Compatibility and consensus in numerical taxonomy

Edwin Diday

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tel (1) 39 63 55 11

Rapports de Recherche

N° 917

COMPATIBILITY AND CONSENSUS IN NUMERICAL TAXONOMY

Programme 5

Edwin DIDAY

Octobre 1988

RR_0917

COMPATIBILITY AND CONSENSUS IN NUMERICAL TAXONOMY

COMPATIBILITE ET CONSENSUS EN CLASSIFICATION AUTOMATIQUE

Edwin DIDAY*

Abstract

In numerical taxonomy one of the most important and difficult problems encountered in practice is that of finding a "consensus" between several classification. The problem arrises when we wish to compare the classifications obtained on the same set of objects by using several sets of variables, several dissimilarity indices, several clustering criteria, etc... The notion of comptability between classifications C_i, order θ_i and measure of dissimilarity s_i shed new light in this framework. We introduce the notion of compatible consensus between several triples (C_i, s_i, θ_i). We give several theoretical results simultaneaously available in the case of partitions, hierarchies and pyramids. These results provide construction procedures for obtaining various kinds of compatible consensus which facilitate the visual comparison of several classifications and make the study of a consensus between them easier.

Résumé

La recherche de consensus entre plusieurs classifications est un problème qui se pose souvent dans la pratique. Le problème se pose quand on desire comparer des classifications obtenues sur le même ensemble d'objets en utilisant plusieurs ensembles de variables, plusieurs indices de dissimilarités, plusieurs critères de classification etc... La notion de compatibiblité entre classifications (C_i, ordres θ_i et indices de dissimilarité si apporte un éclairage nouveau dans cette direction de recherche. Nous introduisons la notion de consensus compatible entre plusieurs triples $(C_i s_i, \theta_i)$ et nous donnons des résultats qui s'appliquent aussi bien aux partitions, aux hierarchies qu'aux pyramides. Ces résultats débouchent sur des algorithmes constructif de consensus compatibles qui facilitent la comparaison visuelle de classifications et la recherche de consensus.

* E. DIDAY is Professor at the University of Paris 9 Dauphine and head of department at INRIA Rocquencourt 78150 LE CHESNAY

1 - INTRODUCTION

In recent years many authors have been interested in the "consensus" between several classifications in numerical taxonomy. Adams (1972) has been interested in comparing trees and dendrograms and has given procedures for finding the most similar dendrogram between several given dendrograms Faris (1973) and Mickevich (1978) define rassemblance measures between dendrograms ; Gordon (1980) proposes a pertinent technic for "pruning trees" in terms of comparing two classification schemes ; this approach is extended by Finden and Gordon (1985) to obtain a common pruned tree from two or more dendrograms. In the case of partitions, Celeux (1984) proposes a "consensus" partition which approximate the middle partition which optimizes Condorcet's majority rule. Rohlf (1981) and more recently Barthelemy, Leclerc, Monjardet (1986) give an interesting synthesis on the subject.

Our approach is based on the notion of compatibility between a classification, a dissimilarity index and an order ; in this paper we study a set Ω of n objects ; a classification C of Ω is a set of subsets $h \in C$ of Ω which covers Ω ; a measure of dissimilarity s is a map $\Omega \times \Omega \to \mathbb{R}^+$ such that $\forall\, x,y \in \Omega$, $s(x,y) = s(y,x)$ and $s(x,x) = 0$. An order on the objects of Ω is denoted by θ.

The compatibility between an order with a measure of dissimilarity has been studied by several authors (see for instance Brossier (1981), Diday (1982)), the notion of compatibility between an order with a classification and a classification with a measure of dissimilarity is also introduced ; we have choosen among various possibilities a natural way to define those notions as figures 1a,..., 1f represents it intuitively in a simple example, where $\Omega = \{w_1, w_2, w_3, w_4\}$ are four points of the plane, s is an euclidean distance and θ is the order defined by the sequence w_1, w_2, w_3, w_4.

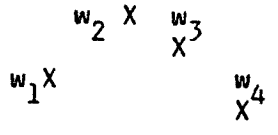


Figure 1.a θ and s are compatible

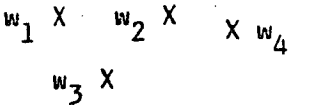


Figure 1.b θ and s are not compatible since $s(w_1,w_3) < s(w_1,w_2)$

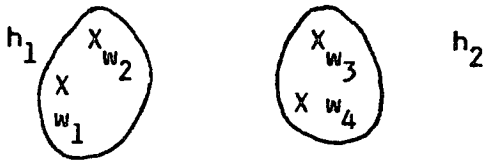


Figure 1c. $C = (h_1, h_2)$ and θ are compatible

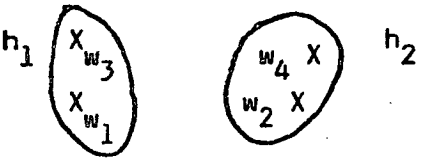


Figure 1d. $C = (h_1, h_2)$ and and θ are not compatible since $[w_1\ w_3]$ is not an interval of θ.

Figure 2

Figure 1e. $C = (h_1, h_2)$ and s are compatible

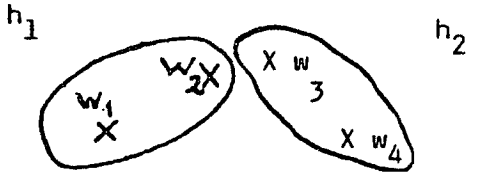


Figure 1f. $C = (h_1, h_2)$ and s are not compatible since $s(w_2, w_3) < s(w_1, w_2)$.

Figure 3

The three precise definitions are given in 2) ; in 3) we recall the notion of induced measure of dissimilarity from a classification (see for instance Diday, Moreau (1984)) and we give properties which relate it to the notion of compatibility. These properties are available in the use of partitions, hierarchies or pyramids ; the pyramids see Diday (1984,85) and Bertrand (1986) are an extension of hierarchies which allow us to represent overlapping clusters by a "pyramidal dendrogram". In 3.3 we study the matrix characterization of the various kinds of compatibilities ; given a measure of dissimilarity it may be shown for instance that the matrix of pairwise dissimilarities between objects ranked in order θ is Robinson if s and θ are compatible.

A general scheme given in table 1 allows us to relate the various kinds of compatibilities and their graphical representation by hierarchical or pyramidal

dendrograms.

Having studied the properties of what we call a compatible triple (C,s,θ) we introduce the notion of compatible consensus between several triples ; we show its existence and we give a construction procedure to obtain it, based on the preceeding theoretical results.

Finally we define the notion of compatible G-consensus between several classifications ; G is any map $\mathscr{C}^L \to \mathscr{C}$ where $\mathscr{C}$ is the set of all the classifications on Ω and $\mathscr{C}^L = \mathscr{C} \times \ldots \times \mathscr{C}$ L times. A compatible G-consensus is a triple $(G(C),s,\theta)$ where $C = (C_1,\ldots,C_L)$. We study in 6 two special cases of compatible G-consensus : the "weak" consensus which leads to a hierarchical representation and the "strong" consensus which leads to a pyramidal representation.

2 - COMPATIBILITY BETWEEN CLASSIFICATIONS, DISSIMILARITY INDICES AND ORDERS

2.1. Compatibility between a classification and a dissimilarity index

Various kinds of "compatibility" between a classification and a dissimilarity index may be considered ; for instance, we may say that a dissimilarity index s and a classification C are compatible if two elements x, y $\in \Omega$ belonging to the same class $h \in C$ are closer to each other than to any element z external to h ; in other words $s(x,y) \leq s(x,z)$ if $x, y \in h \in C$ and $z \notin h$.

To obtain more powerfull results we have finaly decided to use the following definition. Let $h(x,y)$ be (if it exists) a class of smallest cardinality contained in C and which contains x and y. Let $H(x,y)$ be the set of classes of C containing x or y but not x and y simultaneously. We will say that C is "compatible" with s or that (C,s) is a "compatible couple" iff for any x,y $\in I$ for which $h(x,y)$ exist, for any $z \notin H(x,y) \cup h(x,y)$ we have $s(x,y) \leq \text{Min}\{s(x,z), s(y,z)\}$. If there are several classes of minimal cardinality $h(x,y)$ containing x and y is not unique the inequality must be true for all of them. It may happen also that $H(x,y)$ be empty, in that case the inequality must remain true.

2.2 Compatibility between a classification or a dissimilarity index and an order

We say that an order θ and a classification C are "compatible" or that (C,θ) is compatible iff any class $h \in C$ is an interval for θ ; in other words if

two objects are in a class of C, all the objects which are between them with respect to θ and only these are in h.

