $Q_4(\alpha, \beta^*)$

All the n! permutations have been generated for small n. Thus, as it was for the equation (10) giving the mathematical expectation of the random raw index $s'(\alpha, \beta^*)$, that (15) which gives the 2nd absolute moment of this random index, has been exactly verified. Different values of n going from

8 to more than 30 have been considered. For n not enough small, a set of random permutations have been generated as independently as possible.

The objective consists in realizing the general shape of the probability distribution of the new standardized coefficient in comparison with that denoted $Q_\lambda(\alpha, \beta)$, obtained in the previous paper and based on the mean rank coding of a total preoder.

Figure 12 and Figure 13 give respectively the exact probability distributions of $Q_4(\alpha, \beta^*)$ and $Q_\lambda(\alpha, \beta^*)$ in case where α and β are two comblike trees on 8 elements. The vertical hatching density is simply related to the accuracy of unit scale on the horizontal axis. The distribution of $Q_4(\alpha, \beta^*)$ seems more harmoniously balanced than that of $Q_\lambda(\alpha, \beta^*)$. Now, if one wishes to approximate the distribution of $Q_4(\alpha, \beta^*)$ by a probability law having an analytical expression, one may suggest to use an eulerian distribution for $Q_4(\alpha, \beta^*) - \min[Q_4(\alpha, \beta^*)]$.

The two following figures concern respectively the same distributions. Figure 14 is associated with $Q_4(\alpha, \beta^*)$ and Figure 15, with $Q_\lambda(\alpha, \beta^*)$. Here, α and β are two arbitrary tree structures on 30 objects chosen at random and independently. 10000 independent random permutations are generated to simulate the respective probability distributions. The same above remarks hold.

The last two figures concern the case where the α and the β structures correspond to comblike trees on 30 elements. 10000 independent random permutations have generated in order to establish the simulation have generated in order to establish the simulation of the probability distributions of $Q_4(\alpha, \beta^*)$ and $Q_\lambda(\alpha, \beta^*)$. Here also notice the better harmony of the distribution of $Q_4(\alpha, \beta^*)$ with respect to that of $Q_\lambda(\alpha, \beta^*)$. Moreover, these distributions and specially that of $Q_4(\alpha, \beta^*)$ are becoming more alike normal distribution. One reason for this fact is the specificity of the tree structure. The second reason is related to the tree size which is not very small here.

2. COMPARING q-ARY RELATIONS AND CONCLUDING REMARKS

Comparison between q-ary relations is outlined in Lerman (1992). In order to situate the previous development, let us recall the elements of this comparison.

Let $\mathcal{O}^{[q]}$ designate the set of sequences of q objects, mutually distinct. We call such a sequence a q-uple and we indicate it by $(1, 2, ..., i_q)$, where $(i_1, i_2, ..., i_q)$ is a q subset of $I = \{1, 2, ..., n\}$: the set of labels which

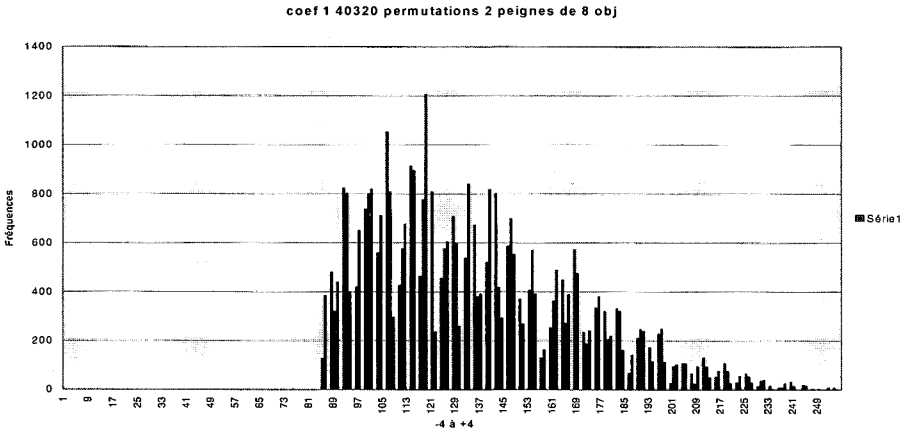

Figure 12. – Probability distribution of $Q_4(\alpha,\beta^*)$ on all the 40320 permutations where α and β are two comblike trees on 8 elements.

