MATHÉMATIQUES ET SCIENCES HUMAINES

EDWIN DIDAY

Croisements, ordres et ultramétriques

Mathématiques et sciences humaines, tome 83 (1983), p. 31-54

<http://www.numdam.org/item?id=MSH_1983__83__31_0>

Math. Sci. hum. (21^{e} année, n°83, *1983*, p.31-54)

CROISEMENTS, ORDRES ET ULTRAMETRIQUES *

Edwin DIDAY **

Résumé

La représentation visuelle d'une hiérarchie induit un ordre sur les singletons. Si l'on désire représenter la même hiérarchie en tenant compte de contraintes extérieures (ordre des singletons induit par une autre hiérarchie, une partition, un indice de dissimilarité, par exemple) des croisements peuvent apparaître. Il y a un croisement dans la représentation visuelle d'une hiérarchie quand une branche horizontale (associée à un palier) est coupée par une branche verticale associée à un singleton. Il s'agit d'étudier les liens entre croisements, ordres, indices de dissimilarité et ultramétriques.

On utilise la notion de compatibilité entre un ordre et un indice de dissimilarité ; on introduit les notions de semi-compatibilité et compatibilité faible. On étudie les aspects matriciels qui débouchent sur une généralisation des matrices de Robinson. On fait le lien entre toutes ces notions et les chaînes de longueur minimales au sens de l'indice de dissimilarité choisi. En introduisant la notion d'élément "compatible à gauche" ou "à droite" d'une chaîne, on donne de nouvelles propriétés concernant les chaînes incluses dans un arbre de longueur minimum. Dans le cas où cet indice est une ultramétrique, on obtient des propriétés intéressantes liant l'ordre des singletons correspondant à la visualisation d'une hiérarchie indicée et l'ultramétrique induite par cette hiérarchie.

---

* INRIA, Domaine de Voluceau, BP 105, Rocquencourt 78153 Le Chesnay Cedex .

** Je tiens à remercier B. Leclerc pour ses judicieuses remarques lors de la rédaction définitive de ce texte.

1 - INTRODUCTION

Parmi les problèmes qui se posent dans la pratique de la classification automatique, l'un des plus fréquents et difficiles est celui de la comparaison de différentes structures inter-classes (hiérarchies, partitions, recouvrements, etc.). Ce problème se pose quand on désire comparer une même population d'individus caractérisés par des tableaux de données différents : tableaux évoluant dans le temps, tableaux correspondants à différents paquets de variables, par exemple. Il se pose aussi quand on désire comparer l'effet d'un changement de codage, le choix de différentes mesures de ressemblance, la robustesse de la structure inter-classe choisie pour les données traitées, etc ...

Les retombées pratiques de tous ces problèmes ont attiré l'attention de nombreux chercheurs ; citons Adams (1972) qui s'intéresse principalement à la comparaison d'arbres et de hiérarchies par recherche de "consensus" (chercher, par exemple, la hiérarchie qui "ressemble" le plus à plusieurs hiérarchies), Farris (1973) et Mikevich (1978) qui s'intéressent à la définition de mesures de ressemblance entre hiérarchies, Hubert et Baker (1977) pour les aspects "robustesse"; citons surtout Rohlf (1981) qui fait la synthèse des approches les plus récentes.

La notion de "croisement" donne un éclairage nouveau à toute cette problématique ; introduites d'abord en classification hiérarchique, les propriétés qui la caractérisent sont ensuite étendues au cas d'indices de dissimilarité autres que des ultramétriques ; la classification hiérarchique fournit une structure inter-classe qui est très utilisée car d'interprétation visuelle facile quand la taille de la population n'est pas trop grande ; par contre, la comparaison visuelle de plusieurs hiérarchies n'est pas très aisée surtout si l'ordre des singletons qui se trouvent à la base de chaque hiérarchie n'est pas le même. Si l'on tente de représenter plusieurs hiérarchies avec le même ordre sur les singletons, des "croisements" peuvent apparaître ; on dit qu'il y a un "croisement" dans la représentation visuelle d'une hiérarchie quand il apparaît une coupure entre une branche horizontale et une branche verticale associée à un singleton.

L'étude théorique de la notion de croisement débouche rapidement sur la recherche de liens entre un indice de dissimilarité et un ordre. On introduit la notion de "compatibilité faible" entre un ordre Θ et un indice de dissimilarité d et on montre que cette notion est équivalente, si d est une ultramétrique, à

l'inexistence de croisements pour la hiérarchie induite par d quand Θ est l'ordre des singletons ; on montre aussi qu'elle est équivalente au fait que la chaîne induite par Θ et valuée par d soit de plus courte longueur ; quand d n'est pas une ultramétrique, il faut introduire une nouvelle condition appelée "semi-compatibilité" qui est plus restrictive que celle de "compatibilité faible" mais moins restrictive que cette de "compatibilité" qui a été clairement rappelée par Brossier (1981).

Les aspects matriciels de ces notions permettent de les relier entre elles et de généraliser les matrices de Robinson utilisées, notamment, par Kendall (1969) et Hubert (1974) à des familles de matrices dites SDR et SDD qui les contiennent. La semi-compatibilité donne un point de vue nouveau à la notion de chaîne T-minimax introduite par Leclerc (1977, 1981) et permet donc de faire apparaître une nouvelle caractérisation des chaînes "incluses" dans un arbre de longueur minimum. La notion de "X-compatibilité" permet de regrouper les trois types de compatibilité que nous étudions entre un ordre et une distance ; on l'utilise pour définir les matrices "quasi-Robinson, SDR, SDD" dans le cas où seule une partie $\Omega' \subset \Omega$ est ordonnée. On introduit pour cela, la notion d'élément à gauche ou à droite d'une chaîne. On montre entre autres que si d et Θ' sont semi-compatibles alors la chaîne induite est contenue dans un arbre de longueur minimum ; réciproquement si une chaîne est sans branches dans un arbre de longueur minimum la distance et l'ordre associés à cette chaîne sont semi-compatibles ; il en résulte que la semi-compatibilité caractérise les chaînes qui sont arbres de longueur minimum. Dans le cas où d est une ultramétrique on montre enfin l'équivalence entre la notion de croisement et les différents types de compatibilité.

Nous supposons connu un certain nombre de notions classiques en classification automatique : indice d'agrégation, ultramétriques, existence d'une bijection entre les hiérarchies indicées et les ultramétriques, etc ...) que le lecteur pourra trouver dans le chapitre 2 du livre "Eléments d'analyse des données", Dunod (1982) ; signalons enfin que la notion de croisement correspond à une "inversion horizontale" dans une hiérarchie, le lecteur intéressé par les problèmes "d'inversion verticale" pourra se reporter à [6].