A dissimilarity index s and an order θ are "compatible" iff for any triple the distance (according to s) between the extreme objects (according to θ) is larger than the distance between the intermediate objects ; in other words, for any triple x y z ordered according to θ', we have $s(x,z) \geq \text{Max}\ (s\ (x,y),\ s(y,z))$. Other kinds of compatibility between an order and a dissimilarity index may be defined ; the two following have been introduced in Diday (1983) : a dissimilarity index s and an order θ are "semi-compatible" iff for any quadruple w_i, w_j, w_{j+1}, w_ℓ ordered with respect to θ and where w_j, w_{j+1} are consecutive we have $s(w_i, w_\ell) \geq s(w_j, w_{j+1})$. We say that s and θ are weakly compatible iff for any ordered triple w_i, w_j, w_k which has two consecutive elements according to θ than the dissimilarity according to s between the consecutive elements is no greater than the dissimilarity between w_i and w_k (for example if k = j+1 than $s(w_j,w_k) \leq s(w_i,w_k)$).

Several links exist among those various kinds of compatibility and have been studied for instance in Diday (1983) ; some of them are given in the appendix.

3 - COMPATIBILITY BETWEEN AN INDEXED CLASSIFICATION AND AN INDUCED DISSIMILARITY INDEX

3.1 Indexed classification and induced dissimilarity index

An indexed classification is a couple (C,f) where C is a classification and f a map $\mathbb{P} \to \mathbb{R}^+$ defined on the set $\mathbb{P}$ of subsets of Ω and taking real "non negative" values ; moreover f satisfies the two following properties : i) f(h) = 0 iff h is a single element of Ω ii) $\forall$ h, h' $\in$ C, h $\subset$ h' strictly ==> f(h) < f(h').

A classical value for f(h) is for instance the within sum of squared deviation.

We define the induced dissimilarity index from an indexed classification by defining the map F such that F(C,f) = s with

$s(x,y) = \text{Min}\ \{f(h) \mid h \in C,\ (x,y) \in h \times h\}$.

If $h \in C$ does not exist in C such that $x, y \in h$ we consider that $s(x,y) = f(\Omega)$. In the remaining part of the paper all the indexed classification will be indexed by a map noted f so to simplifie the notation (C,f) it will be noted C.

3.2 Compatibility between a classification and its induced dissimilarity index in the case of partitions, hierarchies and pyramids

A partition is a set of subsets of Ω which cover Ω and have empty intersection.

Let P be an indexed partition of Ω and $F(P) = s$; we are in the case where $h \in P : x, y \in h$ may not exist ; thus, in this case we have $s(x,y) = f(\Omega)$; in the other cases $x, y \in h \in P$ implies $s(x,y) = f(h) \leq f(\Omega)$; such a dissimilarity index s is called a "caracteristic index".

A hierarchy H is a set of subsets of Ω which contains the single elements and Ω and which satisfies the following property $\forall\ h, h' \in H \Longrightarrow h \cap h' = \phi$ or $h \subset h'$ or $h' \subset h$.

It may be shown that $s = F(H)$ is an ultrametric distance (see for instance Diday (1985)) which means that for any triple $x, y, z \in \Omega$ we have $s(x,y) \leq \text{Max}\ (s(x,z), s(z,y)))$ (ultrametric inequality).

It is easy to show that if we add to a partition P the partition reduced to the single class Ω and the partition of the single elements we obtain a hierarchy.

A pyramid Π is a set of subsets of Ω which contains also the single elements and Ω and which satisfies the following property : $\forall\ h, h' \in H \Longrightarrow h \cap h' = \phi$ or $h \cap h' \in \Pi$ and for which there exists and order θ such that the elements of any $h \in \Pi$ constitute an interval of θ (which means that Π and θ are compatible).

It may then be shown (see Diday (1984) that the set of pyramids contains the set of hierarchies and that $F(\Pi) = s$ is a pyramidal index which means that it exists an order θ such that s and θ be compatible.

In practice the pyramids constitute a natural extension of hierarchies and allows the representation of overlapping clusters (see figure 8, 9, 10, 11 at the end of the paper). For more details on pyramids see appendix 2.

- Some properties of h (x,y) and H (x,y)

According to the definition of h (x,y) and H (x,y) which have been given in 2.2, in the case of a partition, h(x,y) does not exist if x and y are not in the same class and H(x,y) is always empty. In the case of a hierarchy or a pyramid h(x,y) exists always and H (x,y) may be empty. In the three cases (partition, hierarchy, pyramid) when h (x,y) exists it is unique.

We have now the following result, where C may be a partition, a hierarchy or a pyramid.

Proposition 1

1) (C, F(C)) is compatible
2) There exists Θ such that (C,Θ) be compatible
3) If (C,Θ) is compatible then (F(C),Θ) is compatible.

Proof

1) According to the definition of compatibility which has been given in 2.1 we have to prove that if h (x,y) exist $z \notin H(x,y) \cup h(x,y) \Rightarrow s(x,y) \leq \text{Min}(s(x,z)), s(z,y))$ where F(C) = s. We will prove this result in the case of pyramids (it is thus easy to deduce the proof for the other cases). We have to prove that $z \notin H(x,y) \cup h(x,y)$ implies $s(x,y) \leq \min(s(x,z), s(y,z))$; it is clear that $h(x,z) \cap h(x,y) \neq \phi$ implies $h(x,y) \subset h(x,z)$ if $z \notin H(x,y)$ (which implies that $y \in h(x,z)$). It results that $f(h(x,y)) \leq f(h(x,z))$ which implies that $s(x,y) \leq s(x,z)$. In the same way we have also $s(x,y) \leq s(y,z)$ and then $s(x,y) \leq \text{Min}(s(x,z), s(y,z))$.

2) It is easy to see that there is a compatible order with any partition ; in the case of a hierarchy H we can find a compatible order by starting from the largest class and going on until the single elements are reached in the following way : we give an order on the largest classes of H contained in I, we do the same on the largest classes of H contained in each of these classes and so on, until all the classes contains a single element of I. By construction the final order thus obtained is compatible with H. In the case of a pyramid there exists a compatible order by definition.

3) Let $s = F(X)$ in the case of an indexed partition $P = X$ it is easy to see that if θ is a compatible order with P, for any triple x, y, z of θ ordered according to θ, we have $s(x,z) \geq \text{Max}\ (s(x,y), s(y,z))$ because if x, y, z are in the same class the inequality is an equality ; if only x, y or y, z are in the same class we have $s(x,z) = f(\Omega)$ and the inequality remains true.

If X is an indexed hierarchy H the same inequality may be proved as follows ; let θ be a compatible order with H (from 2) we know that it exists) then the smallest class $h(x,z)$ G H which contains x and z contains y if x, y, z are ordered according to θ, hence, we have necessarily $h(x,y) \subset h((x,z)$ and $h(y,z) \subset h(x,z)$ therefore $s(x,z) \geq \text{Max}\ (s(x,y), s(y,z))$.

If X is a pyramid we know from 2) that it exists a compatible order θ with X ; let x, y, z be an ordered triple according to θ ; the smallest class $h(x,z)$ $\in$ X which contains x and z contains necessarily y as $h(x,z)$ must be an interval of θ ; therefore, we have $h(x,y) \subset h(x,z)$ and $h(y,z) \subset h(x,z)$ for the same reason, consequently $f(h(x,y)) \leq f(h(x,z))$ and $f(h(y,z)) \leq f(h(x,z))$ and then $s(x,z) \leq \text{Max}\ (s(x,y), s(y,z))$ which proves that $s = F(X)$ and θ are compatible. The same kind of proof also give the result in the case of a hierarchy.

□

3.3 Matrix characterisation and graphical representation of the various kind of compatibility

Let $M(s,\theta) = [s(w_i,w_j)]$ be the n×n matrix whose elements are the dissimilarity index $s(w_i,w_j)$ values and where rows and columns are ordered according to θ.

In the case of a partition P of k classes, it is easy to see that if θ is compatible with $s = F(P)$ a "characteristic index" $M(s,\theta)$ is represented by a sequence of k rectangles (see figure 2) for which a diagonal is the main diagonal of $M(s,\theta)$. Each rectangle is associated to a class $h_i \in P$ and its elements take the value $f(h_i) = a_i$ except on the main diagonal where the values are 0. The other values are $b_i = f(\Omega)$. We call this kind of matrix a "partition matrix".

In the case of a hierarchy H, if θ is compatible with $s = F(H)$, $M(s,\theta)$ is what we call an "ultrametric matrix" and has been characterized by several

authors (see for instance Lerman (1981)). In such matrixes the values of the elements increase from the main diagonal on each row and column which means that M(s,θ) is Robinson ; moreover these values must satisfy the ultrametric inequality : for any triple x, y, z ∈ Ω, s(x,y) ≤ Max (s(x,z), s(z,y)).

In the case of a pyramid Π if $\hat{\theta}$ is compatible with s = F(Π) it may be shown (see Diday (1984)) that M(s,θ) is Robinson where s is called a "pyramidal index".