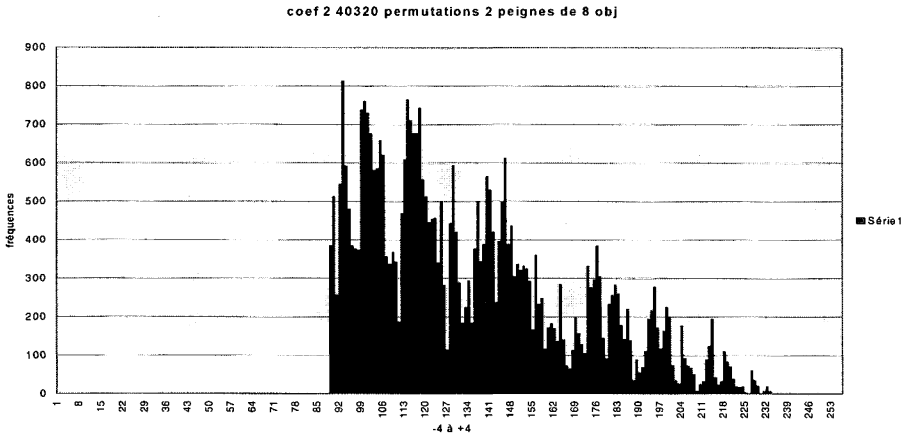

Figure 13. – Probability distribution of $Q_\lambda(\alpha,\beta^*)$ on all the 40320 permutations where α and β are two comblike trees on 8 elements.

codes $\mathcal{O}$. The cardinality of $\mathcal{O}^{[q]}$ is $n(n-1)...(n-q+1)$. For the comparison of two weighted (valued) q-ary relations, denoted

$$\{\mu_{i_1 i_2 \ldots i_q} \mid (i_1, i_2, \ldots, i_q) \in I^{[q]}\}, \tag{21}$$

$$\{\nu_{i_1 i_2 \ldots i_q} \mid (i_1, i_2, \ldots, i_q) \in I^{[q]}\}; \tag{22}$$

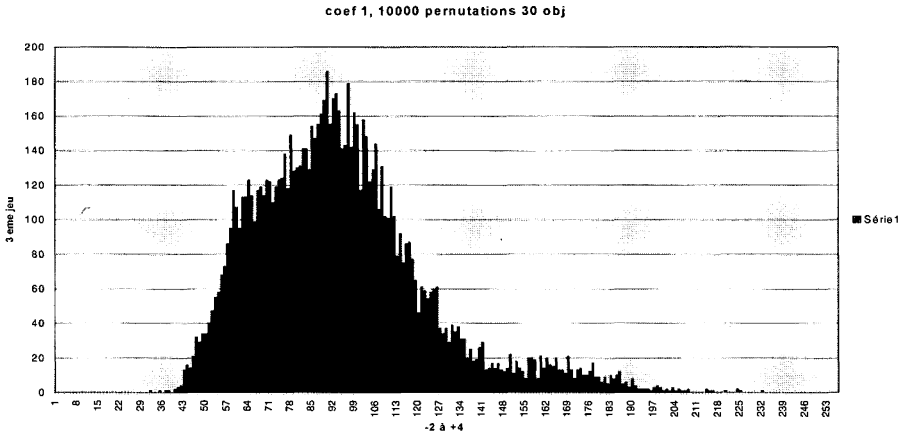

Figure 14. – Probability distribution of $Q_4(\alpha, \beta^*)$ on 10 000 random permutations where α and β are two distinct and arbitrary tree structures on 30 objects.