2 - DEFINITION D'UNE HIERARCHIE ET D'UN CROISEMENT

Définition d'une hiérarchie

Soit Ω un ensemble fini, H un ensemble de parties (appelées paliers) de Ω, H est une hiérarchie sur Ω si et seulement si :

a) $\Omega \in H$ et $\phi \notin H$
b) $\forall w \in \Omega$, $\{w\} \in H$
c) $\forall h,h' \in H$ on a $h \cap h' \neq \phi \Longrightarrow h \subset h'$ ou $h' \subset h$.

Soit une hiérarchie H définie sur un ensemble d'individus Ω et un ordre Θ quelconque sur Ω que nous notons pour simplifier $w_1,...,w_n$ (autrement dit, $w_i \ \Theta \ w_j$ si $i < j$).

Définition

On dit que l'ordre Θ donne lieu à un croisement pour la hiérarchie H, si et seulement si il existe $h \in H$ contenant deux éléments de Ω, w_i et w_ℓ tels qu'il existe $w_j \notin h$ avec $i < j < \ell$.

Exemples

L'ensemble des individus est $\Omega = \{ w_1, w_2, w_3, w_4 \}$, $h_1 = \{w_1, w_3\}$ et $h_2 = \{w_2, w_4\}$ dans la hiérarchie H qui est représentée figures 1 et 2.

Dans la figure 1, l'ordre $w_1w_2w_3w_4$ donne lieu à un croisement ; par contre, dans la figure 2, l'ordre $w_1w_3w_2w_4$ ne donne pas lieu à un croisement pour la même hiérarchie H.

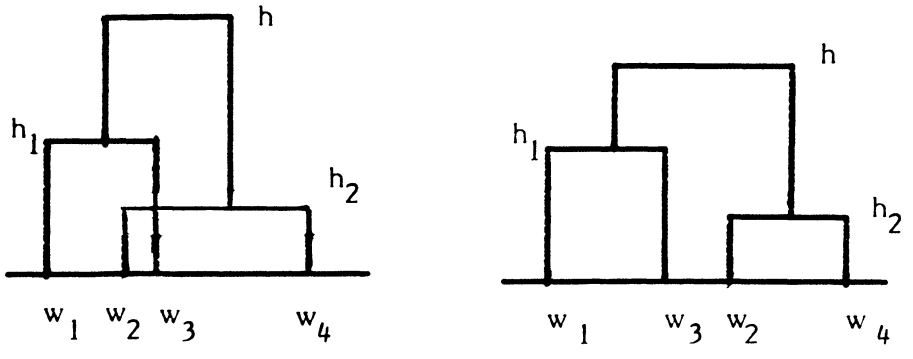


Figure 1

3 - ORDRES SANS CROISEMENTS

3.1. Construction d'un ordre sans croisement

On peut définir un ordre sans croisement à partir d'une hiérarchie H en associant, à chaque palier, un ordre sur les plus grands paliers dont il est la réunion ; en procédant ainsi à partir du palier identique à Ω jusqu'aux paliers qui ne contiennent que des singletons, on obtient un ordre sur les individus qui ne peut comporter de croisement , par construction même. Remarquons que si H est une hiérarchie binaire (autrement dit, chaque palier est formé du regroupement de 2 éléments de H, il y a donc n-1 paliers) alors le nombre d'ordres sans croisement est 2^{n-1} puisque chaque palier peut ordonner de deux façons les deux paliers dont il est la réunion ; c'est beaucoup,mais c'est peu par rapport au nombre de hiérarchies binaires sur Ω qui vaut $n!(n-1)!2^{1-n}$ (voir par exemple Frank et Svensson (1981)).

Exemple

Si l'on reprend la hiérarchie représentée dans la figure 1, l'algorithme indiqué ci-dessus donne la succession d'ordres suivants :

$h \rightarrow h_1\ h_2,\ h_1 \rightarrow w_1\ w_3,\ h_2 \rightarrow w_2\ w_4$

d'où finalement l'ordre $w_1\ w_3\ w_2\ w_4$.

3.2. Une condition nécessaire et suffisante pour avoir un ordre sans croisement: la compatibilité faible

Nous considérons l'ordre Θ sur Ω défini par $w_1\ w_2 \ldots w_n$ et un indice de dissimilarité d.

Définition

Nous dirons que d et Θ sont faiblement compatibles si et seulement si, pour tout triplet $w_i\ w_j\ w_k$ où $i < j < k$ et deux individus sont consécutifs (au sens de Θ), la distance (au sens de d) des deux sommets consécutifs est inférieure[x] à $d(w_i,w_k)$.

---

x Dans tout le texte les notions "d'inférieur" ou de "supérieur" doivent être comprises au sens large.

Autrement dit si $j = i + 1$ par exemple, on doit avoir $d(w_i, w_j) \leq d(w_i, w_k)$.

Rappelons qu'une hiérarchie indicée est un couple (H,f) où H est une hiérarchie et f une application de H dans $\mathbf{R}^+$ telle que :

1° $f(h) = 0 \iff h$ est un singleton.
2° $\forall h, h'$ dans H : $h \subset h'$ on a $f(h) < f(h')$.

On sait qu'il existe une bijection entre l'ensemble des hiérarchies indicées et l'ensemble des ultramétriques (voir, par exemple, E. DIDAY et al (1982)) ; soit δ_H l'ultramétrique associée à une hiérarchie indicée notée H par cette bijection ($\delta_H(w_i,w_j)$ est la hauteur du plus bas des paliers de H contenant w_i et w_j). On a alors la proposition suivante :

Proposition 1

Une c.n.s. pour que l'ordre Θ ne donne pas lieu à un croisement pour la hiérarchie indicée H est que δ_H et Θ soient faiblement compatibles.

Démonstration

Montrons d'abord que s'il n'y a pas de croisement alors δ_H et Θ sont nécessairement faiblement compatibles ; en effet, si δ_H et Θ ne sont pas faiblement compatibles, alors il existe un triplet w_i w_j w_k tel que $\delta_H(w_i,w_j) > \delta_H(w_i,w_k)$ si $j = i+1$ ou $\delta_H(w_j,w_k) > \delta_H(w_i,w_k)$ si $k = j+1$; supposons que $j = i+1$ (la démonstration se fait de façon analogue si $i = j+1$). Soit h_2 (resp. h_3) le plus bas palier de H qui contienne w_i et w_j (resp. w_k) ; si w_j appartenait à h_3, on aurait $h_2 \subset h_3$ ou $h_2 \equiv h_3$ et donc $\delta_H(w_i,w_j) \leq \delta_H(w_i,w_k)$ puisque H est indicée (i.e. pas d'inversions), ce qui est contraire à l'hypothèse ; donc w_j n'appartient pas à h_3 bien que $i < j < k$, il y a donc un croisement.