We say that a matrix M(s,θ) is SDR iff the elements within any above-rectangle containing an element $s(w_i,w_{i+1})$ of the above diagonal (i.e. the diagonal immediatly above the main diagonal), are not less than $s(w_i,w_{i+1})$. We say that a matrix M(s,θ) is SDD iff in the above triangular matrix the elements of any row (resp. column) are not lower than the element of the above diagonal which is lying on that row (resp. column).

w_1	0	1	1	3	3
w_2	1	0	1	3	3
w_3	1	1	0	3	3
w_4	3	3	3	0	2
w_5	3	3	3	2	0

$h_1 = \{w_1, w_2, w_3\}$
$h_2 = \{w_4, w_5\}$
$a_1 = 1 \quad a_2 = 2 \quad b = 3$

Partition-Matrix

w_1	0	1	3	3	4
w_2	1	0	3	3	4
w_3	3	3	0	4	4
w_4	3	3	4	0	4
w_5	4	4	4	4	0

Ultrametric-Matrix

0	1	3	5	6	
1	0	2	4	6	
3	2	0	4	5	
5	4	4	0	1	
6	6	5	1	0	

Robinson

0	1	5	6	2	
1	0	2	6	5	above
5	2	0	4	7	rectangles
6	6	4	0	1	
2	5	7	1	0	

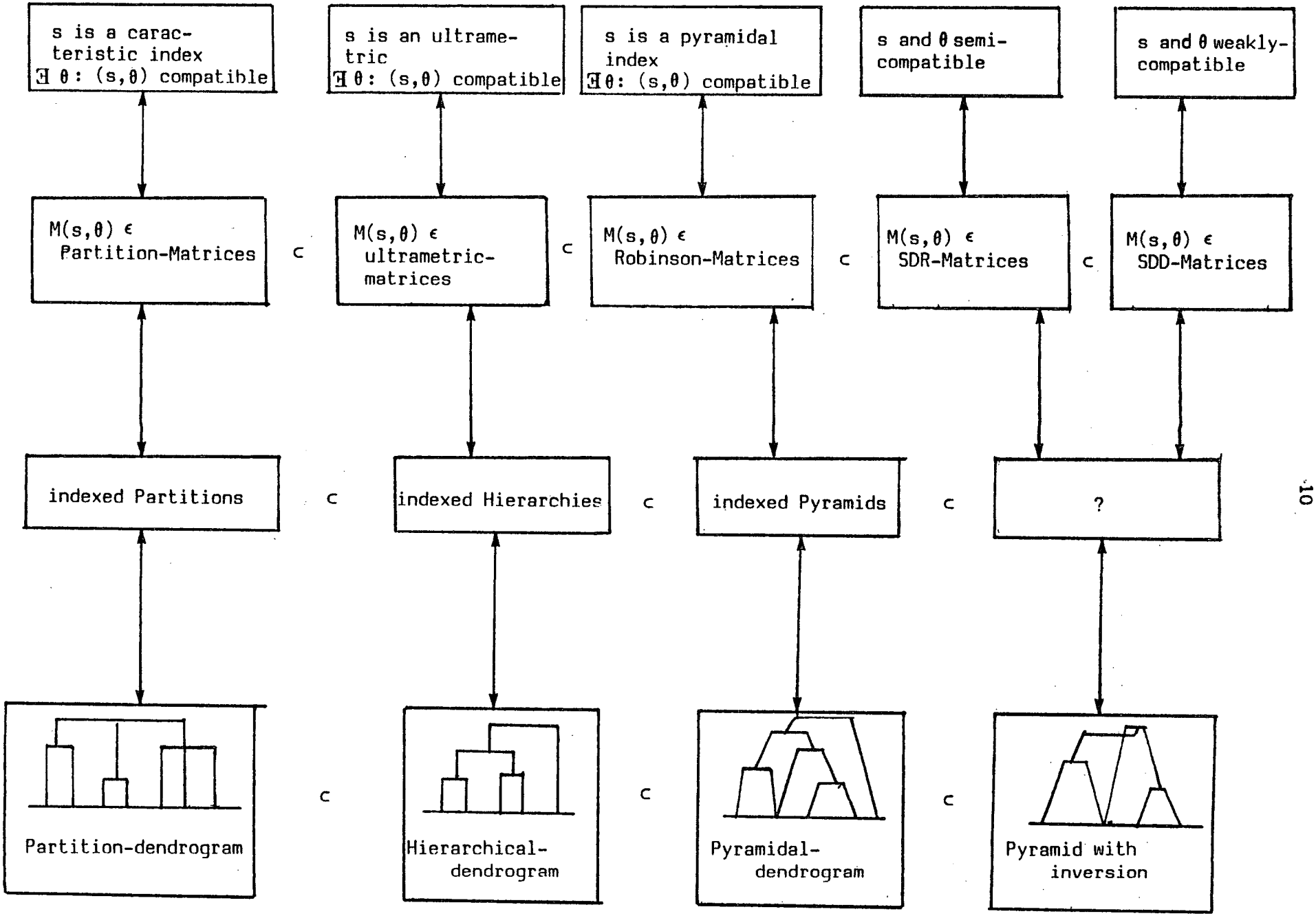
SDR

0	2	5	4	
2	0	4	6	
5	4	0	1	above diagonal
4	6	1	0	

SDD

Figure 2. The matrix characterisation of the various kinds of compatibility

It is easy to show that the set of partition-matrix is included in the set of ultrametric matrices which is included in the set of Robinson matrices. Table 1, permit to sumarize the results of this section and give an idea on the different kinds of graphical representation.



A → B : A implies B, A ↔ B A implies B and B implies A, [X] c [y] : the X set is included in the y set.

4 - DEFINITION AND CONSTRUCTION OF A COMPATIBLE TRIPLE (C,s,θ)

4.1 Définition and properties

A compatible triple (C,s,θ) is a classification C a dissimilarity index s and an order θ which are two by two compatible ; we have then the following general result where X may be an indexed partition a hierarchy or pyramid.

Proposition 2

If (X,θ) or (F(X),θ) is compatible then (X,F(X),θ) is compatible.

Proof :

In Proposition 1 we have proved that (X,F(X)) is compatible if X is a partition a hierarchy or a pyramid and also that if (X,θ) is compatible then (F(X),θ) is compatible ; therefore it remains to be prove that if (F(X),θ) is compatible then (X,θ) is compatible.

Let s = F(X) and (s,θ) be compatible ; let h ∈ X we have to prove that h is an interval of θ ; let w_1, w_2 be the extrem elements of h according to θ for any w ∈ $[w_1,w_2]$. We have $s(w,w_1) \leq s(w_1,w_2)$ since s is compatible with θ therefore we must have w ∈ h because if w ∉ h the smallest class which contains w and h would have to contain h strictly and we would have $s(w,w_1) > s(w_1,w_2)$; this inequality is true for any element w which is out of h therefore w cannot belong to $[w_1,w_2]$.

□

4.2 Construction of a compatible triple : the A algorithm

Starting from any kind of data array (quantitative, qualitative, ordinal...), there are numerous algorithms which make it possible to obtain partitions, hierarchies (see for instance Anderberg (1973), or Diday et al (1985)) or pyramids (see Diday (1984) and Diday, Bertrand (1985)). Having an indexed classification X we have explained in 3.1 how to obtain a compatible dissimilarity index s = F(X) ; having a dissimilarity index s the following algorithm called A allows us to obtain a compatible order θ with s if such an

order exists.

The A algorithm : we start from a randomly choosen element $w \in \Omega$, if at a step of the algorithm we have the partial order $w_i \ldots w_j$, at the next step, we add to the left of w_i or the right of w_j the element w_k depending on if it is the closest element (according to s) of w_i or of w_j among the elements of Ω which have not been already taken ; if w_k is not unique we choose it randomly from the elements which satisfy $s(w_k,w_i) \geq s(w_i,w_j)$ if it is to the right of w_j and $s(w_k,w_j) \geq s(w_i,w_j)$ if it is to the left of w_i. If a compatible order θ with s exist at least one element must satisfy one of these inequalities ; if at one step it does not exist, it is because one of the random choices was wrong at one preceeding step and we have to change it and apply the algorithm by starting again from this step.

If s is an ultrametric the A algorithm becomes easier : we start from a randomly choosen element w_i if at one step we obtain the sequence $w_i \ldots w_j$ the next step will give $w_i \ldots w_j\ w_k$ where w_k is the closest element of w_j among the remaining elements of Ω.

It is easy to show that in both cases the obtained sequence gives an order θ compatible with s.

5 - COMPATIBILITY AND CONSENSUS

5.1 Définition and construction of a compatible consensus

We say that a triple (C,s,θ) is a consensus of L triples (C_i, s_i, θ_i) where $i = 1,2,\ldots,L$ iff $\forall i$, C is compatible with s_i and θ_i, s is compatible with C_i and θ_i and θ is compatible with C_i and s_i. If moreover (C,s,θ) is a compatible triple it is a compatible consensus.