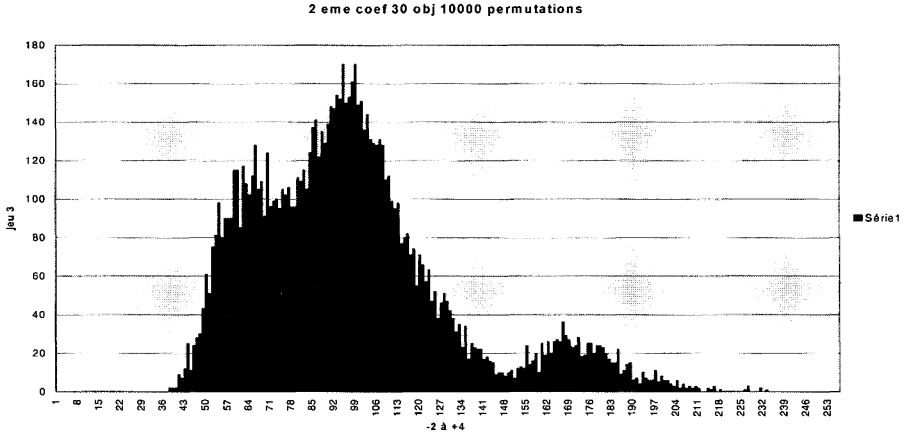

Figure 15. – Probability distribution of $Q_\lambda(\alpha, \beta^*)$ on 10 000 random permutations where α and β are two distinct and arbitrary tree structures on 30 objects.

the raw similarity index takes the following form:

$$s(\mu, \nu) = \sum \{\mu_{i_1 i_2 \ldots i_q} \nu_{i_1 i_2 \ldots i_q} \mid (i_1, i_2, \ldots, i_q) \in I^{[q]}\} \qquad (23)$$

where μ (resp. ν) is a numerical or logical (*i.e.* binary) valuation.

If μ^* and ν^* are independent random valuations, respectively associated with μ and ν, under the permutational model, the random indices $s(\mu, \nu^*)$, $s(\mu^* \nu)$ and $s(\mu^*, \nu^*)$ have the same distribution law.

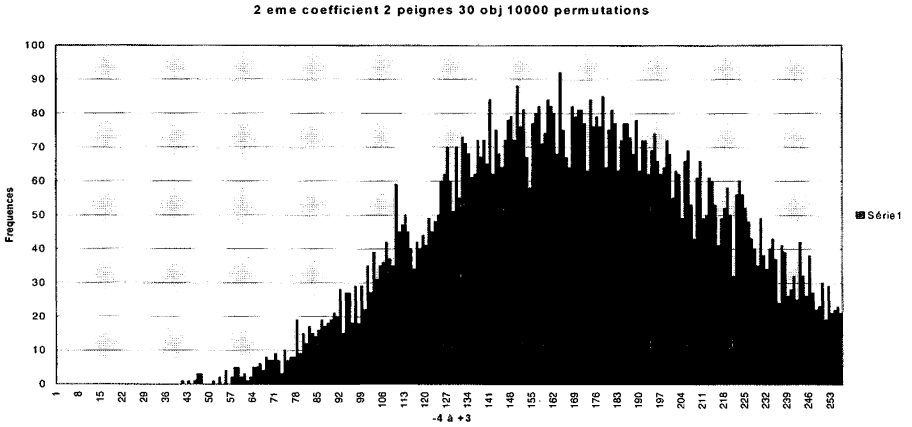

Figure 16. – Probability distribution of $Q_4(\alpha, \beta^*)$ on 10 000 random permutations where α and β are two comblike trees on 30 objects.

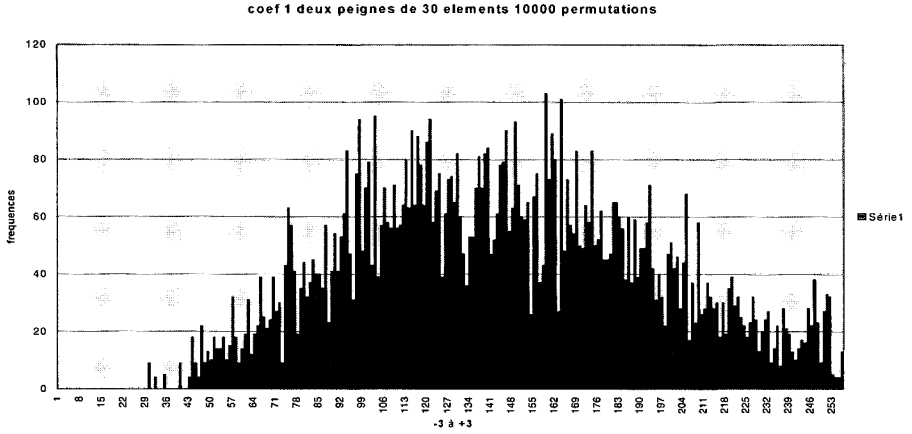

Figure 17. – Probability distribution of $Q_\lambda(\alpha, \beta^*)$ on 10 000 random permutations where α and β are two comblike trees on 30 objects.

The mathematical expectation and the absolute second moment can be expressed as follows.

$$E[s(\mu, \nu^*)] = n^{[q]}\, \bar{\mu}\, \bar{\nu} \tag{24}$$

where we have denoted by $n^{[q]}$, $n(n-1) \times \cdots \times (n-q+1)$ and where

$\bar{\mu}$ ($resp.$ $\bar{\nu}$) designates the mean of the μ (resp. ν) valuation of $O^{[q]}$,

$$E[(s(\mu,\nu^*))^2] = \sum_{0\le r\le q}\sum_{c_r}\frac{1}{n^{[2q-r]}}\left(\sum_{C(c_r)}\mu_{i_1\ldots i_q}\,\mu_{j_1\ldots j_q}\right) \times\left(\sum_{C(c_r)}\nu_{i_1\ldots i_q}\,\nu_{j_1\ldots j_q}\right) \tag{25}$$

where c_r is a configuration of $((i_1,i_2,\ldots,i_q),(j_1,j_2,\ldots,j_q))$ for which r components of $(i_1,i_2,\ldots,i_q)$ are repeated in $(j_1,j_2,\ldots,j_q)$. There are

$$\binom{q}{r}^2 r\,! \tag{26}$$

different configurations c_r. $C(c_r)$ denotes the set of ordered pairs of q-uples $((i_1,i_2,\ldots,i_q),(j_1,j_2,\ldots,j_q))$ having the same configuration c_r. We have

$$\text{card}\,[C(c_r)] = n^{[q]} \times (n-q)^{[q-r]} = n^{[2q-r]}. \tag{27}$$

The total number of configurations is given by

$$\sum_{0\le r\le q}\binom{q}{r}^2 r\,!, \tag{28}$$

its value for $q = 4$ is 209. This number is much greater than the necessary number of configurations (110, see Fig. 6) considered in case of trees comparison. This, because we have taken into account, in the latter case, the specificity of the relations to be associated.

In order to calculate the mathematical expectation and variance of $s(\mu,\nu^*)$ (see Eqs. (24) and (25)), only enumerative algorithmic method can be envisaged. The order of its computational complexity is $\mathcal{O}(n^{2q-1})$. This comes from the fact that the calculation of the cardinality associated with the configuration comprising $2q$ distinct elements can be derived from the cardinalities associated with the other configurations relative to $H_q \times H_q \cdot H_q$ is the set of q-uples including q distinct elements. The argument is analogous to that given in Section 1.4 (see (20)).

The importance of the scale, with respect to which an association coefficient is established, is not enough emphasized in data analysis literature. It is now admited and mainly evocated in the binary case (Hubert 1983; Messatfa 1990), that the numerator of the association coefficient has to be

centralized. The reduction proposed is often based on the maximum of the numerator. This may give rise to very difficult problems of combinatorial optimization (Lerman 1987; Lerman and Peter 1988; Messatfa 1992). In our case and relative to our latter mathematical coding, this leads to the intractable problem of finding the permutation σ which maximizes $s'(\alpha, \beta(\sigma))$ (see (3)). For statistical reasons and according to likelihood linkage analysis (LLA) classification method (Lerman 1993), we have adopted reduction by means of the standard deviation of $s'(\alpha, \beta^*)$. And, we have shown that the computing problem becomes tractable by means of a polynomial algorithmic procedure. Approximating probabilistic distribution of $s(\mu, \nu^*)$ with enough accuracy remains a difficult problem.

3. APPENDIX: STRUCTURAL DECOMPOSITION OF $(G + H) \times (G + H)$

We are going here to make explicit the structural decomposition of $(G + H) \times (G + H)$ (see Sect. 1.4) and then, to justify the content of the table given in Figure 6. On the other hand, we will give the cardinality associated with each substructure defining a given configuration c of an ordered pair of which each component is an ordered pair of unordered object pairs, such as:

$$(\{\{x, y\}, \{z, t\}\}, \{\{x', y'\}, \{z', t'\}\}).$$

An unordered object pair such as $\{x, y\}$ will be denoted here by a word with two letters xy, of which the first letter x precedes lexicographically the second one y.

First recall the general equation (14):

$$(G + H) \times (G + H) = G \times G + G \times H + H \times G + H \times H \quad \text{(set sum)}$$

and let us designate by $\mathcal{U} \times \mathcal{V}$ one of the four subsets of the right member of this equation ($\mathcal{U} = G$ or H and $\mathcal{V} = G$ or H). Our general decomposition strategy consists of organizing the structure of $\mathcal{V}$ with respect to a given element of $\mathcal{U}$. If (ξ, η) belongs to $\mathcal{U} \times \mathcal{V}$, its configuration $c = c(\xi, \eta)$ is conditionned by the manner in which the objects appearing in ξ are repeated in η. The cardinality of c is the number of elements of $\mathcal{U} \times \mathcal{V}$ covered by the configuration c.

3.1. Decomposition of $\mathcal{U} \times \mathcal{V} = G \times G$

Let $\xi = (xy, xz)$ be a given element of G. Consider the set $\{x, y, z\}$ of the three elements which intervene in the constitution of ξ. We have to consider

four general cases according to the number of objects distinct from x, y or z and which intervene in the construction of η. This number can be 0,1,2 or 3; and then, the cases will be denoted according to this number. Now, we are going to give below the different structures of $\eta = (x'y', x'z')$ and the associated cardinalities of $C = C(c)$ where $c = c(\xi, \eta)$ (see Sect. 1.4).

3.1.1. *Structures of η for the case 0*

- (xy, xz), $\text{card}(C) = n(n-1)(n-2)$;
- (xz, xy), $\text{card}(C) = n(n-1)(n-2)$;
- (xy, yz), $\text{card}(C) = n(n-1)(n-2)$;
- (yz, xy), $\text{card}(C) = n(n-1)(n-2)$;
- (xz, yz), $\text{card}(C) = n(n-1)(n-2)$;
- (yz, xz), $\text{card}(C) = n(n-1)(n-2)$.

3.1.2. *Structures of η for the case 1*

- (xy, xu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xu, xz), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xu, xy), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xz, xu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xy, yu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xz, zu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yu, xy), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (zu, xz), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yz, yu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yu, xy), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (zu, xz), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yz, yu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yu, yz), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yz, zu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (zu, yz), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xu, yu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yu, xu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xu, zu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (zu, xu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yu, zu), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (zu, yu), $\text{card}(C) = n(n-1)(n-2)(n-3)$.

3.1.3. *Structures of η for the case 2*

- (xu, xv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (xu, uv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;

- (uv, xv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (yu, yv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (yu, uv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (uv, yu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (zu, zv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (zu, uv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (uv, zu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$.

3.1.4. *Structures of η for the case 3*

- (uv, uw), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)$.

Finally, the number of distinct configurations is 34. On the other hand one may verify that the sum of the cardinalities, which can be put in the following form

$$\begin{aligned} 6n(n-1)(n-2) &+ 18n(n-1)(n-2)(n-3) \\ &+ 9n(n-1)(n-2)(n-3)(n-4) \\ &+ n(n-1)(n-2)(n-3)(n-4)(n-5), \end{aligned}$$

is nothing other than $[n(n-1)(n-2)]^2$ which represents the cardinal of $G \times G$.

3.2. Decomposition of $\mathcal{U} \times \mathcal{V} = G \times H$

As for the preceding Section 7.1, $\xi = (xy, xz)$ will designate a given element of G. We also distinguish here four cases according to the number of elements of the set $\{x, y, z\}$ which intervene in the building of the element η belonging to H. Let us denote by case i, the case for which $(3-i)$ elements of $\{x, y, z\}$ are repeated in η.

3.2.1. *Structures of η for the case 0*

- (xy, zt), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (zt, xy), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (xz, yt), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yt, xz), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (yz, zt), $\text{card}(C) = n(n-1)(n-2)(n-3)$;
- (zt, yz), $\text{card}(C) = n(n-1)(n-2)(n-3)$.

3.2.2. *Structures of η for the case 1*

- (xy, tu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- (tu, xy), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- (xz, tu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;

- (tu, xz), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- (yz, tu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- (tu, yz), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- (xt, yu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (yu, xt), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (xt, zu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (zu, xt), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (yt, zu), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- (zu, yt), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$.

3.2.3. *Structures of η for the case 2*

- (xt, uv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- (uv, xt), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- (yt, uv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- (uv, yt), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- (zt, uv), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- (uv, zt), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$.

3.2.4. *Structures of η for the case 3*

- (tu, vw), $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$.

One may verify that the sum of the above cardinalities of the 25 categories is equal to

$$n(n-1)(n-2) \times (1/4)n(n-1)(n-2)(n-3) = \text{card}(G \times H).$$

It is obvious that the decomposition of $H \times G$ is structurally analogous to that of $G \times H$.

3.3. Decomposition of $\mathcal{U} \times \mathcal{V} = H \times H$

Let $\xi = (xy, zt)$ be a given element of $\mathcal{U} = H$ and let us designate by $E(\xi)$ the set $\{x, y, z, t\}$ including the four elements which appear in ξ. $D(\xi)$ will indicate the complementary subset of $E(\xi)$ and we have $\text{card}(D(\xi)) = n-4$.

As previously, the structural decomposition of $\mathcal{V} = H$ will be elaborated according to the repetitions of x, y, z or t, in the components of the element $\eta = (x'y', z't')$ which belongs to $\mathcal{V} = H$. But in this situation the respective roles of x and y (resp. z and t) are equivalent. To illustrate this point, consider the two following elements of $H \times H$:

$$((xy, zt), (xz, yt)) \quad \text{and} \quad ((xy, zt), (yt, xz))$$

and notice that they belong to the same configuration. Indeed, in both cases $x'y'$ (resp. $z't'$) is formed by taking one element from $\{x, y\}$ and one element from $\{z, t\}$.

Thus, we have to introduce three sets $\{x, y\}$, $\{z, t\}$ and $D(\xi)$ of which the cardinalities are 2, 2 and $(n-4)$ and that we respectively label by 1, 2 and 3. Consequently, the characterization of the configuration c of $(\xi, \eta) = ((xy, zt), (x'y', z't'))$ doe only depend on the set of labels of which x', y', z' and t' are provided. For example, the configuration concerned by the two above elements of $H \times H$ is, for the η definition $(12, 12)$. Therefore, a given configuration will be specified by an ordered pair of two numbers associated with $(x'y', z't')$. The first (resp. second) number can be $11, 12, 13, 22, 23, 33$. Finally notice that if the same label $(1, 2$ or $3)$ appear more than one time in the definition of the configuration of (ξ, η), the concerned objects are necessarily distinct; precisely because η belongs to H. As above we are going to distinguish five cases according to the number of times where the set labeled 3 intervenes for providing η. Case i is that for which the set 3 intervenes i times, $0 \leq i \leq 4$.

3.3.1. *Structures of η for the case 0*

- $(11, 22)$, $\text{card}(C) = n(n-1)(n-2)(n-3)/4$;
- $(22, 11)$, $\text{card}(C) = n(n-1)(n-2)(n-3)/4$;
- $(12, 12)$, $\text{card}(C) = n(n-1)(n-2)(n-3)$.

3.3.2. *Structures of η for the case 1*

- $(11, 23)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(23, 11)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(12, 13)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- $(13, 12)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- $(12, 23)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- $(23, 12)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)$;
- $(22, 13)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(13, 22)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)/2$.

3.3.3. *Structures of η for the case 2*

- $(11, 33)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;
- $(33, 11)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;
- $(13, 13)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(22, 33)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;
- $(33, 22)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;

- $(23, 23)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(12, 33)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(33, 12)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(13, 23)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)$;
- $(23, 13)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)$.

3.3.4. *Structures of η for the case 3*

- $(13, 33)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$;
- $(33, 13)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$;
- $(23, 33)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$;
- $(33, 23)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$.

3.3.5. *Structures of η for the case 4*

- $(33, 33)$, $\text{card}(C) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)(n-7)/16$.

The number of categories is 26 and one may verify that the sum of the above cardinalities is equal to

$$(n(n-1)(n-3)(n-4)/4)^2 = \text{card}(H \times H).$$

REFERENCES

1. P. Arabie and L.J. Hubert, Combinatorial data analysis. *Annual Rev. Psychology* **43** (1992) 169-203.
2. F.B. Baker, Stability of two hierarchical grouping techniques. *J. Amer. Statist. Assoc.* **69** (1974) 440-445.
3. J.P. Benzecri, *L'Analyse des Données*, Tome 1 : *La Taxinomie*. Dunod, Paris (1973).
4. A. Bravais, Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires de l'Institut de France* (1846) 255-332.
5. H.E. Daniels, The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33** (1944) 129-135.
6. F. Daudé, *Analyse et Justification de la Notion de Ressemblance dans l'Optique de la Classification Hiérarchique par AVL*. Thèse de l'Université de Rennes I (1992).
7. E.B. Fowlkes and C.L. Mallows, A method for comparing two hierarchical clusterings. *J. Amer. Statist. Assoc.* **78** (1983) 553-584.
8. O. Frank and K. Svensson, On probability distributions of single-linkage dendrograms. *J. Statist. Comput. Simulation* **12** (1981) 121-131.
9. L.A. Goodman and W.H. Kruskal, Measures of association for cross classification. *J. Amer. Statist. Assoc.* **49** (1954) 732-764.
10. A. Guénoche and B. Monjardet, Méthodes ordinales et combinatoires en analyse des données. *Revue Mathématiques et Sciences Humaines* **25** (1987) 5-47.
11. A. Guénoche, Ordinal properties of tree distances. *Discrete Math.* **191** (in press).
12. J. Hájek and Z. Sidak, *Theory of Rank Tests*. Academic Press, New-York and London (1967).

13. V. HAMANN, Merkmalbestand und verwandtschaft sbeziehungen der farinosae. Ein Beitragzum System der Monokotyledonen. *Willdenowia* **2** (1961) 639-768.
14. L.J. HUBERT, Inference procedures for the evaluation and comparison of proximity matrices, *Numerical Taxonomy*, edited by J. Felsenstein. NATO ASI Series, Berlin, Springer-Verlag (1983) 209-228.
15. L.J. HUBERT, *Assignment Methods in Combinatorial Data Analysis.* Marcel Decker, New-York (1987).
16. A. JOVICIC, Minimal entropy algorithm for solving node problems, IFCS-96, *Data Science Classification and Related Methods.* Abstracts Vol. 2 (1996) 115-116.
17. M.G. KENDALL, *Rank Correlation Methods.* Charles Griffin, Fourth Edition (1965).
18. F.J. LAPOINTE and P. LEGENDRE, Comparison tests for dendrograms: A comparative evaluation. *J. Classification* **12** (1995) 265-282.
19. F.J. LAPOINTE and P. LEGENDRE, A statistical framework to test the congruence of two nested classifications. *Systematic Zoology* **39** (1990) 1-13.
20. G. LECALVÉ, Un indice de similarité pour des variables de types quelconques. *Statist. Anal. Données* **01-02** (1976) 39-47.
21. I.C. LERMAN, *Les Bases de la Classification Automatique.* Gauthier-Villars, Collection Programmation, Paris (1970).
22. I.C. LERMAN, Formal analysis of a general notion of proximity between variables, Congrès Européen des Statisticiens, Grenoble 1976, *Recent Developments in Statistics.* North Holland (1977) 787-795.
23. I.C. LERMAN, *Classification et Analyse Ordinale des Données.* Dunod, Paris (1981).
24. I.C. LERMAN, Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées, *Publications de l'Institut de Statistique de l'Université de Paris*, XXIX, Fasc. 3-4 (1984) 27-57.
25. I.C. LERMAN, Maximisation de l'association entre deux variables qualitatives ordinales. *Revue Mathématiques et Sciences Humaines* **100** (1987) 49-56.
26. I.C. LERMAN, Formules de réactualisation en cas d'agrégations multiples. *RAIRO Oper. Res.* **23** (1989) 151-163.
27. I.C. LERMAN, Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles, I et II. *Revue Mathématiques Informatique et Sciences Humaines:* I **118** (1992) 35-522; II **119** (1992) 75-100.
28. I.C. LERMAN, Likelihood linkage analysis (LLA) classification method (Around an example treated by hand). Elsevier Editions. *Biochimie* **75** (1993) 379-397.
29. I.C. LERMAN, Comparing Classification tree Structures: A Special Case of Comparing q-Ary Relations, *Publication interne* 1078 IRISA (April 1997) and *Rapport de recherche* 3167 INRIA (Mai 1997); 37 pages.
30. I.C. LERMAN and N. GHAZZALI, What do we retain from a classification tree? An experiment in image coding, *Symbolic-Numeric Data Analysis and Learning*, edited by E. Diday and Y. Lechevallier. Nova Science Publishers (1991) 27-42.
31. I.C. LERMAN and Ph. PETER, Structure maximale pour la somme des carrés d'une contingence aux marges fixées ; une solution algorithmique programmée. *RAIRO Oper. Res.* **22** (1988) 83-136.
32. N. MANTEL, Detection of disease clustering and a generalized regression approach. *Cancer Research* **2** (1967) 209-220.
33. F. MARCOTORCHINO and P. MICHAUD, *Optimisation en Analyse Ordinale des Données.* Masson, Paris (1979).
34. H. MESSATFA, *Unification Relationnelle des Critères et Structures Optimales des Tables de Contingence.* Thèse de doctorat de l'Université de Paris 6 (1990).

35. H. MESSATFA, An algorithm to maximize the agreement between partitions. *J. Classification* **9** (1992) 5-15.

36. F. MURTAGH, Counting dendrograms: A survey. *Discrete Appl. Math.* **7** (1984) 191-199.

37. A. OCHIAI, Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of Japanese Society of Scientific Fisheries* **22** (1957) 526-530.

38. M. OUALI,-ALLAH, *Analyse en Prordonnances des Données Qualitatives, Applications aux Données Numériques et Symboliques*. Thèse de doctorat de l'Université de Rennes I (1991).

39. K. PEARSON, Notes on the history of correlation. *Biometrika* **13** (1920) 25-45.

40. S. REGNIER, Sur quelques aspects mathématiques des problèmes de la classification automatique. *International Computing Center Bulletin* **4** (1965) 175-191.

41. F. ROUXEL, *Comparaison d'arbres de classification*, rapport de DEA, Informatique et Recherche Opérationnelle. Université Paris VI (1997).

42. C. SPEARMAN, The proof and measurement of association between two things. *Amer. J. Psychology* **15** (1904) 88.

43. C. SPEARMAN, A footrule for measuring correlation. *British J. Psychology* **2** (1906) 89.

44. R.R. SOKAL and F.J. ROHLF, The comparison of dendograms by objective methods. *Taxon* **11** (1962) 33-40.

45. G.U. YULE, On the methods of measuring the association between two attributes. *J. Roy. Statist. Soc.* **75** (1912) 579-652.