Réciproquement, si l'ordre Θ est sujet à un croisement pour la hiérarchie H, il existe h_1 et h_2 dans H tels que h_1 contienne w_i, w_k et h_2 contienne $w_j \notin h_1$ avec $i < j < k$; si un tel croisement existe, on peut construire un triplet w_ℓ $w_{\ell+1}$ $w_{\ell'}$, comportant donc deux éléments consécutifs dont la distance (au sens de δ_H) est strictement supérieure à $\delta_H(w_\ell, w_{\ell'})$, (i.e. δ_H et Θ ne

sont pas faiblement compatibles) ; en effet, soit ℓ (resp. ℓ') le premier indice supérieur (resp. inférieur) à i (resp. k) tel que $w_{\ell+1} \notin h_1$ (resp. $w_{\ell'-1} \notin h_1$), voir la figure 2.

On a nécessairement $\ell \neq \ell'$ puisque $w_j \notin h_1$ et $i < j < k$.

Soit h'_1 le plus bas des paliers contenant w_i et w_{i+1} ; on a donc $h_1 \subset h'_1$ puisque $h_1 \cap h'_1 \neq \Phi$ (à cause de w_ℓ) et $w_{\ell+1}$ appartient à h'_1 et non à h_1; il en résulte que $\delta_H(w_\ell, w_{\ell'}) < \delta_H(w_\ell, w_{\ell+1})$ puisque H est une hiérarchie indicée ; c'est bien la condition cherchée. Il en résulte qu'une condition suffisante pour que Θ ne donne pas lieu à un croisement est que δ_H et Θ soient faiblement compatibles. □

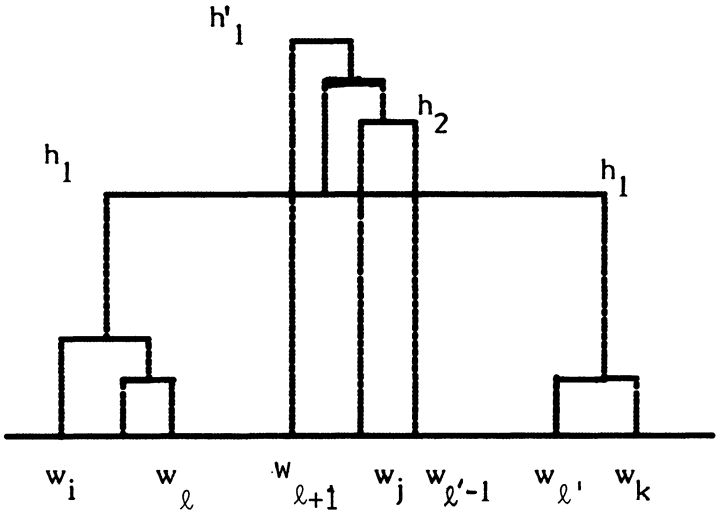


Figure 2

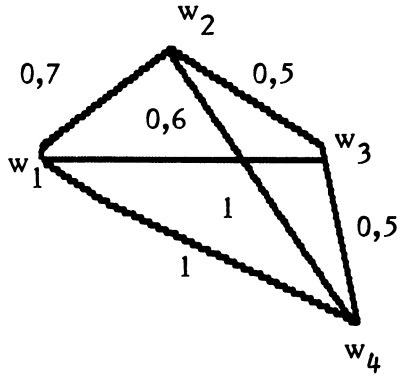
3.3. Chaînes de plus courte longueur et compatibilité faible

Un ordre Θ étant donné, on peut lui associer la chaîne hamiltonienne définie sur Ω et dont les arêtes sont formées de deux sommets consécutifs de l'ordre. Un indice de dissimilarité d étant donné, on associe à chaque arête $w_i\ w_{i+1}$ le poids $d(w_i, w_{i+1})$. Une chaîne ainsi valuée sera notée $C(d,\Theta)$. Une chaîne $C(d,\Theta)$ est de plus courte longueur si la somme des poids des arêtes est minimum.

Le fait que d et Θ soient faiblement compatibles n'est ni nécessaire, ni suffisant pour que la chaîne $C(d,\Theta)$ soit de plus courte longueur. Ceci est prouvé par les deux exemples suivants. Le premier exemple montre que la condition n'est pas nécessaire et le second qu'elle n'est pas suffisante.

Exemple 1

Considérons la matrice de dissimilarité indiquée figure 3 ; la chaîne $C(d,\Theta)$ induite par l'ordre Θ : $w_1\ w_2\ w_3\ w_4$ est de plus courte longueur bien que d et Θ ne soient pas faiblement compatibles puisque : $d(w_1,w_2) \quad d(w_1,w_3)$.



	w_1	w_2	w_3	w_4
w_1	0	0,7	0,6	1
w_2		0	0,5	1
w_3			0	0,5
w_4				0

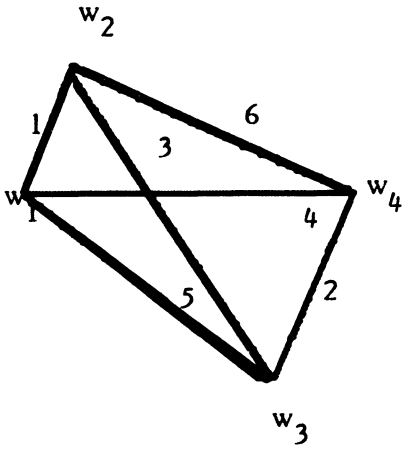
Figure 3

Exemple 2*

Considérons la matrice de dissimilarité indiquée figure 4 ; la chaîne $C(d,\Theta)$ définie par l'ordre Θ : $w_1\ w_2\ w_3\ w_4$ (de longueur égale à 7) est plus longue que la chaîne $C(d,\Theta')$ définie par Θ' : $w_2\ w_1\ w_4\ w_3$ (de longueur égale à 6) bien que d et Θ soient faiblement compatibles :

---

* ce contre exemple m'a été signalé par B. Monjardet que je tiens à remercier ici.



	w_1	w_2	w_3	w_4
w_1	0	1	5	3
w_2		0	4	6
w_3			0	2
w_4				0

Figure 4

4 - AUTRES TYPES DE COMPATIBILITE ENTRE UN ORDRE ET UN INDICE DE DISSIMILARITE

4.1. Définition de la semi-compatibilité

La semi-compatibilité entre un ordre Θ et une distance d est une condition plus restrictive que la compatibilité faible car elle assure (comme nous le verrons en 7) à la chaîne $C(d,\Theta)$ d'être de plus courte longueur.

Définition

Nous dirons que d et Θ sont semi-compatibles si et seulement si tout quadruplet $w_i\ w_j\ w_k\ w_\ell$ avec $i \leq j < k \leq \ell$ et $k = j+1$ est tel que $d(w_j,w_k) \leq d(w_i,w_\ell)$.

Il résulte de cette définition que si d et Θ sont semi-compatibles alors d et Θ sont faiblement compatibles, l'inverse n'étant pas nécessairement vrai.

4.2. Compatibilité entre un ordre et un indice de dissimilarité

Soit d un indice de dissimilarité sur Ω et Θ : $w_1 \ \ldots \ w_n$ un ordre total sur Ω ; on pose $d(w_i,w_j) = d_{ij}$.