To obtain such a consensus we use L times the algorithm A ; more precisely the L×A algorithm may be defined as follows : we start from an element $w \in \Omega$ choosen randomly, if at a step of the algorithm we have obtained the sequence $w_i \ldots w_j$ at the next step we add to the left of w_i or to the write of w_j the element w_k depending whether if it is the closest element of w_i or of w_j among

the element of Ω which are in $A_i^1 \cap \ldots \cap A_i^L$ or $A_j^1 \cap \ldots \cap A_j^L$ where A_i^ℓ is the subset of Ω containing the closest elements according to s_ℓ of w_i among the elements of Ω which are not yet in the sequence $w_i \ldots w_j$.

Let X_i be a partition, a hierarchy or a pyramid indexed by f_i (see 3.1) and let (X_i, s_i, θ_i) be a compatible triple for $i = 1,\ldots,L$ such that $s_i = F(X_i)$ (where F has been defined in 3.1), we have the following result.

Proposition 3

If for any i it exists a compatible order θ with X_i and s_i then θ is given by the L×A algorithm and if $X = \bigcap_i X_i$ and $f(h) = \text{Max } f_i(h)\ \forall\ h \in X$ (where f_i is the indexing map of X_i) then $(X, F(X), \theta)$ is a compatible consensus.

Proof

We have seen in 4.2 that θ is a compatible order with s_i if it is given by the A algorithm ; therefore, as $\forall i\ A_i^1 \cap \ldots \cap A_i^L$ are not empty if θ exist it results that θ is a compatible order with s_i for $i = 1,\ldots,L$; therefore by applying the proposition 2 we deduce that the triples (X_i, s_i, θ) are compatible and hence that θ is compatible with X_i for $i = 1,\ldots,L$; this implies that any $h \in X_i$ for $i = 1,\ldots,L$ is an interval of θ which proves that θ is compatible with $X_1 \cap \ldots \cap X_L = X$.

It is easy to see that X is necessarily a partition if there is i such that X_i be a partition ; if none of the X_i is a parition and at least one of them is a hierarchy then X is a hierarchy ; if all the X_i are pyramids X is a pyramid ; therefore, in any case we can apply proposition 2 and say that $(X, F(X), \theta)$ is compatible because (X, θ) is compatible if (X, f) is an indexed classification ; it is easy to see that (X, f) is an indexed classification because : i) $\{f_i(h) = 0 \iff h \in \Omega\}$ implies $\{f(h) = \sup \{f_i(h) \mid i = 1,\ldots,L\} = 0 \iff h \in \Omega\}$ and ii) $h, h' \in X$ and $h \subset h' \Rightarrow \forall i\ h, h' \in X_i$ and $h \subset h' \Rightarrow f_i(h) \leq f_i(h') \Rightarrow \sup \{f_i(h) \mid i = 1,\ldots,L\} \leq \sup \{f_i(h') \mid i = 1,\ldots,L\} \Rightarrow f(h) \leq f(h')$.

□

5.2 Other kind of compatible consensus

The compatible consensus which has been obtained by choosing $X = \cap_i X_i$ is not unic, many other choices may be made ; for instance, $X = X_i / C_i$ where $C_i = \{X_1,\ldots X_{i-1}, X_{i+1},\ldots,X_L\}$ represents the "point of view of X_i on C_i" may be defined as follows :

$h \in X \Leftrightarrow \{ \exists h' \in \{X_\ell \mid \ell \neq i, \ell = 1,2,\ldots,L\}$ such that $h' \subset h$ and h is of minimum cardinality in $X_i\}$. In other words $h \in X$ if it is the class of X_i of minimum cardinality which contains at least one class of C_i.

It may be shown as it has been made for proposition 3 that $(X_i \mid C_i, F(X_i/C_i),\Theta)$ is a compatible consensus if Θ is compatible with X_i for $i = 1,\ldots,L$ and X_i/C_i is indexed by f_i the indexing map of X_i.

5.3 The case where there exists only partial compatible consensus.

If there is no order Θ compatible with all the s_i at a step of the algorithm L×A it happen that $A_i^1 \cap \ldots \cap A_i^L = \emptyset$; in this case we apply the algorithm again on the remaining part of Ω by starting from a new element taken in a non empty intersection of a maximum number of A_i^ℓ or A_j^ℓ. This is made until a new empty intersection appears and so on until all the elements have been considered ; by this way we obtain several partial orders which are compatible with the s_i ; as it can be seen on the appendix, each of these partial orders defined a path of a minimum spaning tree of the graph $\Gamma = (\Omega', s_i)$.

5.4 Construction procedure to obtain a compatible consensus

We use table 2 where the notation are defined as follows : A_i^ℓ and the common order Θ are obtained as defined in 5.1 by using the L×A algorithm. $h_{\ell\ \ell+1}$ is the smallest common class of $X_1\ldots X_\ell$ countaining w_ℓ and $w_{\ell+1}$ if Θ is the order $w_1\ldots w_n$; $h^i_{\ell\ \ell+1}$ is the smallest class of the classification X_i containing the smallest class of C_i which contains w_ℓ and $w_{\ell+1}$. It may be proved (except in the case where X is a pyramid) that :

S_1 S_L	A_1^1 A_1^L		A_i^1 A_i^L	A_{n-1}^1 A_{n-1}^L
w_1	w_2		w_3	w_n
$\bigcap_i X_i$	h_{12}		h_{23}	h_{n-1n}
X_i/C_i	h_{12}^i		h_{23}^i	h_{n-1n}^i
$F(\bigcap_i X_i)$	$f(h_{12})$		$f(h_{23})$	$f(h_{n-1n})$
$F(X_i/C_i)$	$f(h_{12})$		$f(h_{23})$	$f_i(h_{n-1n})$

Table 2. Constructive procedure to obtain a compatible consensus

i) the set of the classes $h_{\ell\ \ell+1}$ for $\ell = 1,\ldots,n-1$ produces the classification $X = \cap_i X_i$;

ii) the set of the classes $h^i_{\ell\ \ell+1}$ for $i = 1,\ldots,L$ and $\ell = 1,\ldots,n-1$ produces the classification $X = X_i/C_i$.

Example : the case of several partitions

Let $\Omega = \{w_1,\ldots,w_6\}$; let X_1, X_2, X_3 be three partitions defined as follows $X_i = \{P_i^1, P_i^2, P_i^3\}$ with :

$P_1^1 = \{w_2,w_3\}$ $P_1^2 = \{w_4,w_5\}$ $P_1^3 = \{w_1,w_6\}$

$P_2^1 = \{w_4,w_5\}$ $P_2^2 = \{w_1,w_6\}$ $P_2^3 = \{w_2,w_3\}$

$P_3^1 = \{w_1,w_6\}$ $P_3^2 = \{w_2,w_3\}$ $P_3^3 = \{w_4,w_5\}$.

The three partitions are indexed in the following way :

$f_1(P_1^1) = 1$ $f_1(P_1^2) = 2$ $f_1(P_1^3) = 1$ $f_1(\Omega) = 3$

$f_2(P_2^1) = 2$ $f_2(P_2^2) = 1$ $f_2(P_2^3) = 3$ $f_2(\Omega) = 4$

$f_3(P_3^1) = 1$ $f_3(P_3^2) = 1$ $f_3(P_3^3) = 2$ $f_3(\Omega) = 5$.

By using the construction procedure defined in table 2 we obtain the table 3.

S_1	w_2	$w_1\ w_4\ w_5\ w_6$	w_6	$w_4\ w_5$	w_3
S_2	w_2	$w_1\ w_4\ w_5\ w_6$	w_6	$w_4\ w_5$	w_3
S_3	w_2	$w_1\ w_4\ w_5\ w_6$	w_6	$w_4\ w_5$	w_3
$\theta : w_3$	w_2	w_1	w_6	w_4	w_3
$\cap X_i$	$w_2\ w_3$	Ω	$w_1\ w_6$	Ω	$w_4\ w_5$
X_2/C_2	$w_2\ w_3$	Ω	$w_1\ w_6$	Ω	$w_4\ w_5$
$F(\cap X_i)$	3	5	1	5	2
$F(X_2/C_2)$	3	4	1	4	2

Table 3

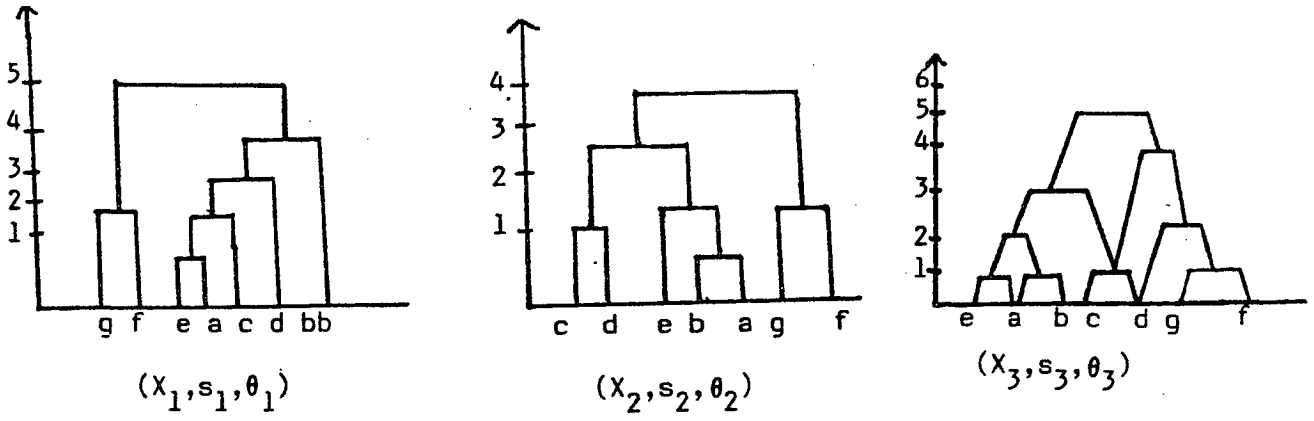


Figure 3

s_1	e a c	e a	a	b	g f	f
s_2	c	e b a	b a	b	g f	f
s_3	c	e a b	a b	b	g f	f
θ : d	c	e	a	b	g	f
$\cap X_i$	dceab	dceab	dceab	dceab	g	g f
X_1/C_1	deca	dceab	dceab	dceab	g	g f
$F(\cap X_i)$	4	4	4	4	5	2
$F(X_1/C_1)$	3	4	4	4	5	2

Table 4

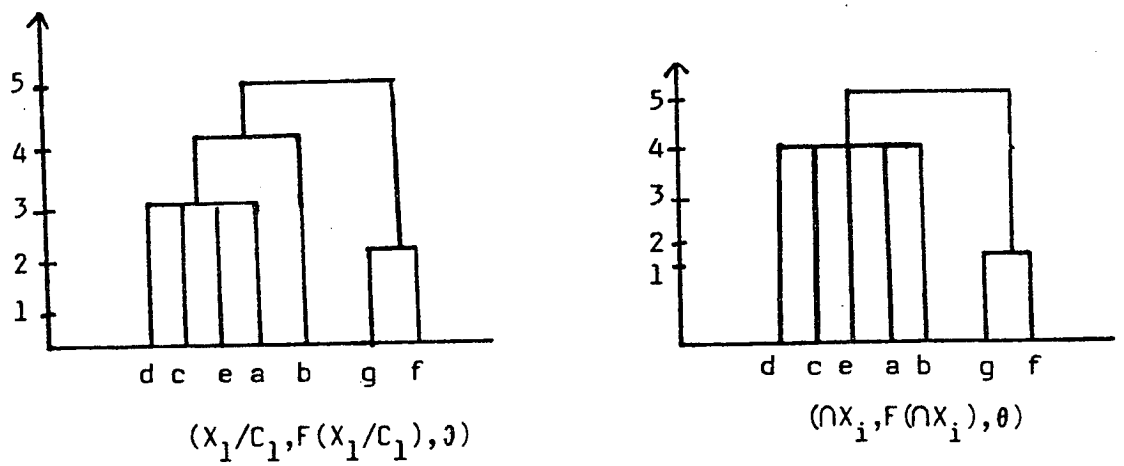


Figure 4

From the table 3 we deduce two compatible consensus : ($\cap X_i$, $F(\cap X_i),\theta$) and (X_2/C_2, $F(X_2/C_2),\theta$) where the order θ is defined by the sequence : w_3, w_2, w_1, w_6, w_4, w_5.

Example : the case of hierarchies and pyramids.

Let Ω = {a, b, c, d, e, f, g} ; let X_1, X_2 be two hierarchies and X_3 a pyramid ; the compatible triple (X_i, s_i, θ_i) where $s_i = F(X_i)$ are given in figure 3 ; table 4 give the construction procedure of the compatible consensus $(\cap_i X_i, F(\cap_i X_i), \theta)$ and $(X_1/C_1, F(X_1/C_1), \theta)$ which are described in figure 4.

6 - COMPATIBLE G-CONSENSUS

6.1 Définition of a compatible G-consensus

Let $\mathscr{C}^L = \mathscr{C} \times \ldots \times \mathscr{C}$ (L time) where $\mathscr{C}$ is the set of all the classifications of Ω.

Let G be a map $\mathscr{C}^L \to \mathscr{C}$ which associates to L classifications $C = (C_1, \ldots, C_L)$ a unic classification $B = G(C)$.

We say that the triple (C, s, θ) is a compatible G-consensus of $(C_\ell, s_\ell, \theta_\ell)$ for $\ell = 1, \ldots, L$ if it exists a dissimilarity index s and an order θ such that (G(C), s, θ) be compatible. It is easy to see that the cases $G(C) = \cap_i C_i$ and $G(C) = X_i/C_i$ studied in 5) permit to define particular cases of G-consensus.

Many other compatible G-consensus may be defined ; let us give two of which we shall consider here : the "weak" and the "strong" G-consensus.

6.2 "Weak" consensus

Let G_i be a map $\mathscr{C}^L \to \mathscr{C}$ such that any class $b \in B = G_i(C)$ where $C = (C_1, \ldots, C_L)$ contains all the elements of Ω for which there exists a path which connect them such that two elements w, w' $\in \Omega$ which are consecutive in this path appear in the same class for at least L-i different classifications taken from $\{C_1, \ldots, C_L\}$.

We have then the following result :

Proposition 4

$G_i(C)$ is a partition and therefore there exists s and θ such that

$(G_i(C),s,\theta)$ be a compatible G_i-consensus of $(C_\ell,s_\ell,\theta_\ell)$ for $\ell = 1,\ldots,L$.

Proof

It is easy to see that $B = G_i(C)$ is a partition because if two classes b, b' $\in$ B have non-empty intersection, they are identical ; therefore, we deduce from proposition 1 that there exists s and θ such that $(G_i(C),s,\theta)$ be compatible.

Let $G(C) = \{\bigcup_i G_i(C)/i = 1,\ldots,L\}$ and $f(b) = L-i$ if $b \in G_i(C)$ then if $(G(C),s,\theta)$ is compatible we say that it is a "weak" consensus of $(C_1,C_2,\ldots,C_L)$.

Let $\Delta(w,w') = L-i$ if w and w' are in the same class for i different classifications taken from $\{C_1,\ldots,C_L\}$.

Proposition 5

There exists s and θ such that the triple $(G(C),s,\theta)$ be a weak consensus of $(C_1,\ldots,C_L)$ moreover G(C) is the single link hierarchy which induces the sub-dominant s of Δ and θ is a compatible order with G(C).

Proof

G(C) is a hiearchy because it contains $G_L \equiv \Omega$ and the singleton $G_0 \equiv \{w_1,\ldots,w_n\}$; moreover h, h' $\in$ G(C) ==> i,j : h $\in G_i(C)$, h' $\in G_j(C)$; suppose for instance i < j then $h \cap h' \neq \emptyset$ ==> $h \subset h'$ because if two consecutive elements of any path contained in h are contained at least in L-i different classifications they are also necessarily in L-j.

If any class h of the hierarchy G(C) is indexed by $f(h) = L-i$ when $h \in G_i(C)$, the induced dissimilarity index $s = F(G(C))$ is the sub-dominant because we have $s(w_i,w_j) = \text{Min } \ell(C_{ij})$ where $\ell(C_{ij})$ is the maximum distance according to Δ between the consecutive elements contained in the path $C_{i,j}$ which connects w_i to w_j in h (this result is given for instance in Diday et al (1985)).

Any graphical representation without crossing of the single link hierarchy H defines an order θ which is compatible with H.

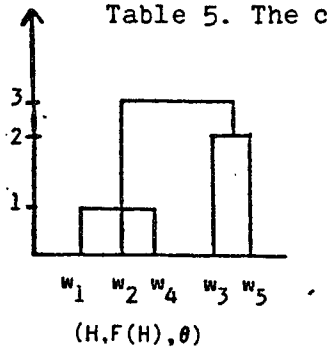
Example

Let $\Omega = \{w_1,\ldots,w_5\}$ and $C = (C_1, C_2, C_3, C_4)$ four classifications defined on Ω by the array of Table 5 where in case (i,j) there is the number of the class of C_j in which w_i appears ; for instance, w_1 belongs in class 2 of the classification C_2.