Définition

d et Θ sont dits compatibles si et seulement si :

$$\{ i < j < k \} \iff \{ d_{ij} \leq d_{ik} \text{ et } d_{jk} \leq d_{ik} \}.$$

Il résulte facilement de cette définition que si une distance et un ordre sont compatibles, ils sont faiblement compatibles ; en effet, la condition (1) est a fortiori satisfaite si $j = i+1$ ou $k = j+1$. Par contre, si d et Θ sont faiblement compatibles, ils ne sont pas nécessairement compatibles car la condition (1) peut ne pas être satisfaite pour les triplets ne comportant pas de sommets consécutifs. Nous verrons que la semi-compatibilité est une notion intermédiaire entre la compatibilité et la compatibilité faible. Nous verrons également que si d est une ultramétrique, alors les notions de compatibilité, semi-compatibilité et compatibilité faible sont équivalentes.

4.3. Aspects matriciels : matrices de Robinson et matrices SDR et SDD

4.3.1. Matrices de Robinson

Soit D la matrice de dissimilarité associée à d, autrement dit :

$$D = \{ d_{ij} \}_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

La matrice D étant symétrique, on peut définir la matrice de Robinson et les matrices SDR et SDD en ne considérant que la partie triangulaire supérieure de D.

Définition d'une matrice de Robinson (voir, par exemple, Kendall (1969), Hubert (1974)).

Une matrice est dite de Robinson si et seulement si les termes des lignes et des colonnes sont croissants à partir de chaque terme de la diagonale (voir figure 5).

0	2	4	6	0	2	6	5	0	2	5	3
2	0	4	5	2	0	4	7	2	0	4	6
4	4	0	1	6	4	0	1	5	4	0	1
6	5	1	0	5	7	1	0	4	6	1	0
matrice de Robinson				matrice SDR				matrice SDD			

Figure 5

4.3.2. Matrices à sur-diagonales "rectangle"

Considérons la matrice triangulaire supérieure déduite de D ; la diagonale de D étant exclue, la sur-diagonale est la plus grande diagonale de cette matrice. Par exemple, dans la matrice de Robinson, indiquée figure 5, la sur-diagonale est : 2 4 1.

A chaque terme de la sur-diagonale, on peut associer un rectangle dont les côtés sont formés de la ligne et de la colonne contenues dans la matrice triangulaire supérieure et issues de ce terme.

Par exemple, dans la matrice de Robinson, indiquée figure 5, le rectangle (qui est dans ce cas un carré) issu du terme 4 de la sur-diagonale est $\begin{matrix} 4 & 6 \\ 4 & 5 \end{matrix}$.

Définition d'une matrice SDR

Une matrice est dite SDR (sur-diagonale "rectangle") si chaque terme de la sur-diagonale est inférieur aux termes du rectangle qui lui est associé (voir un exemple figure 5).

4.3.3. Matrices à sur-diagonale dominée

Définition d'une matrice SDD

Une matrice est dite SDD (sur-diagonale dominée) si dans la matrice triangulaire supérieure associée à D les termes des lignes et des colonnes sont plus grands que le terme de la sur-diagonale qu'elles contiennent (voir figure 5).

4.3.3. Matrices à sur-diagonale dominée

Définition d'une matrice SDD

Une matrice est dite SDD (sur-diagonale dominée) si dans la matrice triangulaire supérieure associée à D les termes des lignes et des colonnes sont plus grands que le terme de la sur-diagonale qu'elles contiennent (voir figure 5).

On peut résumer ces trois définitions par le tableau 1 (où les intervalles de variation de i, j et ℓ entre 1 et n se déduisent immédiatement des différentes formules ; ainsi par exemple, pour la condition Robinson lignes, on a i=1, ..., n et j=1, ..., n-1).

pour $i \leq j$

Robinson $\Longleftrightarrow$

- lignes $d_{ij} \leq d_{ij+1}$ SDR $\Longleftrightarrow$ $\{d_{j\ j+1} \leq d_{i\ell\ j+1} \leq \ell\}$
- colonnes $d_{ij} \leq d_{i-1j}$ SDD $\Longleftrightarrow$
  - lignes $d_{ii+1} \leq d_{ij+1}$
  - colonnes $d_{j-1j} \leq d_{i-1j}$

Tableau 1

Il résulte facilement de ces trois définitions que l'ensemble des matrices de Robinson est inclus dans l'ensemble des matrices SDR qui est lui même inclus dans l'ensemble des matrices SDD. On a de plus les trois propriétés suivantes :

(1) { M(d, Θ) Robinson } $\Longleftrightarrow$ { d et Θ sont compatibles }
(2) { M(d, Θ) SDR } $\Longleftrightarrow$ { d et Θ sont semi-compatibles }
(3) { M(d, Θ) SDD } $\Longleftrightarrow$ { d et Θ sont faiblement compatibles }

5 - COMPATIBILITE ENTRE UN INDICE DE DISSIMILARITE ET UN ORDRE DEFINI SUR UNE PARTIE $\Omega' \subset \Omega$.

Dans tout ce qui suit on note $\Theta' : w_1 \ldots w_\ell$, $\ell < n$ un ordre total défini sur une partie Ω' de Ω.

5.1. Définition de la notion d'élément compatible à gauche ou à droite

On dit qu'un élément w de Ω est compatible "à gauche" de la chaîne $C(d,\Theta')$ si et seulement si pour $j = 1, \ldots, \ell$ on a :

(1) $d(w,w_j) \geq d(w_i,w_j)$ avec $1 \leq i < j$.

L'élément w est dit compatible "à droite" si et seulement si pour $j = 1, \ldots, \ell$ on a :

(2) $d(w,w_j) \geq d(w_i,w_j)$ avec $j < i \leq \ell$.

Un élément w est dit semi-compatible à gauche (resp. à droite) de $C(d,\Theta')$ si on peut remplacer les inégalités (1) (resp. (2)) par les inégalités (3) (resp. (4)) suivantes :

(3) $d(w,w_j) \geq d(w_i,w_{i+1})$ avec $i = 1, \ldots, j-1$

(4) $d(w,w_j) \geq d(w_i,w_{i+1})$ avec $i = j, \ldots, \ell-1$.

De même en remplaçant l'inégalité (1) (resp. (2)) par l'inégalité (5) (resp. (6)) on obtient les conditions de la compatibilité faible à gauche (resp. à droite) de la chaîne $C(d,\Theta')$, pour un élément w de Ω.

(5) $d(w,w_j) \geq d(w_j,w_{j-1})$

(6) $d(w,w_j) \geq d(w_j,w_{j+1})$

5.2. Définition de la X-compatibilité entre un indice de dissimilarité et un ordre défini sur une partie de Ω.

Afin de simplifier l'énoncé de la définition générale qui suit, le terme "X-compatibilité" est utilisé pour désigner soit la "compatibilité" soit la "semi-compatibilité", soit la "compatibilité faible".

Définition

Un indice de dissimilarité d et un ordre Θ' sur une partie Ω' de Ω sont X-compatibles sur Ω si et seulement si :

a) d et Θ' sont X-compatibles
b) tous les éléments n'appartenant pas à Ω' sont X-compatibles, soit à gauche, soit à droite de la chaîne $C(d,\Theta')$
c) la distance entre un élément à gauche et un élément à droite de cette chaîne n'appartenant pas à Ω' est supérieure à la distance entre deux éléments consécutifs quelconques selon l'ordre Θ'.

5.3. Aspect matriciel

Notons $M(d,\Theta,\Theta')$ une matrice $M(d,\Theta)$ où Θ est un ordre sur Ω, identique à Θ' sur Ω'.

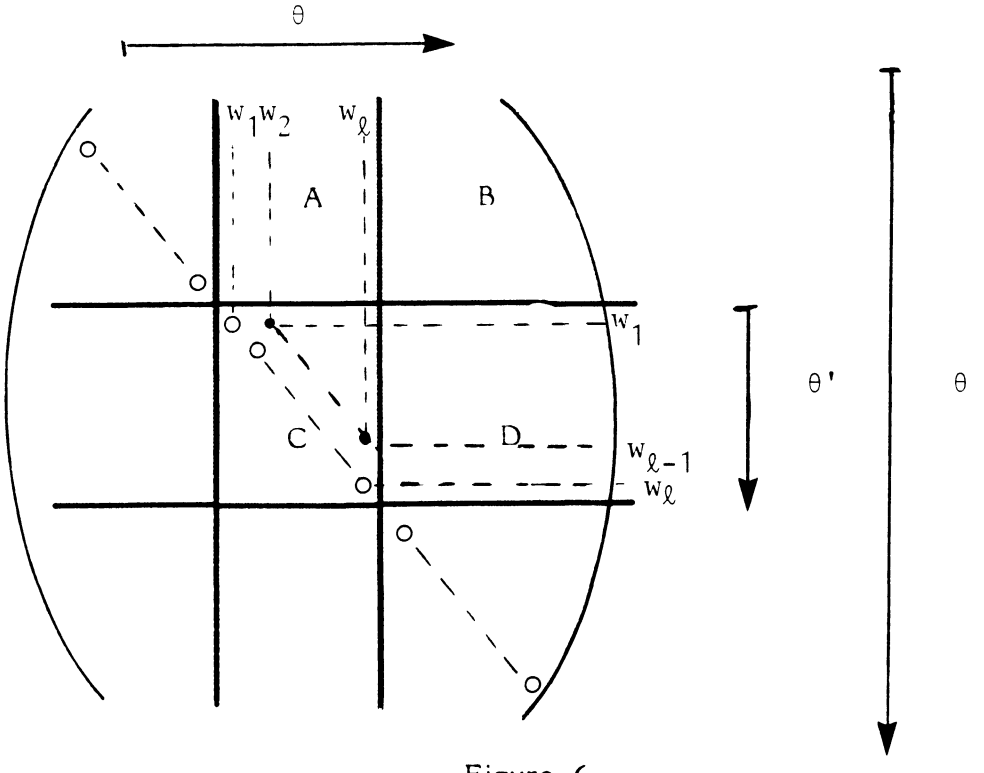


Figure 6

A partir de la matrice $M(d,\Theta,\Theta')$ on peut définir quatre matrices A, B, C, D comme indiquées figure 6. Une matrice $M(d,\Theta,\Theta')$ est dite quasi-Robinson si et seulement si :

a) la matrice C est Robinson
b) tous les termes de B sont supérieurs aux termes de la sur-diagonale de C
c) tous les termes de A (resp. D) sont supérieurs au terme de la première ligne (resp. dernière colonne) de C qui leurs correspond sur la même colonne (resp. ligne).

Une matrice $M(d,\Theta,\Theta')$ est dite quasi-SDR si et seulement si :

a) C est SDR
b) B satisfait à la même propriété que précédemment
c) tous les termes de A (resp. D) sont supérieurs à tous les termes de la sur-diagonale de C qui précèdent (resp. suivent) selon l'ordre la colonne (resp. ligne) correspondante.

Une matrice $M(d,\Theta,\Theta')$ est dite quasi-SDD si et seulement si :

a) C est SDD
b) B satisfait à la même propriété que précédemment
c) tous les termes de A (resp. D) sont supérieurs au terme de la sur-diagonale de C qui leur correspond sur la même colonne (resp. ligne).

Il résulte facilement de ces trois définitions que :

a) l'ensemble des matrices quasi-Robinson est inclus dans l'ensemble des matrices quasi-SDR qui est lui même inclus dans l'ensemble des matrices SDD
b) la compatibilité, la semi-compatibilité et la compatibilité faible de d et Θ' sur Ω équivaut à dire que $M(d,\Theta,\Theta')$ est respectivement quasi-Robinson, quasi-SDR et quasi-SDD.

6 - PROPRIETES DE LA X-COMPATIBILITE

6.1. Arbre de longueur minimum et semi-compatibilité de d et '

Proposition 2

Si d et Θ' sont semi-compatibles sur Ω alors la chaîne $C(d,\Theta')$ est incluse dans un arbre de longueur minimum.

Démonstration

Rappelons que l'ordre Θ' est noté $w_1 \ldots w_\ell$; en raisonnant par récurrence, il suffit de montrer que si les éléments $w_1 \ldots w_j$, $j < \ell$ sont consécutifs dans l'arbre alors w_{j+1} se connecte nécessairement à w_j ; en effet, si d et Θ' sont semi-compatibles sur Ω, w_{j+1} ne peut se connecter obligatoirement avec un élément w_g à gauche, ni avec un élément w_d à droite de $C(d,\Theta')$ puisque :

$d(w_g,w_{j+1}) \geq d(w_j, w_{j+1})$ et $d(w_d,w_j) \geq d(w_j, w_{j+1})$;

d'autre part comme d et Θ' sont semi-compatibles sur Ω, ils le sont sur Ω' et donc parmi les éléments qui suivent w_j dans l'ordre Θ', le plus proche est w_{j+1}. □

Une chaîne $C(d,\Theta')$ incluse dans un arbre est dite "sans branche" si les éléments w_2 , ..., $w_{\ell-1}$ sont connectés à l'arbre uniquement par l'élément qui précède et l'élément qui suit dans la chaîne (voir figure 7)

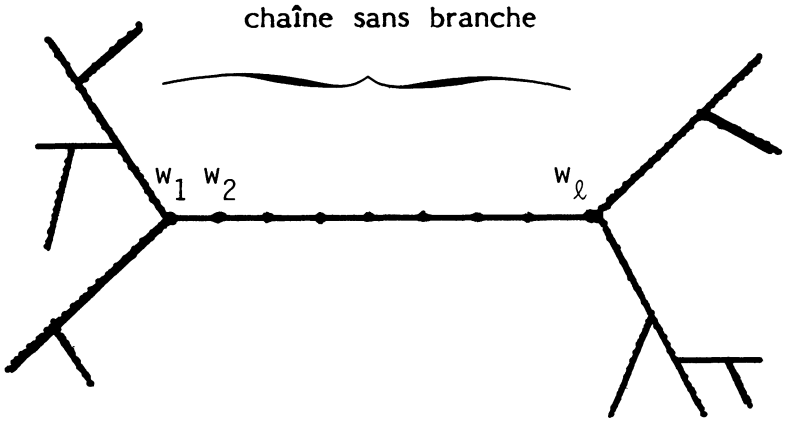


Figure 7

Proposition 3

Si $C(d,\Theta')$ est une chaîne sans branche incluse dans un arbre de longueur minimum alors d et Θ' sont semi-compatibles sur Ω .

Démonstration

Si $C(d,\Theta')$ est une chaîne sans branche incluse dans un arbre de longueur minimum on a les trois propriétés suivantes :

a) d et Θ' sont semi-compatibles sur Ω'. En effet, la distance entre deux éléments quelconques w et w' de Ω' est supérieure à la distance de deux éléments consécutifs w_j, w_{j+1} compris entre w et w' selon l'ordre Θ', sinon l'arbre ne serait pas de longueur minimum. (l'arbre obtenu en coupant $w_j w_{j+1}$ et reliant w et w' serait plus court)

b) L'ensemble des éléments notés A_1 (resp. A_2) qui sont sommets de la partie connexe de l'arbre contenant w_1 (resp. w_ℓ) obtenu en coupant l'arête $w_1 w_2$ (resp. $w_{\ell-1} w_\ell$) sont à gauche (resp. à droite) de $C(d,\Theta')$; en effet, soit w_g l'un des éléments de A_1, on a $d(w_g,w_j) \geq d(w_{i-1},w_i)$ $\forall$ i $\in$ { 2,...,j } car sinon on pourrait remplacer l'arête w_{i-1} w_i par l'arête $w_g w_j$ et l'arbre ne serait pas de longueur minimum ;donc w_g est un élément à gauche de la chaîne $C(d,\Theta')$. Par un raisonnement analogue, on montre de même que tout élément w_d de A_2 est à droite de $C(d,\Theta')$; il en résulte que tous les éléments de Ω n'appartenant pas à Ω' sont semi-compatibles soit à gauche, soit à droite.

c) On voit enfin que la distance entre un élément $w_g \in A_1$ et un élément $w_d \in A_2$ est supérieure à la distance de deux éléments consécutifs de $C(d, \Theta')$ car si elle était strictement inférieure l'arbre ne serait pas de longueur minimum. □

6.2. Croisement et semi-compatibilité de d et Θ'

Définition

On dit qu'un ordre Θ' sur $\Omega' \subset \Omega$ donne lieu à un croisement pour une hiérarchie H quand il existe h $\in$ H et trois éléments $w_\ell w_j w_k$ de Ω' avec $\ell < j < k$ tels que $w_j \notin h$ alors que w_ℓ et w_k sont dans h.

Proposition 4

Si la chaîne $C(d,\Theta')$ est contenue dans un arbre de longueur minimum, alors l'ordre Θ' sur $\Omega' \subset \Omega$ ne donne pas lieu à un croisement pour une hiérarchie du saut minimum.

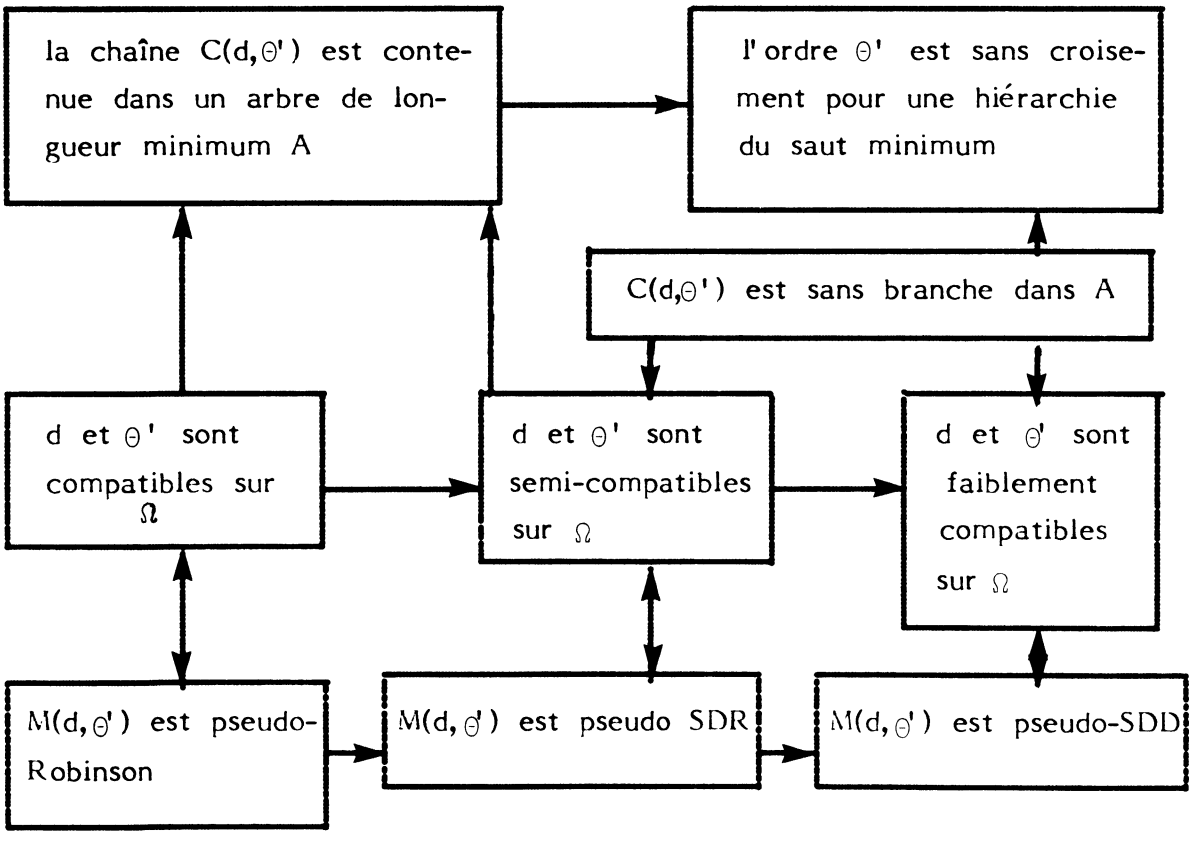
Démonstration

Etant donné un arbre de longueur minimum A, on sait (voir 4) que la hiérarchie du saut minimum H_A est telle qu'à chaque palier $h \in H_A$ on puisse associer une partie connexe de l'arbre A.

Considérons une chaîne $C(d,\Theta')$ où $\Theta' : w_1 \ldots w_\ell$ avec $\ell \leq n$ contenue dans A ; si Θ' donnait lieu à un croisement pour la hiérarchie H_A, il existerait un palier $h \in H_A$ tel que $w_p \in h$, $w_q \in h$ et $w_j \notin h$ avec $p < q < j$, la partie de l'arbre de longueur minimum A qui contient les éléments de h ne serait pas connexe, ce qui est contraire au choix de H_A. □

Il résulte de cette proposition que si d et Θ' sont semi-compatibles, alors Θ' est sans croisement pour une hiérarchie du saut minimum.

On peut résumer l'ensemble des résutats concernant un ordre partiel Θ' par le tableau 2.



$A \leftrightarrow B$: A et B équivalents ; $A \rightarrow B$: A implique B et B n'implique pas A

Tableau 2

7 - LES PRINCIPALES PROPRIETES DANS LE CAS OU Θ EST UN ORDRE SUR Ω.

Dans tout ce qui suit Θ est un ordre total sur Ω . Il résulte facilement des propositions 2 et 3, le résultat suivant, qui complète l'étude de Hubert (1974) sur le lien entre arbre de longueur minimum et compatibilité .

Proposition 5

Une condition nécessaire et suffisante pour qu'une chaîne $C(d,\Theta)$ soit un arbre de longueur minimum est que d et Θ soient semi-compatibles. De même on déduit de la proposition 4, la proposition suivante :

Proposition 6

Si la chaîne $C(d, \Theta)$ est un arbre de longueur minimum, alors Θ est sans croisement pour la hiérarchie du saut minimum.

8 - LIEN AVEC LA NOTION DE CHAINE t-MINIMAX

Rappelons d'abord la définition de cette notion qui a été introduite par B. Leclerc (1977).

Définition

Soit, $w, w' \in \Omega$; une chaîne $C(d,\Theta')$ où Θ' : $w = w_1, \ldots, w_\ell = w'$ est un ordre total sur $\Omega' \subset \Omega$ est une chaîne minimax si et seulement si pour tout ordre Θ'_1 : $w = w^1_1 \ldots w^1_{\ell_1} = w'$ avec w^1_i quelconque dans Ω , on a

$$\max_{(w_i,w_{i+1}) \in \Theta'} d(w_i,w_{i+1}) \leq \max_{(w^1_j,w^1_{j+1}) \in \Theta'_1} d(w^1_j, w^1_{j+1})$$

$C(d,\Theta')$ est t-minimax si toute chaîne $C(d, \Theta'')$ où Θ'' : $w_i \ldots w_j$ est une suite d'éléments consécutifs de Θ' est une chaîne minimax.

B. Leclerc a démontré (1977) qu'une caractérisation des chaînes qui sont dans un arbre de longueur minimum est qu'elles soient t-minimax. Il résulte de ce résultat et de la proposition 5 qu'il y a équivalence entre le fait que $C(d,\Theta)$ soit t-minimax et que d et Θ soient semi-compatibles.

9 - CAS OU L'INDICE DE DISSIMILARITE EST UNE ULTRAMETRIQUE

Dans le cas où d est une ultramétrique que nous noterons δ, les divers types de compatibilité sont équivalents. On a en effet, les deux propositions suivantes :

Proposition 7

Si δ est une ultramétrique et si $M(\delta,\Theta)$ est SDD alors $M(\delta,\Theta)$ est une matrice de Robinson.

Démonstration

Si $M(\delta,\Theta)$ est SDD, on a $\delta_{ij+1} \geq \delta_{jj+1}$ pour $i \leq j$, si de plus δ est une ultramétrique, on a : $\delta_{ij} \leq \text{Max}\,(\delta_{ij+1},\delta_{jj+1})$; d'où $\{\delta_{ij} \leq \delta_{ij+1}$ pour $i \leq j\}$ qui est la condition "ligne" (voir tableau 1) à satisfaire par $M(\delta,\Theta)$ pour être Robinson.
De même si $M(\delta,\Theta)$ est SDD, on a $\delta_{i-1j} \geq \delta_{j-1j}$, en utilisant de même le fait que δ est une ultramétrique, on a $\delta_{ij} \leq \max\,(\delta_{i-1j},\delta_{i-1i}) = \delta_{i-1j}$ si $i \leq j$ d'où la condition "colonne" à satisfaire par $M(\delta,\Theta)$ pour être de Robinson.

Signalons que I.C. Lerman (1981) a donné aussi une caractérisation d'une matrice $M(\delta,\Theta)$.

Nous dirons qu'une matrice $M(d,\Theta)$ est SDDL (resp. SDDC) si dans la matrice triangulaire supérieure associée à $M(d,\Theta)$ chaque terme de la sur-diagonale est inférieur à tous les termes de la ligne (resp. colonne) qui le contient (voir figure 8).

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

Matrice SDDL

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

Matrice SDDC

Figure 8

Plus précisément une matrice est SDDL (resp. SDDC) si elle satisfait à la condition "ligne" (resp. "colonne") donnée dans le tableau 1 ; autrement dit, $M(d,\Theta)$ est SDDL si $d_{ii+1} \leq d_{ij+1}$ pour $i \leq j$, elle est SDDC si $d_{j-1j} \leq d_{i-1j}$ pour $i \leq j$. A l'aide de ces définitions, on peut énoncer le résultat suivant :

Proposition 8

Une condition nécessaire et suffisante pour qu'un ordre Θ soit sans croisement pour une hiérarchie H est que la chaîne induite par Θ soit de plus courte longueur au sens de l'ultramétrique δ_H induite par H.

Démonstration

La condition suffisante résulte de la proposition 4. La condition nécessaire est prouvée par la suite des implications suivantes :

$$\delta_{ii+1} \leq \delta_{ii+2} \leq \cdots \leq \delta_{ij-1} \leq \delta_{ij} \leq \text{Max}\,(\delta_{jj+1}, \delta_{ij+1}) = \delta_{ij+1}$$

d'où $\{\delta_{ii+1} \leq \delta_{ij+1}$ pour $i \leq j\}$ qui prouve que $M(\delta,\Theta)$ est SDDL.

Ainsi quand δ est une ultramétrique, une matrice $M(\delta,\Theta)$ qui est SDDL (resp. SDDC) étant aussi SDDC (resp. SDDL) elle est SDD. □

On peut résumer l'ensemble des résultats obtenus dans le cas où Θ est un ordre total sur Ω, à l'aide du tableau 3. On note δ_d la sous-dominante de d (i.e. ultramétrique inférieure maximum de d) et H_d une hiérarchie du saut minimum.

Remarquons que dans le cas où d est une ultramétrique ou a $d = \delta_d$ et toutes les relations du tableau 3 se transposent en équivalences.

Conclusion

De nombreuses directions de recherche restent ouvertes : Etendre les résultats obtenus au cas des populations différentes, à la comparaison d'arbres, à des cycles. Voir ce qui se passe dans le cas de distances autres que des ultramétriques (quasi-ordres, semi-ordres, etc...).

Il serait intéressant de voir plus précisément comment sont situés dans le plan ou dans un espace à trois dimensions les points à gauche et à droite suivant le différent type de chaîne et dans les différents cas de compatibilité entre l'ordre et la distance. Les aspects algorithmiques des résultats théoriques obtenus ici sont abordés dans [5].

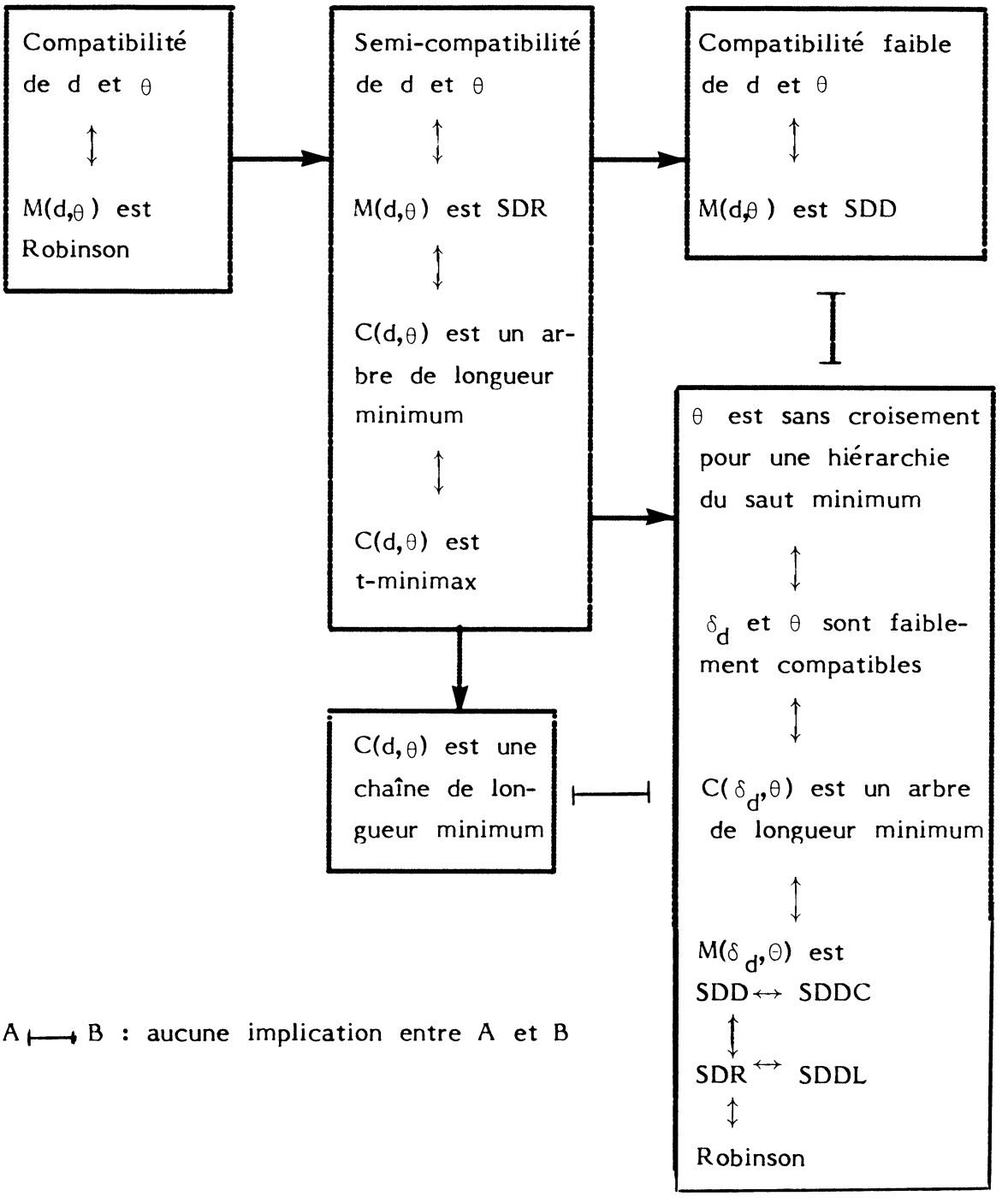


Tableau 3

BIBLIOGRAPHIE

[1] E.N. ADAMS., "Consensus techniques and the comparison of taxonomic trees", Syst. Zool, 21, pp. 390-397, (1972).

[2] G. BROSSIER., "Représentation ordonnée des classifications hiérarchiques", Statistique et Analyse des données, Vol. 2., pp. 31-44, (1980).

[3] J.L. CHANDON., J. LEMAIRE., J. POUGET., Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés, RAIRO, 14, 2, pp. 157-170, (1980).

[4] E. DIDAY., J. LEMAIRE., J. POUGET., F. TESTU., Elements d'analyse des données, Paris, DUNOD, (1982).

[5] E. DIDAY., "Croisements, ordres et ultramétriques : application à la recherche de consensus en classification automatique", Rapport de Recherche INRIA, n° 144, (1982).

[6] E. DIDAY., "Problèmes d'inversions en classification hiérarchique", Revue de Statistique appliquée, Vol. XXXI, n° 1, pp. 45, (1982).

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

[8] O. FRANK., K. SVENSSON., "On probability distributions of single-linkage dendograms", J. Stat. Comput. Simul., 12, pp. 121-131, (1981).

[9] L. HUBERT., "Some applications on graph theory and related non-metrics techniques to problems of approximate seriation", The British Journal of Mathematical and Statistical Psychology, Tome 27, pp. 133-153, (1974).

[10] L. HUBERT., F. BAKER., "The comparison and filting of given classification schemes", J. Math. Psychol. 16, pp. 233-253, (1977).

[11] D.G. KENDALL., "Incidence matrices : interval graphs and seriation in archeologic", Pacific J. Math. 28, (1969).

[12] M.F. MICKEVICH., Taxonomic congruence, Ph. D. Dissertation, State Univ. of New York at Stony Brook, 70 pp., (1978).

[13] B. LECLERC., "An application of combinatorial theory to hierarchical classification" in : BARA J.L. et Al. Eds., Recent Developments in Statistics, Amsterdam North Holland, (1977).

[14] B. LECLERC., "Description combinatoire des ultramétriques". Math. Sci. hum.19ème année, n° 73, pp. 5-37, (1981).

[15] I.C. LERMAN., Classification automatique et analyse ordinale des données, Paris Dunod, (1981).

[16] F.J. ROHLF., "Consensus indices for comparing classsifications", IBM Research Report R.C. 8940, (1981).