If H is the single link hiearchy given figure 5 which induces the sub-dominant $s = F(H)$ of Δ also given figure 5 and θ the order defined by the sequence w_1, w_2, w_4, w_3, w_5, then the triple (H,s,θ) is a weak consensus of (C_1, C_2, C_3, C_4).

	C_1	C_2	C_3	C_4
w_1	1	2	3	1
w_2	1	3	3	1
w_3	3	1	1	1
w_4	2	2	3	1
w_5	3	1	2	2

Table 5. The classifications (C_1,C_2,C_3,C_4).



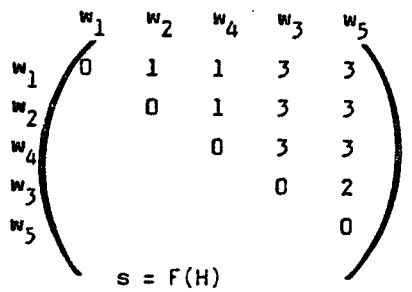


Figure 5. A weak consensus of (C_1,C_2,C_3,C_4)

6.3 "Strong" consensus

Let G_i be a map $\mathscr{C} \xrightarrow{L-i} \mathscr{C}$ such that for any class $b \in B_i = G_i(C)$ two elements w, w' $\in$ b are simultaneously in the same class of at least L-i different classifications taken from $\{C_1,\ldots,C_L\}$.

Such a classification is not necessarily a partition as the following example shows.

Example

Let $\Omega = \{w_1,w_2,w_3\}$ and $C = (C_1,C_2)$ where $C_1 = \{\{w_1\}, \{w_2,w_3\}\}$ and $C_2 = \{\{w_1,w_2\}, \{w_3\}\}$ then $G_1(C) = B_1$ with $B_1 = \{b_1,b_2\}$ where $b_1 = \{w_1,w_2\}$ and $b_2 = \{w_2,w_3\}$.

Let $G(C) = \{\bigcup_i G_i(C) / i = 1,\ldots,L\}$ and $f(b) = L-i$ if $b \in G_i(C)$; we say that $G'(C) \subset G(C)$ if for any $b' \in G'(C)$. There exists $b \in G(C)$ such that $b' \subset b$; finally we say that a triple $(G'(C),s,\theta)$ is a strong consensus of $(C_1,\ldots,C_L)$ iff it is a compatible triple. We will give now an algorithm which makes it possible to build up a pyramid from which a strong consensus may be defined.

The following pyramidal ascending clustering algorithm is called : PAC(d).

1) The elements of Ω are called groups.

2) The two groups of lowest dissimilarity d are merged and their elements are ordered.

3) Step 2 is repeated (until a group becomes identical to Ω) by taking account of the two following contraints.

i) any group may not be the intersection of more than two groups

ii) the "connexity" of any already created group cannot be destroyed ; in other words, it is not possible to merge a group which is internal to a group already created to any external group (we say that a group is internal to h if it does not contain the extreme elements of h according to the order defined in 2)).

Other algorithms which build up pyramids may be given, for more details see. Bertrand (1986). Let $d(b_1,b_2) = L-i$ if i is the greatest value such that b_1 and b_2 belong to $G_i(C)$; it is then easy to obtain the following result.

Proposition 6

A pyramid P given by the PAC(d) algorithm defines a strong consensus $(P,F(d),\theta)$ of $(C_1,\ldots,C_L)$ where θ is a compatible order with P, F is defined by f : $f(b) = d(b_1,b_2)$ if $b, b_1, b_2 \in P$ and $b = b_1 \cup b_2$.

Example

Let $\Omega = \{w_1,w_2,\ldots,w_8\}$ and $C = (C_1, C_2, C_3)$ where C_i for i = 1, 2, 3 is a two classes partition of Ω defined by the table 6 as it has been done in table 4. (for instance this table shows that w_1 appears in the first class of C_1, in the second class of C_2 and in the first class of C_3).

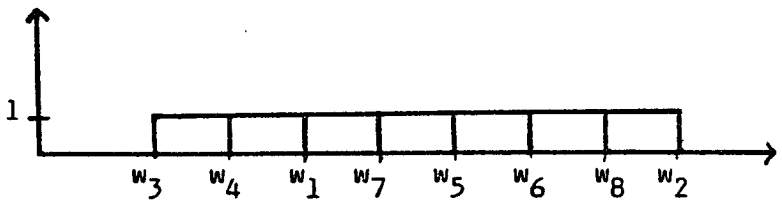


Figure 6a. Weak consensus of (C_1, C_2, C_3)

	C_1	C_2	C_3
w_1	1	2	1
w_2	2	1	2
w_3	1	1	2
w_4	1	1	1
w_5	2	2	2
w_6	2	2	1
w_7	1	2	2
w_8	2	1	1
w_9	2	1	1

Table 6. The classifications (C_1,C_2,C_3)

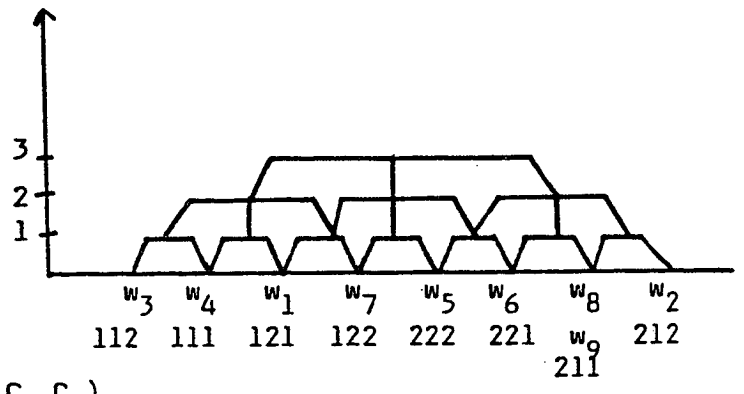


Figure 6b. Strong consensus of (C_1,C_2,C_3)

The pyramid obtained by the PAC(d) algorithm is given figure 6b ; it can be seen that each class obtained is included in G(C) ; all the classes of G(C) are not obtained, for instance $\{w_8, w_9, w_2, w_3, w_4\}$; so, if G'(C) is the pyramid we have $\forall b' \in G'(C) \quad \exists\, b \in G(C)$ such that $b' \subset b$; let θ be the order given in

figure 6a. $\Theta = w_3, w_4, w_1, w_7, w_5, w_6, w_8, w_9, w_2$ and $s(w_i,w_j)$ the height of the lowest level which contains w_i, w_j in this pyramid. Then $(G'(C),s,\Theta)$ is a strong consensus.

The single link hierarchy H built up with the distance Δ given in induces a triple $(H,F(H),\Theta)$ which is a weak consensus ; the result of this example shows that the weak consensus is not interesting when the proportion of stable classes (as $\{w_8,w_9\}$) is too large.

Conclusion

In this paper we have only considered the case of a set of elements $w \in \Omega$ on which several classifications have been made, it is also possible to consider many other cases, for instance the two following : i) several sets of elements $\Omega_1,\ldots,\Omega_n$ are given ; ii) the particular case where $\Omega_1,\ldots,\Omega_n$ defines the elements of the various dimensions of a multi-way data array. In the first case the problem may be for instance to find a unique compatible triple (C,s,θ) on $\bigcup_i \Omega_i$; in the second case the problem may be to find a "cross-compatibility". To keep this paper not too long we don't develop these aspects here, but we give in figure 8, 9, 10 two examples of how in each of these two cases, these problems may be approached by using pyramids (for more detail see Bertrand (1986)).

Many other directions of research are open : the study of compatible consensus in the case of other kinds of compatibility between classification, dissimilarities and orders. If we say that a crossing is a triple of elements belonging in Ω for which there is no compatibility between s and Θ, C and Θ or C and s, how can we find the various kinds of compatible consensus which minimize the number of crossings ? What are the properties of other kinds of compatible -G consensus ? etc...

REFERENCES

ANDERBERG M.R. (1973). Cluster analysis for applications. Academic Press, New-York.

ADAMS E.N. (1972). Consensus Techniques and the comparison of taxonomic trees. Syst. Zool, 21 : 390-397.

BARTHELEMY J.P., LECLERC B., MONJARDET B. On the use of ordered sets in problems of comparing and consensus on classification. Submitted to the Journal of Classification.

BERTRAND P. (1986). Etude de la représentation pyramidale. Thèse de 3ème Cycle. Université Paris Dauphine et INRIA Rocquencourt 78150.

BROSSIER G. (1980). Représentation ordonnée des classifications hiérarchiques. Statistiques et Analyse des Données Vol. 2.

CELEUX G. (1984). Approximation rapide et interprétation d'une partition centrale pour les algorithmes de partitionnement. Rapport INRIA n°301.

DIDAY E., MOREAU J.V. (1984). Learning hierarchical clustering from examples. Rapport de recherche n°289 INRIA Rocquencourt (78150).

DIDAY E. (1982). Croisements ordres et ultramétriques : application à la recherche de consensus. Rapport de recherche n°144. INRIA Rocquencourt 78150 FRANCE.

DIDAY E. (1983). Croisements ordres et ultramétriques. Math. Sc. Hum (21ème année, n°83, 31-54).

DIDAY E. (1985). "Orders and overlapping clusters by pyramids" in J. de Leeuw ed. Multidimensional Data Analysis. Proceedings of a Workshop, Pembroke college, Cambridge University. DSWOO Press/university of Leiden. The Netherlands.

DIDAY E., J. LEMAIRE, J. POUGET, F. TESTU (1985). Elements d'Analyse des données. Dunod.

FARRIS J.G. (1973). On comparing the shape of taxonomic trees. Syst. Zool. 22, 50-54.

FINDEN C.R., GORDON A.D. (1985). Obtaining Common Pruned Trees. Journal of Classification 2 : 255-276.

GORDON A.D. (1980). On the assessment and comparison of classifications. In R. TOMASSONE (ed.). Analyse des données et Informatique. INRIA Rocquencourt 78150 FRANCE.

LERMAN I.C. (1981). Classification et analyse ordinale des données. Dunod.

MICKEVICH M.F. (1978). Taxonomic congruence, Ph. D. Dissertation state University of New York and Stony Brook.

Rohlf F.J. (1981). Consensus indices for comparing classifications. IBM Research Report R.C. 8940.

APPENDIX 1

Let θ' be an order on a subset $\Omega' \subset \Omega$ and $\Gamma = (\Omega', s)$ a graph whose vertices are the elements of Ω' and the edges (w_i, w_j) have the weight $s(w_i, w_j)$.

We denote by $P(A, \theta')$ a path of Γ such that the weight of an edge w_i, w_{i+1} is $s(w_i, w_{i+1})$; A is a minimum spaning tree on Γ ; we say that there is a crossing for a couple (C, θ) if there is a class $h \in C$ such that h is not an interval for θ ; for instance if C is a hierarchy we give in figure 7 an example of hierarchy with or without any crossing.

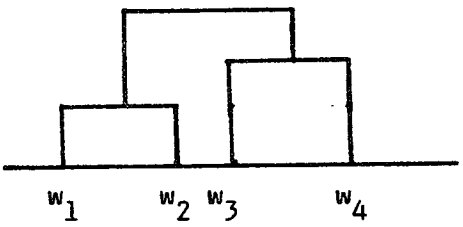


Figure 7a. No crossing

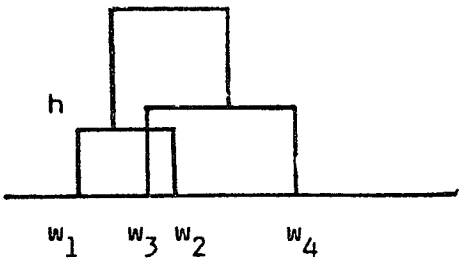


Figure 7b. Crossing

In 7b, h is not an interval for θ : w_1 w_2 w_3 w_4 therefore there is a crossing. The table 7 summarize some of the properties of the different kind of compatibilities.

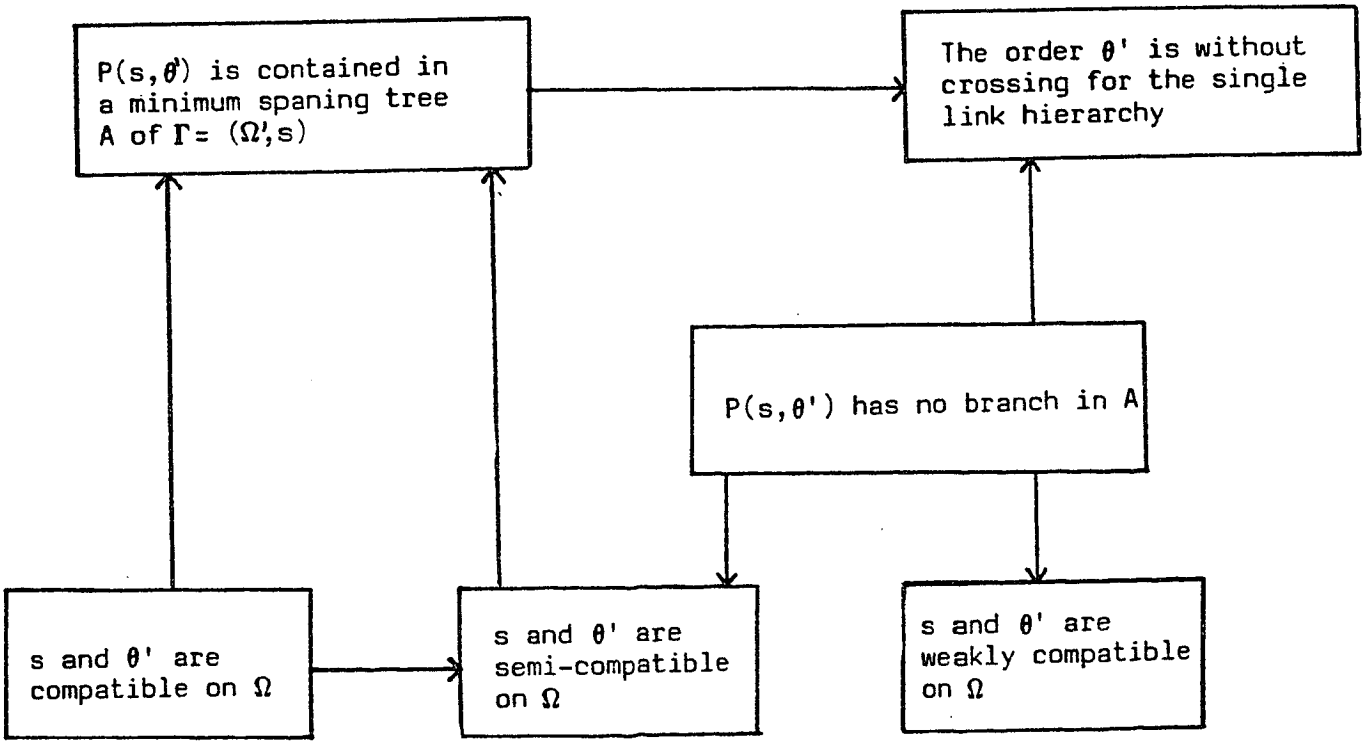


Table 7. $P(s, \theta)$ is a path of the graph $\Gamma(\Omega', s)$.

APPENDIX 2

Indexed hierarchies (H,f) and indexed pyramids (P,f) are special cases of indexed classification. It may be shown that the set of indexed hierarchies is in bijection with the set of ultrametrics and that there also exists a bijection between the set of pyramids and the set of pyramidal indices. Let $M(d,\mathcal{O}) = [d(w_\ell,w_j)]$ be the nxn matrix whose terms are the dissimilarity index $d(w_\ell,w_j)$ values, and where rows and columns are ordered according to $\mathcal{O}$. It may be shown that for any ultrametric δ there exists an order $\mathcal{O}$ such that δ and $\mathcal{O}$ be compatible. We say that a matrix is Robinson iff the terms of the rows and columns never decrease when moving away, in either direction, from the main diagonal. It may be shown that $M(\delta,\mathcal{O})$ is Robinson if $\mathcal{O}$ is compatible with δ and that there exists a bijection between the Robinson matrices and the pyramidal indices. All these results are summarized in table 8. In figure 8 we give an example of cut off pyramid. For more details on pyramids, see Diday (1984), Diday, Bertrand (1986) and Durand, Fichet (1987).

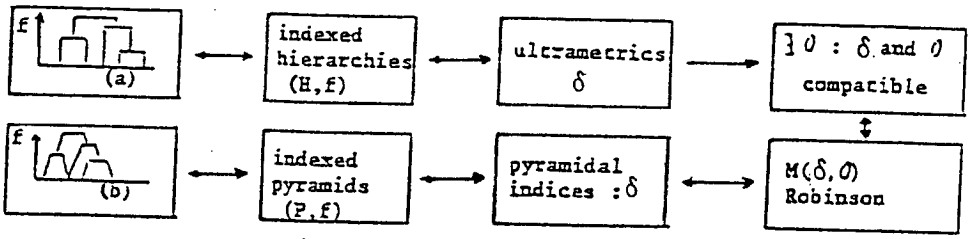


Table 8 : (a) is an hierarchical dendrogram (H,f);
(b) is a pyramidal dendrogram (P,f);
A⟷B means "bijection" between A and B
A⟹B means : A implies B .

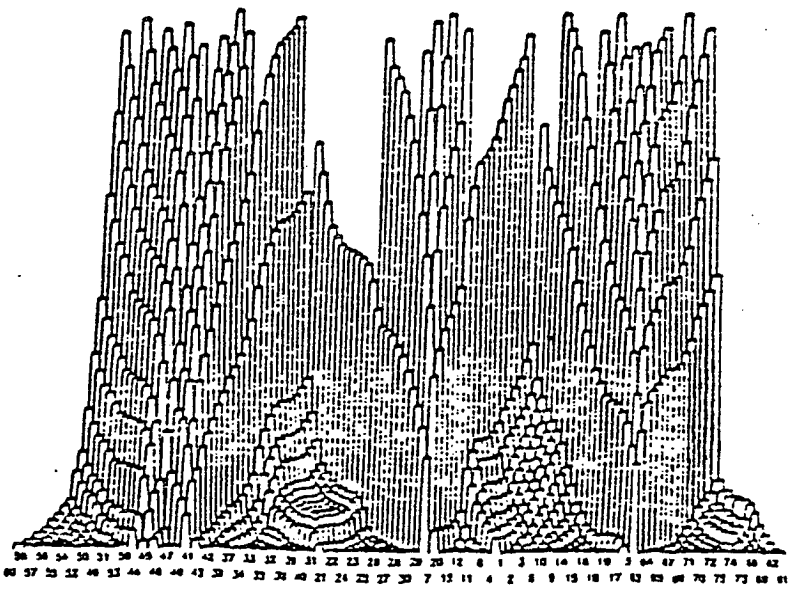
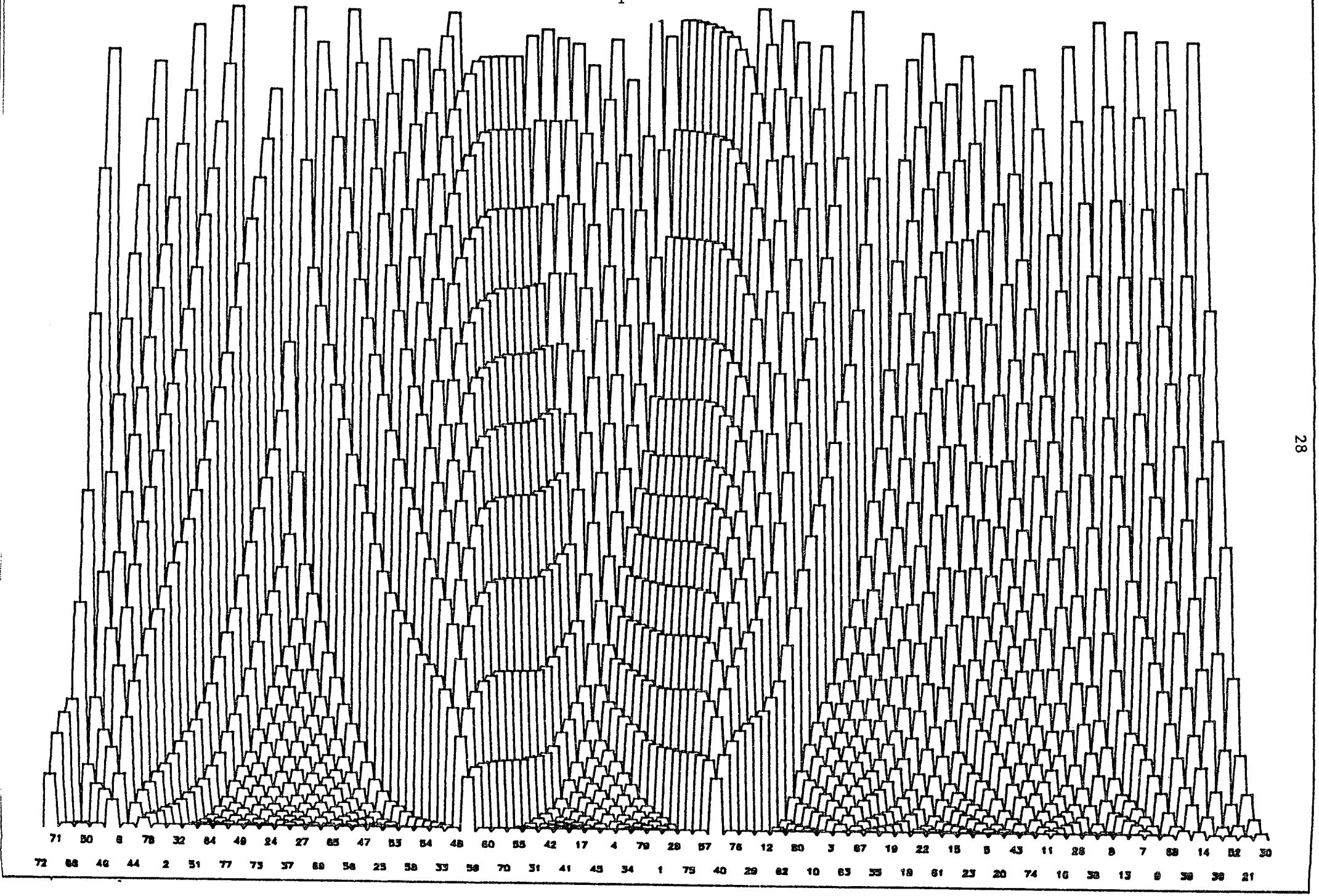


Figure 8 : An example of cut off pyramid

Figure 8. (C,S,B) on Π_1 Π_2 were Π_1 is distributed according to a Gaussian law.



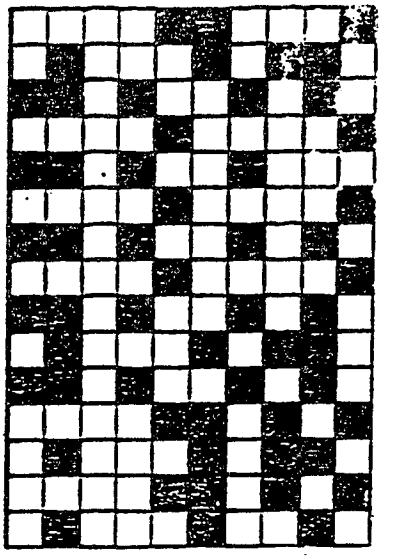


Figure 9 - Initial array

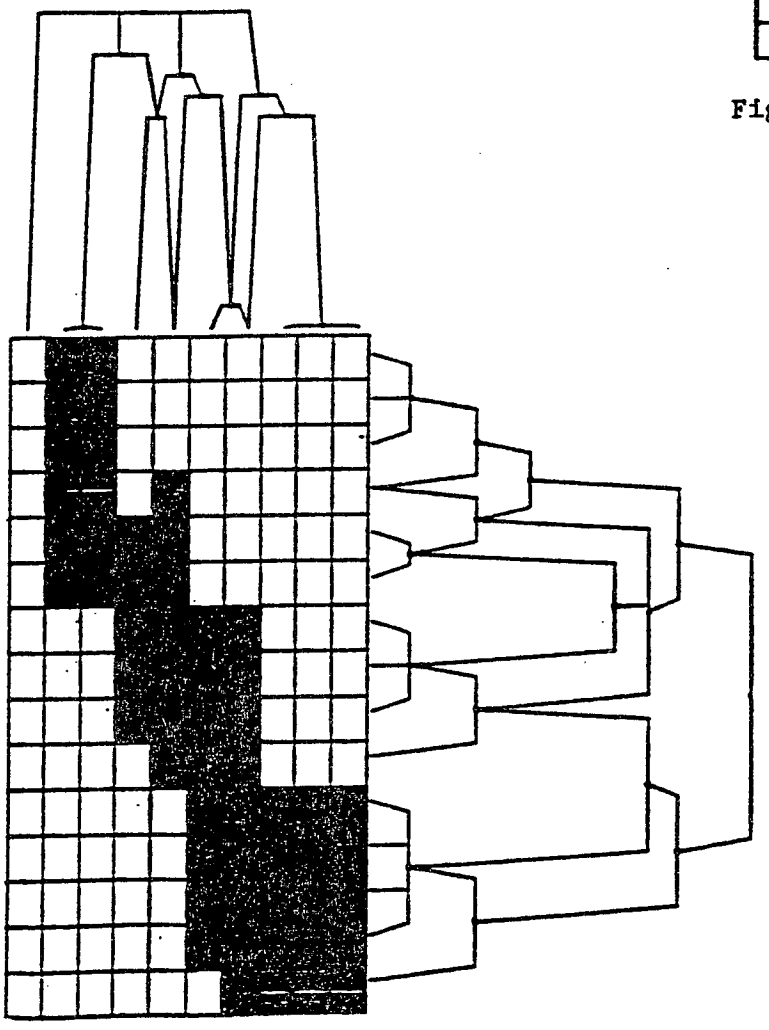


Figure 10. "Cross-compatibility" (Π_1, s_1, θ_1) and (Π_2, s_2, θ_2) on the initial data given figure 9.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique