MATHÉMATIQUES ET SCIENCES HUMAINES

BRUNO LECLERC

Arbres minimums communs et compatibilité de données de types variés

Mathématiques et sciences humaines, tome 98 (1987), p. 41-67

<http://www.numdam.org/item?id=MSH_1987__98__41_0>

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1987, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (http://msh.revues.org/) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Math. Sci. hum. (25^{e} année, n°98, *1987*, p.41-67)

ARBRES MINIMUMS COMMUNS ET COMPATIBILITE DE DONNEES DE TYPES VARIES

Bruno LECLERC *

INTRODUCTION

La reconnaissance d'ordres totaux compatibles avec des données de divers types est l'un des objets de la sériation. Par exemple, un problème classique est de repérer si un ensemble de parties d'un ensemble fini X est, ou non, un ensemble d'intervalles d'un certain ordre total sur X (cf. e.g. Golumbic 1980, ch.8). Un algorithme efficace a été proposé pour résoudre ce problème (Booth et Leuker 1976). De même, Diday (1984) a étudié diverses formes de compatibilité d'un ordre total sur X avec une classification ou une dissimilarité sur X. Il a posé le problème de la reconnaissance d'un ordre total sur X (ou d'ordres totaux sur des parties de X) simultanément compatible avec plusieurs classifications hiérarchiques indicées ou avec plusieurs pyramides (Diday 1982, 1986). Un problème analogue, mais où les dissimilarités ne sont pas symétriques, est étudié par Doignon et al. (1986).

On aborde ici une question de ce type, mais en cherchant un arbre sur X (au sens de la théorie des graphes) simultanément compatible avec plusieurs données, de types divers, sur X. Si l'on spécifie que l'arbre est une chaîne, on retrouve le problème de la recherche d'un ordre total. En considérant des arbres, le problème d'existence de la configuration recherchée recevra donc plus fréquemment une réponse positive. De plus, cette réponse sera en général plus facile à obtenir, la différence entre les deux problèmes étant du même ordre qu'entre deux autres bien connus : déterminer si un graphe donné admet un arbre (c'est-à-dire en fait s'il est connexe) est bien plus aisé que de reconnaître s'il admet une chaîne hamiltonienne, c'est-à-dire un arbre qui est une chaîne.

* Centre d'Analyse et de Mathématique Sociales, E.H.E.S.S., Paris.

On veut aussi élargir le champ des données, ou des structures, qui peuvent ou non être compatibles avec un arbre. Il faut donc partir d'une formalisation adéquate de l'idée de compatibilité, utilisable dans des situations diverses. On considère ici la suivante : à des données de types variés (qualitatives, ordinales, classificatoires, ...), on associe canoniquement des préordonnances, c'est-à-dire des préordres sur l'ensemble $P = X(2)$ des paires d'éléments de X. On le fait de sorte que la paire xy est avant la paire zt lorsque, et seulement lorsque, la donnée considérée fait apparaître x et y comme au moins aussi proches, ou aussi ressemblants, entre eux que ne le sont z et t. Les préordonnances obtenues ne sont pas totales en général : si, par exemple, on a deux classes distinctes Y et Z d'une partition Π sur X, avec $xy \subseteq Y$ et $zt \subseteq Z$, il n'y a pas de raison d'affirmer que la paire xy est équivalente à la paire zt, encore moins de les ordonner. Mais par contre x et y seront jugés plus ressemblants entre eux que, par exemple, x et z.

Avec cet "mise en préordonnances", on trouve alors une expression générale de l'idée de compatibilité : un arbre A sur X est compatible avec une certaine donnée sur X si et seulement si A est un arbre minimum pour la préordonnance sur X induite par cette donnée. Les exemples de compatibilité donnés au début de cette introduction entrent bien dans ce cadre : l'ordre sur X associé à une dissimilarité de Robinson (i.e. un "indice pyramidal" au sens de Diday) correspond à un arbre minimum qui est une chaîne (Hubert 1974); une famille d'intervalles est aussi une famille arborée, c'est-à-dire dont chaque élément correspond à un sous-arbre d'un certain arbre A : celui-ci est un arbre minimum pour une certaine préordonnance sur X induite par la famille considérée (Flament 1975, 1978).

Formellement, le problème posé s'énonce donc maintenant comme suit : rechercher un arbre minimum commun à plusieurs préordonnances, totales ou partielles. Ce problème se rencontre aussi à propos du consensus de classifications : la médiane d'un ensemble d'ultramétriques, ou de partitions, a de bonnes propriétés lorsqu'il y a un arbre minimum commun (Barthélemy, Leclerc et Monjardet 1984b, 1986). On va montrer qu'il a une solution simple, en s'appuyant essentiellement sur l'étude par Flament et Leclerc (1983) des arbres minimums et minimaux pour une préordonnance quelconque. On va en particulier présenter une procédure de calcul de cette solution ne mettant en jeu que des opérations de routine (addition de dissimilarités, calculs d'arbres minimums), et souvent utilisable. Lorsque de tels arbres existent, ils fournissent des sériations partielles cor-

respondant à leurs chaînes, et des dichotomies correspondant à leurs arêtes. Les unes et les autres ont des propriétés intéressantes; leur existence peut aussi être un argument pour la validation d'hypothèses structurelles sur l'ensemble X étudié.

La section 1 de l'article présente les outils : graphes, préordonnances et dissimilarités (par. 1.1), et arbres minimums pour une préordonnance quelconque, avec leurs propriétés (par. 1.2). Au par. 1.3, on décrit les préordonnances, en général partielles, que l'on associe à divers types de données qualitatives ou ordinales et on observe qu'elles admettent des arbres minimums, en général fort nombreux.

On aborde le problème de la recherche d'un arbre minimum commun dans la section 2. On y propose une procédure (par. 2.1) que l'on illustre par un exemple (par. 2.2), où un tel arbre est effectivement produit (par. 2.3). La section 3, plus théorique, apporte les justifications nécessaires. Au par. 3.1, on caractérise les arbres minimums communs et on propose trois procédures pour les obtenir lorsqu'ils existent; elles diffèrent par les opérations combinatoires mises en jeu. L'objet du par. 3.2 est l'étude du cas particulier des familles arborées, dont on retrouve des caractérisations déjà connues. On conclut par deux brèves discussions, d'abord du problème de la recherche d'un arbre minimum-chaîne, c'est-à-dire d'un ordre total compatible (par. 3.3), puis du cas où il n'y a pas d'arbre minimum commun (par. 3.4).

Ce travail se place dans le prolongement de l'ensemble d'approches et de méthodes combinatoires assez variées rassemblées dans l'analyse de similitude. On trouve une présentation récente de celle-ci, avec des exemples de son usage, dans un numéro spécial de la revue *Informatique et Sciences humaihumaines* (n° 67, décembre 1985).

1. ARBRES MINIMUMS DEFINIS A PARTIR D'UNE PREORDONNANCE

1.1. Définitions

1.11 Graphes, chaînes, arbres. On considère un ensemble fini de cardinal n, et l'ensemble $P = X(2)$ des *paires* d'éléments de X c'est-à-dire des parties de X de cardinal 2. Les éléments de P seront parfois appelés *arêtes*. Pour $Y \subset X$, on écrira $Y(2) = P_Y$. Pour alléger, on notera xy, plutôt que $\{x,y\}$, l'élément de P correspondant à $x,y \in X$ (distincts).

Un *graphe* (simple) sur X est un couple (X,P'), avec $P' \subseteq P$, le couple $K_X = (X,P)$ étant le graphe complet sur X. Une *chaîne* sur X est ici une partie de P de la forme $C = \{x_0x_1, x_1x_2, \ldots, x_{k-1}x_k\}$ où tous les éléments x_i, sauf éventuellement x_0 et x_k, sont distincts. Si $x_0 = x_k$, la chaîne C est un *cycle* sur X ; sinon, c'est une chaîne *entre* x_0 et x_k. Pour $P' \subset P$, C est une chaîne, ou un cycle du graphe (X,P') si et seulement si $C \subseteq P'$. Le graphe (X,P') est *connexe* si, pour tous $x,y \in X$ (distincts), il a une chaîne entre x et y.

Un *arbre* A sur X est ici une partie de P telle que le graphe (X,A) est connexe et sans cycles. On sait qu'alors $|A| = n-1$, et que pour tous $x,y \in X$ (distincts), il y a une chaîne unique, notée $A(x,y)$, entre x et y et incluse dans A. On associe à tout arbre A sa relation d'*échangeabilité* Δ_A, définie sur P. Pour $xy \in A$ et $zt \in P-A$, on a $(xy,zt) \in \Delta_A$ si et seulement si $(A - \{xy\}) \cup \{zt\}$ est encore un arbre sur X. On a $\Delta_A \subset A \times (P-A)$; on voit facilement que $(xy,zt) \in \Delta_A$ équivaut à $xy \in A(z,t)$.

La figure 1 montre un arbre A sur $X = \{\alpha,\beta,\gamma,\delta,\varepsilon,\zeta,\eta\}$. La chaîne $A(\alpha,\beta)$ y est en traits renforcés. On voit que chacune de ses paires, $\alpha\gamma$ ou $\gamma\eta$ ou $\eta\beta$, peut être échangée avec $\alpha\beta$ de façon à obtenir un nouvel arbre.

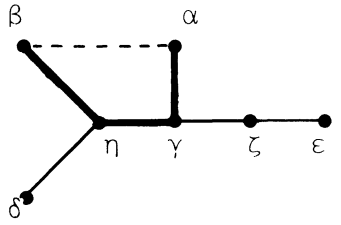

Figure 1.

1.12 Préordonnances et dissimilarités. Une *préordonnance* R sur X est un préordre, c'est-à-dire une relation réflexive et transitive, sur P.

$(\forall xy \in P) \quad (xy,xy) \in R$

$(\forall xy,zt,vw \in P) \quad (xy,zt) \in R \text{ et } (zt,vw) \in R \Rightarrow (xy,vw) \in R$.

Dans la suite, on écrira généralement $xy \leq zt$ plutôt que $(xy,zt) \in R$. Si plusieurs préordonnances sont considérées, elles seront notées $R_1, R_2, \ldots$ et l'on écrire $xy \leq_1 zt$, $xy \leq_2 zt, \ldots$ La notation $xy < zt$ correspondra au cas où $(xy,zt) \in R$ et $(zt,xy) \notin R$, celle $xy \sim zt$ à celui où $(xy,zt) \in R$ et $(zt,xy) \in R$ (paires équivalentes). Avec la définition ci-dessus, le cas de paires incomparables ($(xy,zt) \notin R$ et $(zt,xy) \notin R$) est possible : la préordonnance R n'est pas a priori supposée totale. On dira parfois d'une préordonnance qu'elle est *partielle* pour bien préciser qu'elle n'est pas totale.

Une *dissimilarité* sur X est ici une application $d : P \to \mathbb{R}^+$. Cette définition est clairement équivalente à celle, plus usuelle, selon laquelle une dissimilarité est une application de X^2 dans $\mathbb{R}^+$ vérifiant, pour tous $x,y \in X$, $d(x,x) = 0$ et $d(x,y) = d(y,x)$.

La dissimilarité d induit naturellement une préordonnance totale $R(d)$ sur X . Pour $xy,zt \in P$, on a

$$(xy,zt) \in R(d) \Leftrightarrow d(xy) \leq d(zt) .$$

On dira qu'une dissimilarité d et une préordonnance R sont *compatibles* si $R(d)$ préserve la comparabilité $\leq$ et la comparabilité stricte $<$ selon R , c'est-à-dire si :

$$(\forall xy,zt \in P) \quad xy \leq zt \Rightarrow d(xy) \leq d(zt) \quad \text{et}$$
$$xy < zt \Rightarrow d(xy) < d(zt) .$$

1.2 Arbres minimums

1.21 Définition. Etant donnée une préordonnance R sur X , on définit sur l'ensemble $\mathfrak{A}$ de tous les arbres sur X une relation de dominance DR . Pour deux arbres A et A' , on a $(A,A') \in DR$ si et seulement si il existe une bijection $\beta : A \to A'$ telle que :

$$(\forall a \in A) \quad a \leq \beta a .$$

On montre que DR est un préordre sur $\mathfrak{A}$. Ce préordre n'est pas total en général, même lorsque R est une préordonnance totale. Il peut donc ne pas avoir d'éléments minimums.

Un arbre A sur X est un *arbre minimum* pour la préordonnance R si et seulement si il est minimum pour le préordre DR, c'est-à-dire que l'on a :

$$(\forall A' \in \mathfrak{A}) \quad (A,A') \in DR .$$

Etant donnée une dissimilarité d sur X , on peut préordonner (totalement cette fois) les arbres sur X selon leurs longueurs, la longueur $\ell(A)$ d'un arbre étant définie, comme il est naturel, comme étant la somme des dissimilarités de ses éléments. Les arbres de longueur minimum sont les éléments minimums pour ce préordre.

Pour toute préordonnance totale R , il existe au moins un arbre minimum pour R (Rosenstiehl 1967). C'est en particulier le cas si $R = R(d)$ est la préordonnance associée à la dissimilarité d . On voit facilement, à partir

des définitions ci-dessus, qu'un arbre A est de longueur minimum si et seulement si il est minimum pour $R(d)$; on dira aussi dans ce cas qu'il est minimum pour d. Les algorithmes qui permettent de déterminer de tels arbres n'utilisent effectivement que des comparaisons de longueurs d'arêtes (voir par exemple celui donné dans Hartigan 1975). C'est ce caractère purement ordinal et combinatoire que l'expression "arbre de longueur minimum", souvent utilisée dans la littérature relevant de l'analyse des données, a l'inconvénient de masquer.

Dans le cas d'une préordonnance partielle P, il n'y a évidemment plus coïncidence entre arbres minimums pour R et arbres de longueur minimum, ceux-ci n'étant plus définis, mais un lien persiste, ce qu'établit la proposition suivante :

PROPOSITION 1.1. Soient A un arbre sur X et R une préordonnance sur X. Les trois conditions suivantes sont équivalentes.

(1) A est un arbre minimum pour R.

(2) A est un arbre de longueur minimum pour toute dissimilarité d compatible avec R.

(3) La relation Δ_A d'échangeabilité de l'arbre A est incluse dans R.

Résumé de la preuve. (1) $\Rightarrow$ (2) découle immédiatement des définitions données ci-dessus.

(2) $\Rightarrow$ (3). S'il existe $xy \in A$, $zt \in P-A$ tels que $(xy,zt) \in \Delta_A - R$, on peut construire une préordonnance totale R_1 telle que $R \subset R_1$ et que, pour tous $p,p' \in P$, on a $p < p' \Rightarrow p <_1 p'$. Alors, pour toute dissimilarité d' compatible avec R_1, et donc avec R, on a $\ell((A-\{xy\}) \cup \{zt\}) < \ell(A)$, ce qui contredit (2).

La démonstration de (3) $\Rightarrow$ (1) est plus technique. Il s'agit d'établir qu'un arbre localement minimum (si (3) est vraie, on a $(A,A') \in DR$ pour tout arbre A' ne différant de A que par un élément) est aussi globalement minimum. On en trouve une preuve complète dans Flament et Leclerc (1983), où il est aussi indiqué que c'est une conséquence d'un résultat de Brualdi (1969), donné dans le cadre plus général des bases d'un matroïde. □

L'équivalence de (1) et de (2) donne une première indication du fait que la définition précédente des arbres minimums est bien une extension au cas de préordonnances quelconques de celles des arbres minimums pour une dissimilarité.

L'équivalence des conditions (1) et (2) avec (3) donne une caractérisation de ces arbres minimums qui est simple et utilisable en pratique, car la relation Δ_A se construit facilement.

Ainsi, on peut déterminer si une préordonnance R admet un arbre minimum et, dans l'affirmative, obtenir un tel arbre, de la façon suivante :

- on détermine une dissimilarité d compatible avec R ;
- on construit un arbre A minimum pour d ;
- on détermine Δ_A et on vérifie si l'on a $\Delta_A \subset R$.

Si oui, A est un arbre minimum pour R . Sinon, la proposition 1.2 suivante permet d'affirmer qu'il n'y a pas d'arbre minimum pour R .

PROPOSITION 1.2. Soient R une préordonnance, d une dissimilarité compatible avec R et A un arbre sur X minimum pour d . Alors, si R admet un arbre minimum, A est un tel arbre.

Preuve. Ce résultat est une conséquence du fait suivant (Flament et Leclerc 1983) : un arbre est *minimal* pour le préordre DR si et seulement si il est *minimum* pour un préordre total compatible avec DR , tel celui induit par d . □

La caractérisation (3) de la proposition 1.1 apparaîtra aussi plusieurs fois dans la suite comme outil de démonstration.

1.22 Intérêt des arbres minimums en analyse des données

Le fait que les arbres minimums sont associés à la méthode de classification hiérarchique dite du lien simple est bien connu (Gower et Ross 1969). Plus généralement, ils constituent un outil pour la représentation et la manipulation (e.g. le calcul latticiel, la comparaison) des ultramétriques, ou hiérarchies indicées (Hubert 1977, Leclerc 1979, 1981a, 1986, Barthélemy, Leclerc, Monjardet 1986).

En fait, la donnée d'un arbre minimum pour une dissimilarité d apporte des informations sur la façon dont l'ensemble X s'organise selon d : Du point de vue de la *sériation*, les chaînes d'un arbre minimum se caractérisent par une propriété d'optimalité correspondant à l'idée de ressemblance de proche en proche (Leclerc 1977, 1981a). On dit qu'elles sont totalement minimax (Leclerc), ou semi-compatibles avec d (Diday 1983) ou très minimales (Giraudet 1982).

Du point de vue de la *classification*, la recherche, pour toute paire xy d'éléments de X, d'une dichotomie de X séparant x et y et optimale, a une solution donnée par l'une des bipartitions associées aux arêtes d'un arbre minimum (Leclerc 1977, 1981a).

Ces propriétés vont être rappelées, en se situant dans l'hypothèse, beaucoup plus faible, où ce n'est pas une dissimilarité, mais seulement une préordonnance R, non forcément totale, qui est définie sur X.

Malgré la faiblesse d'une telle hypothèse, il reste des critères pour comparer des ensembles de paires d'éléments de X, comme des chaînes. On va établir que, pour un certain critère ordinal, l'existence d'un arbre minimum implique celle de chaînes optimales entre toute paire d'éléments de X. Commençons par rappeler la définition d'un cocycle sur X, ensemble d'arêtes associé à une dichotomie de X :

Soient $x,y \in X$ (distincts) et soit $\{X',Y'\}$ une dichotomie de X telle que $x \in X'$ et $y \in Y'$. Alors, l'ensemble $D = D(X',Y') = \{x'y' \in P / x' \in X'$ et $y' \in Y'\}$ est un *cocycle sur* X *séparant* x et y. On remarque que pour toute chaîne C entre x et y on a $C \cap D \neq \emptyset$.

Pour un arbre A et une paire $xy \in A$, il y a un cocycle unique, noté $D_A(xy)$, séparant x et y et tel que $D_A(xy) \cap A = \{xy\}$. C'est le cocycle associé à la dichotomie $\{X',Y'\}$ de X où X' et Y' sont les sous-ensembles de X correspondant aux composantes connexes du graphe $(X, A - \{x,y\})$ (figure 2). On a $D_A(xy) = \{zt \in P / xy = zt$ ou $(xy,zt) \in \Delta_A\}$
$= \{zt \in P / xy = zt$ ou $xy \in A(z,t)\}$ (cf. par exemple Leclerc 1981a et Flament et Leclerc 1983).

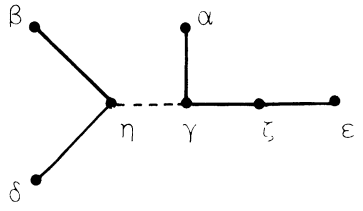

Figure 2.

$D_A(\eta\gamma) = D(\{\alpha,\gamma,\varepsilon,\zeta\},\{\beta,\delta,\eta\}) = \{\eta\gamma,\alpha\beta,\alpha\delta,\alpha\eta,\beta\gamma,\beta\varepsilon,\beta\zeta,\gamma\delta,\delta\varepsilon,\delta\zeta,\varepsilon\eta,\zeta\eta\}$.

Deux chaînes C et C' entre x et y représentent des suites de "sauts" pour passer, de proche en proche, de x à y. Si pour toute arête p de C, on trouve une arête $p' \in C'$ telle que $p \leq p'$, c'est-à-dire si pour tout saut de C, il y a un saut au moins aussi important de C', on

considèrera que C définit une suite d'intermédiaires entre x et y au moins aussi pertinente que celle définie par C' . C'est pour ce critère que toute chaîne incluse dans un arbre minimum est optimale.

PROPOSITION 1.3. Soient A et R respectivement un arbre et une préordonnance sur X . Si A est un arbre minimum pour R , alors pour tous $x,y \in X$ (distincts) et pour tout $a \in A(x,y)$, il existe, pour toute chaîne C entre x et y , une arête $p \in C$ telle que $a \leq p$.

Preuve. Considérons un élément $p \in D_A(a) \cap C$. On a soit $p = a$, soit $(a,p) \in \Delta_A$. Dans les deux cas on a $a \leq p$ puisque A est un arbre minimum et donc que $\Delta_A \subset R$. □

Quand $R = R(d)$ est la préordonnance totale associée à une dissimilarité d , on retrouve la propriété définissant les chaînes totalement minimax. Pour tous $x,y \in X$, pour toute chaîne C entre x et y , on a :

$$\max\{d(a) / a \in A(x,y)\} \leq \max\{d(p) / p \in C\} .$$

Cherchons maintenant à comparer les dichotomies de X , en fait les cocycles sur X , à partir de la seule donnée d'une préordonnance sur X . Soient $x,y \in X$ (distincts), avec deux cocycles D et D' sur X séparant tous deux x et y . On dira que D sépare x et y au moins aussi bien que D' si, pour tout $p \in D$, il existe $p' \in D'$ tel que $p' \leq p$. Pour ce critère, tout cocycle $D_A(xy)$ est, si A est un arbre minimum pour R , optimal pour la séparation de x et de y .

PROPOSITION 1.4. Soient A et R respectivement un arbre et une préordonnance sur X . Si A est un arbre minimum pour R , pour tout $xy \in A$ et pour tout $p \in D_A(xy)$, il existe, pour tout cocycle D' sur X séparant x et y , une paire $p' \in D'$ telle que $p' \leq p$.

Preuve. Si $p \in D_A(xy)$, on a $(xy,p) \in \Delta_A$, donc $xy \leq p$. Or, si D sépare x et y , on a $xy \in D$. □

Pour x et y quelconques, il peut ne pas exister de cocycle optimal selon le critère précédent pour la séparation de x et de y . Par contre, lorsque $R = R(d)$ est la préordonnance totale associée à une dissimilarité d , il y a, pour la séparation de tous $z,t \in X$ (distincts), un cocycle optimal associé à l'arbre minimum. On retrouve dans ce cas la propriété ca-

ractéristique des *cocycles maximum* (Leclerc 1977, 1981a) : soit $xy \in A$ tel que $d(xy) = \max\{d(a) / a \in A(z,t)\}$ et soit un cocycle D sur X séparant z et t. Alors :

$$\min\{d(p) / p \in D\} \leq \min\{d(p') / p' \in D_A(xy)\} = d(x,y) .$$

1.3 Préordonnances associées à divers types de données qualitatives ou ordinales.

1.31 Parties. Une partie F de X peut par exemple être l'ensemble des éléments de X qui possèdent un certain caractère. On lui associe la préordonnance R définie par $xy \leq zt$ si et seulement si $xy \subseteq F$ ou $zt \not\subseteq F$, c'est-à-dire :

$$(\forall p,p' \in P) \quad (p,p') \in R \Leftrightarrow [p' \subseteq F \Rightarrow p \subseteq F] .$$

C'est une préordonnance totale à deux classes. Elle admet donc des arbres minimums. Un arbre A est minimum pour R si F détermine une partie connexe de A, c'est-à-dire si $A \cap P_F$ est un arbre sur F.

Posons $f = |F|$. Une formule due à Moon (1967) permet de calculer le nombre $\tau(R)$ des arbres minimums pour R. On a :

$$\tau(R) = n^{n-f-1} f^{f-1} .$$

1.32 Partitions. Soit $\Pi = \{X_1,\ldots,X_k\}$ une partition de X.

On lui associe la préordonnance R définie par $xy \leq zt$ si et seulement si : x, y, z et t sont dans la même classe de Π, ou z et t ne sont pas dans la même classe. Si le nombre de classes de Π ayant au moins deux éléments est supérieur à 1, R n'est pas totale : en prenant, par exemple $x,y \in X_1$ et $z,t \in X_2$, avec $x \neq y$ et $z \neq t$, on a les paires incomparables xy et zt.

Dans cet exemple, comme dans les deux suivants on a une famille de parties de X. Un arbre A sur X sera considéré comme compatible avec cette famille si chaque partie de X considérée (ici, chaque classe de Π) détermine une partie connexe de A. Il est équivalent de dire que A est un arbre minimum pour R : un tel arbre contient en effet, pour chaque X_i, $i = 1,\ldots,k$, un arbre sur X_i. Un calcul du même type que le précédent donne le nombre d'arbres minimums pour R :

$$\tau(R) = n^{k-2} \prod_{i=1}^{k} n_i^{n_i-1} ,$$

en posant $|X_i| = n_i$ pour tout $i = 1,\ldots,k$. Notons que tout ceci s'applique au cas d'une partition incomplète, dont les classes sont disjointes, mais ne recouvrent pas X tout entier.

1.33 Hiérarchies. Une *hiérarchie* sur X est une famille H de parties de X telle que : $\emptyset \notin H$, $X \in H$, $\{x\} \in H$ pour tout $x \in X$, et, pour $Y,Z \in H$, $Y \cap Z \in \{Y,Z,\emptyset\}$. Alors, l'ordre d'inclusion sur les éléments de H est arborescent. Les hiérarchies sont les familles de parties correspondant aux arbres de classification (Benzecri 1967).

Pour $x,y \in X$, notons alors xHy l'élément de H contenant x et y qui est minimum pour l'inclusion avec cette propriété. On associe à H la préordonnance R définie par : $xy \leq zt \Leftrightarrow xHy \subseteq zHt$.

On trouve dans la littérature des caractérisations de ces préordonnances hiérarchiques (Barthélemy, Leclerc et Monjardet 1984a, Leclerc 1985). Elles ne sont pas totales en général, mais elles ont des arbres minimums, qui ont été dénombrés (Leclerc 1985; la formule de Moon signalée ci-dessus est rappelée dans cet article). Pour les hiérarchies binaires (ou maximales pour l'inclusion), ce nombre $\tau(R)$ est compris entre $4^{n-1}/n^2$ et $(n-1)$!

1.34 Familles d'intervalles. Une famille I de parties de X est une famille d'intervalles si et seulement si il existe une indiciation $x_1,\ldots,x_n$ des éléments de X telle que tout élément J de I soit de la forme $\{x_k,x_{k+1},\ldots,x_{k+|J|-1}\}$, pour un certain entier k compris entre 1 et n .

Supposons que de plus on ait $X \in I$ et, pour tous $J,K \in I$, $J \cap K = \emptyset$ ou $J \cap K \in I$. Alors I est un ensemble de parties du type intervenant dans les pyramides, un modèle associant classification et sériation (Diday 1984, 1986, Bertrand 1986) et englobant les classifications hiérarchiques. On peut aussi dire que I lui-même est une pyramide (Batbedat 1986). Comme dans l'exemple précédent, dont celui-ci constitue d'ailleurs une extension, le plus petit élément xIy de I contenant x et y existe et on pose $xy \leq zt \Leftrightarrow xIy \subseteq zIt$. Clairement, cette préordonnance n'est pas totale et admet $A = \{x_1x_2,x_2x_3,\ldots,x_{n-1}x_n\}$ pour arbre minimum. Il ne semble pas que l'on puisse donner une formule générale simple pour le nombre d'arbres minimums pour une telle préordonnance.

1.35 Préordres totaux. Le cas de la donnée d'un préordre total T sur X peut être traité à partir du précédent, en prenant pour famille I celle des intervalles de T, c'est-à-dire des unions de classes consécutives de T. On laisse au lecteur le soin de vérifier que la préordonnance obtenue ainsi est identique à la préordonnance "de ressemblance" induite de la façon la plus naturelle par T, que l'on définit comme suit :

Soit par exemple un préordre total T sur X, dont les classes $X_1,\ldots,X_k$, ordonnées selon T, sont de cardinal $n_1,\ldots,n_k$ respectivement. Pour $x,y \in T$, on écrit $x \leq y$ pour $(x,y) \in T$. On associe à T la préordonnance R définie par $p \leq p'$ si et seulement si $p = xy$ et $p' = zt$ avec $z \leq x \leq y \leq t$. Une telle préordonnance n'est pas totale en général. Indiçons les éléments de X de façon à avoir $x_1 \leq x_2 \leq \ldots < x_n$. On vérifie facilement que la chaîne $C = \{x_1x_2, x_2x_2, \ldots, x_{n-1}x_n\}$ est un arbre minimum pour R. En fait, un arbre minimum sur R est la réunion d'arbres sur chacun des X_i, $i = 1,\ldots,k$, liés par des paires y_iy_i' telles que $y_i \in X_i$ et $y_i' \in X_{i+1}$, pour $i = 1,\ldots,k-1$.

On en déduit, avec la formule de Cayley (1857), selon laquelle le nombre d'arbres sur X est n^{n-2}, que le nombre d'arbres minimums pour R est ici $\tau(R) = \dfrac{1}{n_1 n_k} \prod_{i=1}^{k} n_i^{n_i}$.

2. LA RECHERCHE D'UN ARBRE MINIMUM COMMUN

2.1 Le problème

A divers types de données qualitatives, ordinales, ou classificatoires on a associé des préordonnances qui admettent des arbres minimums. On a vu que ceux-ci peuvent être fort nombreux, et donc les intermédiarités ou dichotomies associées à l'un d'entre eux peuvent apparaître comme étant peu significatives.

Par contre, si l'on considère simultanément plusieurs préordonnances de ce type, l'existence d'un arbre minimum commun est une observation importante. Elle met en évidence une compatibilité entre ces préordonnances, donc entre les données dont elles peuvent être issues : celles-ci s'accordent alors sur les intermédiarités et les dichotomies optimales étudiées au par. 1.22 ci-dessus. De plus, à l'origine d'un tel accord, il peut y avoir une structure arborescente sous-jacente liée, par exemple, à un processus de bifurcation.

En particulier, un arbre minimum commun qui est une chaîne indique un ordre total compatible avec chacune des préordonnances de départ. Cet ordre peut, par exemple, être de nature chronologique.

Il y a un autre domaine où l'existence d'un arbre minimum commun est un fait important : celui du consensus de classifications, précisément de partitions ou d'ultramétriques. En effet, l'ensemble des ultramétriques sur X qui admettent un arbre donné A comme arbre minimum est, ordonné par l'inclusion, un treillis distributif isomorphe à $(\mathbb{R}^+)^A$ (Leclerc 1979). On trouve dans Barthélemy, Leclerc et Monjardet (1986) une revue des propriétés caractéristiques, de la calculabilité et de l'optimalité de la médiane dans une telle structure. Dès qu'un ensemble d'ultramétriques admet un arbre minimum commun, on dispose d'une bonne solution au problème de les agréger, tandis que l'on ne connaît pas actuellement de solution équivalente pour l'agrégation d'un ensemble quelconque d'ultramétriques. Le cas important des partitions est un cas particulier de celui des ultramétriques.

Une procédure pour résoudre le problème de l'existence d'un arbre minimum commun, et pour trouver effectivement un tel arbre lorsqu'il existe, va être présentée, d'abord sous la forme d'un exemple sur des données d'exercice de divers types : une famille de parties, un préordre total, une partition et une hiérarchie, toutes définies sur le même ensemble $\{\alpha,\beta,\gamma,\delta,\varepsilon,\zeta\}$.

A chacune de ces données, on fera correspondre, conformément à ce que l'on a vu en 1.3 ci-dessus, une préordonnance admettant des arbres minimums. On associera à cette préordonnance une dissimilarité compatible, et on calculera la longueur d'un arbre minimum pour cette dissimilarité.

Ensuite une dissimilarité globale, tout simplement la somme des précédentes, sera calculée, un arbre minimum pour cette dissimilarité et sa longueur seront déterminés. Cette longueur sera comparée à la somme des longueurs des arbres valués précédemment obtenus et le résultat permettra de conclure quant à l'existence ou l'inexistence d'un arbre minimum commun aux préordonnances de départ.

2.2 Un jeu de données

2.21 Une famille de parties. On considère les parties F_1 , F_2 , F_3 , F_4 de X (fig.3) et les préordonnances totales à deux classes R_1 , R_2 , R_3 et R_4 qui leur sont respectivement associées comme en 1.31 ci-dessus.

$F_1 = \{\alpha,\gamma,\zeta\}$, $F_2 = \{\beta,\gamma,\eta\}$, $F_3 = \{\gamma,\delta,\zeta,\eta\}$, $F_4 = \{\gamma,\varepsilon,\zeta,\eta\}$.

Par exemple, la classe minimum de la préordonnance R_1 est l'ensemble des paires d'éléments de F_1 . On lui associe la dissimilarité d_1 , compatible avec R_1 , en posant $d_1(xy) = 0$ si $xy \subset F_1$ et $d_1(xy) = 1$ sinon. Un arbre A_1 minimum pour R_1 est de longueur $\ell_1(A_1) = n-|F_1| = 4$ pour la dissimilarité d_1 et il y a $\tau(R_1) = 3\ 087$ tels arbres. On associe de même à F_2 , F_3 et F_4 des dissimilarités d_2 , d_3 et d_4 , dont les arbres minimums, respectivement A_2 , A_3 et A_4 sont de longueurs $\ell_2(A_2) = 4$, $\ell_3(A_3) = \ell_4(A_4) = 3$. On a aussi $\tau(R_2) = \tau(R_1)$ et $\tau(R_3) = \tau(R_4) = 3\ 136$.

La table 1 ci-dessous donne la dissimilarité $d' = d_1 + d_2 + d_3 + d_4$. Pour $xy \in P$, $d'(xy)$ est le nombre des parties considérées qui ne contiennent pas la paire xy .

Table 1. La dissimilarité $d' = d_1 + d_2 + d_3 + d_4$

β	γ	δ	ε	ζ	η	
4	3	4	4	3	4	α
	2	4	4	4	3	β
		3	3	1	1	γ
			4	3	3	δ
				3	3	ε
					2	ζ

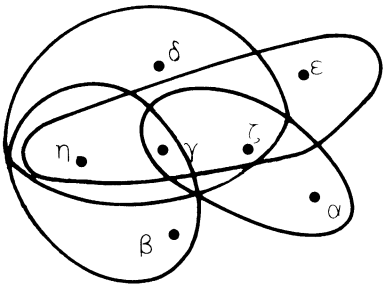

Figure 3. La famille de parties $\{F_1, F_2, F_3, F_4\}$

2.22 Une partition. A la partition $\Pi = \{\{\alpha,\gamma,\delta,\eta\},\{\beta\},\{\varepsilon,\zeta\}\}$ sur X est associée, comme en 1.32 ci-dessus, une préordonnance, notée R_5 , dont le diagramme est donné par la figure 4 ci-dessous.

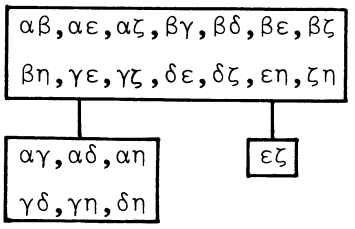

Figure 4. La préordonnance R_5

L'arbre A_5 de la figure 5 est l'un des $\tau(R_5) = 896$ arbres sur X minimums pour R_5 . Il est de longueur $\ell_5(A_5) = |\Pi| - 1 = 2$ pour la dissi-

milarité d_5 compatible avec R_5 obtenue en posant $d_5(xy) = 0$ si x et y sont dans la même classe de Π et $d_5(xy) = 1$ sinon. La dissimilarité d_5 est donnée par la table 2 :

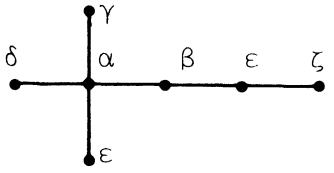

Figure 5. L'arbre A_5

β	γ	δ	ε	ζ	η	
1	0	0	1	1	0	α
	1	1	1	1	1	β
		0	1	1	0	γ
			1	1	0	δ
				0	1	ε
					1	ζ

Table 2. La dissimilarité d_5

2.23 Une classification hiérarchique. On considère la hiérarchie $H = \{\{\alpha,\gamma,\zeta\},\{\alpha,\gamma,\varepsilon,\zeta\},X,\{\beta,\delta,\eta\},\{\gamma,\zeta\},\{\delta,\eta\}\}$ représentée par la figure 6.

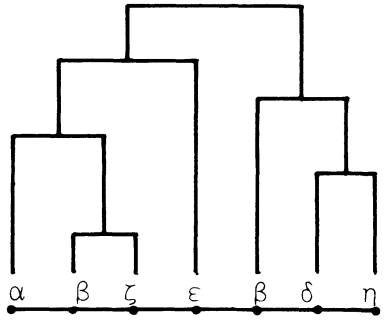

Figure 6. La hiérarchie H et l'arbre A_6

La préordonnance hiérarchique R_6 qui lui est associée est donnée par le diagramme de la figure 7. Elle a $\tau(R_6) = 144$ arbres minimums, dont celui indiqué dans la figure 6 . Celui-ci est une chaîne, qui correspond à un ordre sur X , compatible avec H au sens de Brossier (1980) et Diday (1984). C'est l'ordre d'alignement des éléments de X qui a été choisi pour la représentation de H dans la figure 6.

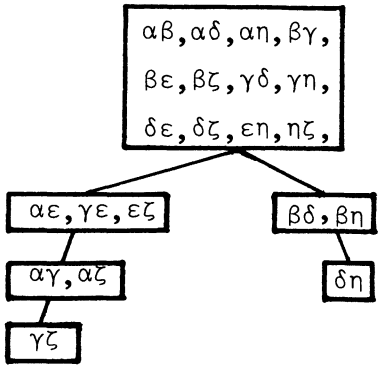

Figure 7. La préordonnance hiérarchique R_6

Toute ultramétrique ayant H pour hiérarchie est une dissimilarité compatible avec R_6 , par exemple celle, notée d_6 , donnée par la table 3 ci-dessous. La longueur de l'arbre A_6 pour d_6 est égale à 21 .

β	γ	δ	ε	ζ	η	
6	3	6	5	3	6	α
	6	4	6	6	4	β
		6	5	1	6	γ
			6	6	2	δ
				5	6	ε
					6	ζ

Table 3. La dissimilarité d_6

2.24 Un préordre total. On considère le préordre total T sur $X = \{\alpha,\beta,\gamma,\delta,\varepsilon,\zeta,\eta\}$ suivant : $T = \{\delta\}<\{\beta,\eta\} < \{\alpha,\gamma,\zeta\} < \{\varepsilon\}$, T étant décrit ci-dessus par ses classes et par l'ordre qu'il induit sur celles-ci. La préordonnance R_7 associée à T comme en 1.35 ci-dessus est donnée par le diagramme de la figure 8.

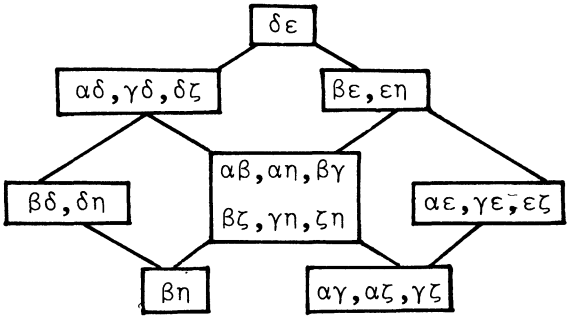

Figure 8. La préordonnance R_7

L'arbre chaîne A_7 de la figure 9 est minimum pour R_7 . C'est aussi le cas de l'arbre A'_7 dont on vérifie qu'il est équivalent à A_7 pour le préordre DR_7 , qui admet $\tau(R_s) = 108$ arbres minimums.

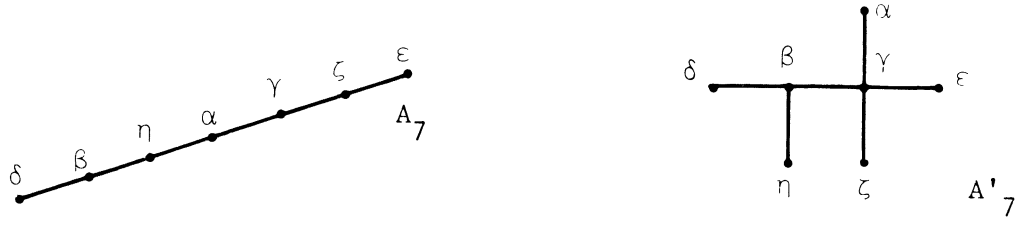

Figure 9. Deux arbres minimums pour la préordonnance R_7

Une dissimilarité d_7 compatible avec R_7 s'obtient en posant $d_7(xy) = 0$ si x et y sont dans la même classe du préordre T, $d_7(xy) = 1$ si x et y sont dans deux classes consécutives, et ainsi de suite. La dissimilarité d_7 est donnée par la table 4 suivante :

β	γ	δ	ε	ζ	η	
1	0	2	1	0	1	α
	1	1	2	1	0	β
		2	1	0	1	γ
			3	2	1	δ
				1	2	ε
					1	ζ

Table 4. La dissimilarité d_7

La longueur de l'arbre A_7 (ou de l'arbre A'_7) pour cette dissimilarité est égale à $\ell_7(A_7) = 3$.

2.3 Obtention d'un arbre minimum commun.

La dissimilarité $d = \sum_{i=1}^{7} d_i$ est donnée par la table 5. Pour obtenir un arbre sur X minimum pour d, on peut utiliser par exemple, l'algorithme "glouton" de Kruskal (1956). On prend les éléments de P dans un ordre non décroissant selon d :

$d(\alpha\zeta) = 3$; $d(\alpha\gamma) = 6$; $d(\delta\eta) = 6$; $d(\alpha\zeta) = 7$, mais la paire $\alpha\zeta$ ne peut être retenue : avec les précédentes, elle formerait le cycle $\{\alpha\zeta, \zeta\gamma, \gamma\alpha\}$, alors que c'est un arbre que l'on cherche.

$d(\beta\eta) = 8$; $d(\gamma\eta) = 8$; $d(\varepsilon\zeta) = 9$;

on a retenu un ensemble A de $n-1 = 6$ paires ne contenant pas de cycle. C'est un arbre minimum (ici, le seul en fait) pour d. Sa longueur est $\ell(A) = 40$. Cet arbre est représenté par la figure 10.

β	γ	δ	ε	ζ	η	
12	6	12	11	7	11	α
	11	10	13	12	8	β
		11	10	3	8	γ
			14	12	6	δ
				9	12	ε
					10	ζ

Table 5. La dissimilarité $d = \sum_{i=1}^{7} d_i$

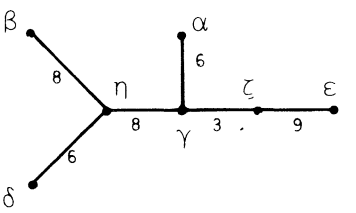

Figure 10. L'arbre A minimum pour D.

On vérifie que l'arbre A possède simultanément les deux propriétés suivantes :

(1) $\ell(A) = \sum_{i=1}^{7} \ell_i(A_i) = 4 + 4 + 3 + 3 + 3 + 2 + 21 = 40$

(2) A est minimum pour chacune des préordonnances R_i , pour $1 \leq i \leq 7$.

La figure 11 illustre la propriété (2) : les classes de la famille $\mathcal{F} = \{F_1, F_2, F_3, F_4\}$ (a) , de la partition Π (b), de la hiérarchie H (c) et du préordre T (d) déterminent des parties connexes de l'arbre A . De plus, pour toute arête $xy \in A$, x et y sont soit dans la même classe du préordre T , soit dans deux classes consécutives.

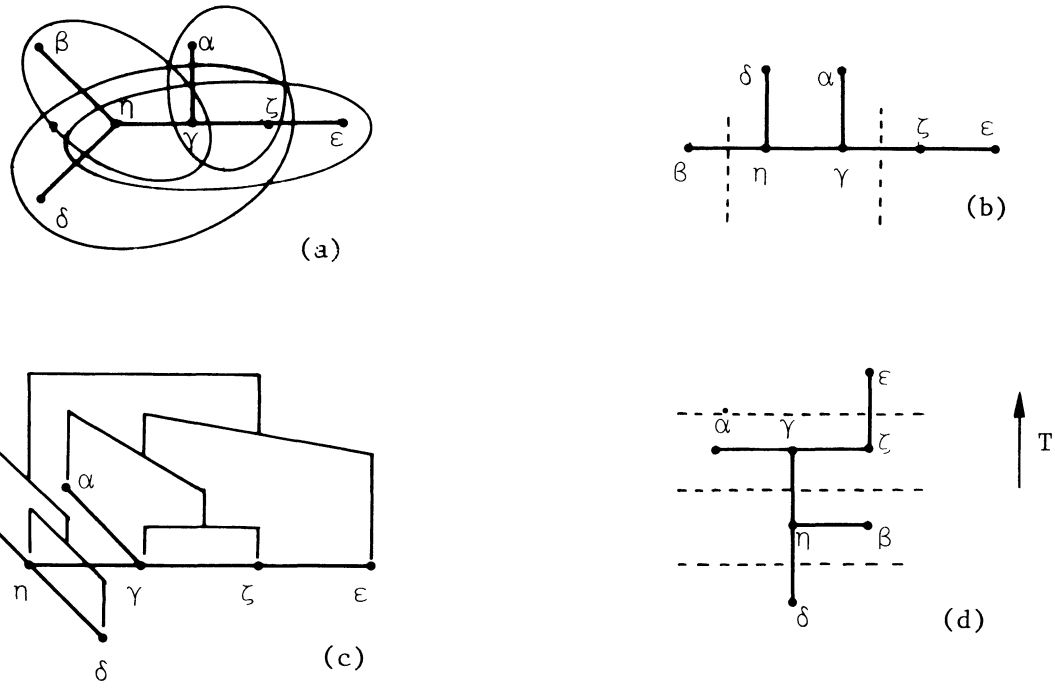

Figure 11. L'arbre A est compatible avec chacune des structures $\mathcal{F}$, T , Π et H sur X .

Evidemment, des deux propriétés (1) et (2) ci-dessus, (1) est, de beaucoup, celle dont la vérification est la plus aisée. Nous allons montrer qu'elles sont en fait équivalentes et ce indépendamment des dissimilarités particulières d_i , compatibles avec les préordonnances R_i , choisies.

3. RESULTATS SUR LES ARBRES MINIMUMS COMMUNS

3.1 Le résultat principal

3.11 Une inégalité. Soit v un entier positif et soient v dissimilarités $d_1, d_2, \ldots, d_v$ sur X . Pour $i = 1, \ldots, v$, on considère un arbre A_i minimum

pour d_i et sa longueur $\ell_i(A_i) = \sum_{a\in A_i} d_i(a)$.

Soit $d = \sum_{i=1}^{v} d_i$ la dissimilarité somme des précédentes. La longueur pour d d'un arbre quelconque A sur X est

$$\ell(A) = \sum_{a\in A} d(a) = \sum_{a\in A} \sum_{i=1}^{v} d_i(a) = \sum_{i=1}^{v} \sum_{a\in A} d_i(a) = \sum_{i=1}^{v} \ell_i(A)$$

PROPOSITION 3.1 Avec les définitions et notations ci-dessus, on a, pour tout arbre A sur X , l'inégalité :

$$\ell(A) \geq \sum_{i=1}^{v} \ell_i(A_i)$$

Preuve. Pour chaque $i = 1,\ldots,v$, on a, puisque A_i est un arbre minimum pour d_i , et que A n'en est pas un en général, $\ell_i(A) \geq \ell_i(A_i)$, d'où

$$\sum_{i=1}^{v} \ell_i(A) \geq \sum_{i=1}^{v} \ell_i(A_i) \ . \quad \square$$

Dans le cas où l'inégalité de la proposition 3.1 devient une égalité, nous dirons que l'arbre A vérifie l'égalité (AMC) pour les dissimilarités $d_1,\ldots,d_v$.

(AMC) $\sum_{a\in A} \sum_{i=1}^{v} d_i(a) = \sum_{i=1}^{v} \sum_{a\in A_i} d_i(a)$, où, pour $i = 1,\ldots,v$, A_i est un arbre minimum pour d_i .

Un arbre A vérifiant l'égalité (AMC) est un arbre minimum pour la dissimilarité $d = \Sigma d_i$. Sinon, l'inégalité de la proposition 3.1 serait contredite pour tout arbre minimum pour d .

3.12 Caractérisation des arbres minimums communs

THEOREME 3.2. Soient $R_1,R_2,\ldots,R_v$ des préordonnances sur X et $A_1,A_2,\ldots,A_v$ des arbres sur X tels que, pour $i = 1,\ldots,v$, A_i est un arbre minimum pour R_i . Les quatre propriétés suivantes sont équivalentes pour un arbre A sur X .

(1) Pour tout $i = 1,\ldots,v$, A est un arbre minimum pour R_i .

(2) A est un arbre minimum pour la préordonnance $R_1 \cap R_2 \cap \ldots \cap R_v$.

(3) Il existe des dissimilarités $d_1,\ldots,d_v$, compatibles avec $R_1,\ldots,R_v$ respectivement, telles que A vérifie l'égalité (AMC) pour $d_1,\ldots,d_v$.
(4) Pour toutes dissimilarités $d_1,\ldots,d_v$, compatibles avec $R_1,\ldots,R_v$ respectivement, A vérifie l'égalité (AMC) pour $d_1,\ldots,d_v$.

Preuve. (1) $\Leftrightarrow$ (2) est une conséquence de l'équivalence (1) $\Leftrightarrow$ (3) de la proposition 1.1. Ici (1) $\Leftrightarrow (\forall\, i = 1,\ldots,v)\ \Delta_A \subset R_i \Leftrightarrow \Delta_A \subset \bigcap_{i=1}^{v} R_i \Leftrightarrow$ (2) .
(1) $\Rightarrow$ (4) provient de l'implication (1) $\Rightarrow$ (2) de cette même proposition 1.1. D'après elle, si (1) est vraie, A est un arbre minimum pour chacune des dissimilarités d_i , et vérifie donc l'égalité (AMC) en prenant $A_i = A$ pour tout i .
(4) $\Rightarrow$ (3) est évident.
(3) $\Rightarrow$ (1) résulte de la proposition 1.2. Si A vérifie l'égalité (AMC) , on a $\ell_i(A) = \ell_i(A_i)$ pour tout $i = 1,\ldots,v$ et donc A est minimum pour chacune des dissimilarités d_i . Alors, A est encore minimum pour chacune des préordonnances R_i , puisqu'on a fait l'hypothèse qu'elles admettent toutes au moins un arbre minimum. □

Remarques.
1. La vérification de l'égalité AMC lorsque l'on n'est pas assuré que chacune des préordonnances R_i admet un arbre minimum permet seulement d'affirmer que A est un arbre minimal commun, c'est-à-dire, pour tout $i = 1,2,\ldots,v$, un élément minimal de $\mathcal{A}$ préordonné par DR_i .
2. Le théorème 5.1 s'étend directement à la caractérisation de bases minimums communes à plusieurs préordres donnés sur un même matroïde (cf. Flament et Leclerc 1983). On ne parlera plus alors de dissimilarités, mais de pondérations compatibles.

3.13 Obtention d'arbres minimums communs. On tire du théorème 3.2 et des propositions 1.1 et 1.2 plusieurs façons d'établir s'il existe ou non un arbre minimum commun à plusieurs préordonnances $R_1,\ldots,R_v$ et, si oui, d'obtenir un tel arbre. La procédure la plus directe est la procédure (A) suivante :

A1. Calcul de la préordonnance $R = R_1 \cap \ldots \cap R_v$.
A2. Détermination d'une préordonnance totale T compatible avec R .
A3. Détermination d'un arbre A minimum pour T .
A4. Détermination de la relation d'échangeabilité Δ_A .

A5. Si l'on a $\Delta_A \subset R$, alors A est un arbre minimum commun. Sinon, il n'y a pas de tel arbre.

Justification : Un arbre minimum commun est minimum pour R , donc pour T (théorème 3.2 et proposition 1.1). Un tel arbre minimum pour T est minimal pour R (proposition 1.2), donc minimum pour R si un tel arbre existe.

En substituant à chacune des R_i une dissimilarité compatible d_i , on a la procédure (B) suivante, où l'on fait l'économie de l'étape A_2 , celle-ci et A1 étant remplacées par B1.

B1. Calcul de la dissimilarité $d = \sum_{i=1}^{v} d_i$.

B2. Détermination d'un arbre A minimum pour d .

B3 et B4 comme A4 et A5.

Justification : analogue à celle de la procédure (A), dès que l'on remarque que la préordonnance associée à d est compatible avec R .

La procédure (C) suivante remplace la détermination de Δ_A et les vérifications $\Delta_A \subset R_i$ par des opérations plus simples portant sur des valeurs numériques. Elle suppose, outre l'introduction des d_i , la connaissance des longueurs $\ell_i(A_i)$. On a vu que dans bien des cas (préordonnances issues de préordres totaux, ou de partitions, ou de parties, ...) celle-ci ne nécessite pas la détermination des arbres A_i .

C1 et C2 comme B1 et B2.

C3. Calcul de la longueur $\ell(A)$ de l'arbre A pour la dissimilarité d .

C4. Si $\ell(A) = \sum_{i=1}^{v} \ell_i(A_i)$, l'arbre A est un arbre minimum commun. Sinon, il n'y a pas de tel arbre.

Justification : par le théorème 3.2.

3.2. Familles arborées.

Soit $\mathcal{F} = \{F_1,\ldots,F_v\}$ une famille de parties de X . Elle est dite *arborée* ou *rigide sur un arbre* si et seulement si il existe un arbre A sur X tel que, pour tout $F \in \mathcal{F}$, F induit une partie connexe de A , c'est-à-dire que le graphe $(F, P_F \cap A)$ est connexe.

Dans l'exemple de la section II ci-dessus, la famille $\mathcal{F} = \{F_1,F_2,F_3,F_4\}$ est arborée : la figure 11(a) illustre le fait qu'elle

est rigide sur l'arbre A .

Si de plus, elle est rigide sur un arbre qui est une chaîne $\{x_1x_2, x_2x_3, \ldots, x_{n-1}x_n\}$, la famille $\mathcal{F}$ est une *famille d'intervalles* (de l'ordre $x_1, x_2, \ldots, x_n$). Enfin, une famille d'intervalles totalement ordonnée par l'inclusion est une *échelle de Guttmann*.

Posons $n_i = |F_i|$, pour $i = 1,\ldots,v$ et associons à $\mathcal{F}$ la préordonnance $R(\mathcal{F})$ et la dissimilarité $d = d(\mathcal{F})$ définies par :
$(\forall p,p' \in P)\quad (p,p') \in R(\mathcal{F}) \Leftrightarrow [(\forall i = 1,\ldots,v)\quad p' \subseteq F_i \Rightarrow p \subseteq F_i]$
$(\forall p \in P)\quad d(p) = |\{F \in \mathcal{F} / p \not\subseteq F\}|$.

On note ℓ la longueur d'un arbre minimum pour d , valué par d . A partir du théorème 3.2, on obtient les caractérisations suivantes des familles arborées :

COROLLAIRE 3.3. Les trois propriétés suivantes sont équivalentes pour une famille $\mathcal{F} = \{F_1,\ldots,F_v\}$ de parties de X .
(1) $\mathcal{F}$ est une famille arborée.
(2) La préordonnance $R(\mathcal{F})$ admet un arbre minimum.
(3) $\ell = \sum_{i=1}^{v} (n-n_i)$.

Preuve. Considérons les préordonnances R_i et les dissimilarités d_i associées, comme en 1.31, aux parties F_i de X . Dire que F_i induit une partie connexe d'un arbre A équivaut à dire que A est un arbre minimum pour R_i . Du fait que $R(\mathcal{F}) = \bigcap_{i=1}^{v} R_i$ et $d = \sum_{i=1}^{v} d_i$, le corollaire est une conséquence immédiate du théorème 3.2. □

La caractérisation (2) est due à Flament (1975,1978). La caractérisation (3) est implicite dans Acharya et Las Vergnas (1982) et est explicitement donnée par Leclerc (1984). Elles sont données sous forme duale dans ces références. Flament appelle *trace* de $\mathcal{F}$ la préordonnance duale de $R(\mathcal{F})$. La famille $\mathcal{F}$ est arborée si et seulement si sa trace admet un arbre maximum. Leclerc considère la similarité $s = v(n-1)-d$ et la longueur ℓ' d'un arbre maximum pour s , valué par s , la condition (3) devenant :

$$\sum_{i=1}^{v} (n_i-1) = \ell' .$$

Acharya et Las Vergnas appellent le nombre $-\ell' + \sum_{i=1}^{v} (n_i - 1)$ *nombre cyclomatique* de la famille $\mathcal{G}$ de parties de $\{1,\ldots,v\}$, "duale" de $\mathcal{F}$, définie par :

$$\mathcal{G} = \{G \subseteq \{1,\ldots,v\} / (\exists x \in X)(\forall i \in G)\ x \in F_i\} .$$

Leur caractérisation des familles pour lesquelles ce nombre est nul, rapprochée d'une caractérisation combinatoire des familles arborées due à Duchet (1978) et Flament (1978) correspond bien à ce que $\mathcal{F}$ est arborée.

3.3. La recherche d'ordres compatibles

On a évoqué, au début de la section 2, le cas particulièrement intéressant où il y a un arbre minimum commun $C = x_1x_2, x_2x_3, \ldots, x_{n-1}x_n$ qui est une chaîne. Alors on a un ordre $x_1x_2\ldots x_n$ sur X, une *sériation* de X, qui est compatible avec toutes les préordonnances R_i, donc avec les données dont elles sont éventuellement issues. Dans le cas où celles-ci sont des classifications (partitions, hiérarchies, pyramides) le problème de la reconnaissance et de l'obtention d'un tel ordre a été abordé par Diday (1982,1986).

Les résultats du paragraphe précédent peuvent parfois permettre de conclure sur ce problème. Ainsi la dissimilarité d obtenue en 2.3 admet un seul arbre minimum A et celui-ci n'est pas une chaîne.

Le cas de la dissimilarité d' du par. 2.21 (table 1) est plus complexe. Elle admet plusieurs arbres minimums (arbres minimums communs à R_1, R_2, R_3 et R_4). On peut voir que ce sont tous les arbres du graphe de la figure 12 qui contiennent les paires $\gamma\eta$ et $\gamma\zeta$. Il y en a 36, mais, à nouveau, aucun n'est une chaîne.

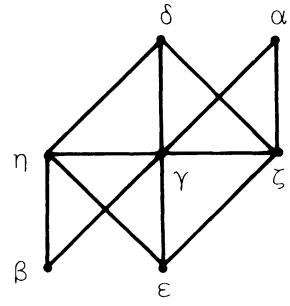

Figure 12.

Dans certains cas, on peut ainsi déterminer tous les arbres minimums communs et les examiner pour voir s'il y a une chaîne. On montre qu'on les obtient en déterminant tous les arbres de certains graphes, obtenus par "contractions" ou suppressions de certaines arêtes (Leclerc 1981b). On peut donc

utiliser un algorithme produisant tous les arbres d'un graphe, comme celui de Read et Tarjan (1975), pourvu qu'il n'y ait pas un trop grand nombre d'arbres minimums communs. Sinon, une approche plus directe est nécessaire.

Il existe un algorithme en temps linéaire en n pour reconnaître si une famille de parties $\mathcal{F}$ est une famille d'intervalles (Booth et Leuker 1976, cf. Golumbic 1980), et produire un ordre sur X correspondant. On peut chercher à étendre un tel algorithme à d'autres types de données, en notant toutefois que le cas d'une préordonnance totale à deux classes quelconques équivaut à celui de la recherche d'une chaîne hamiltonienne dans un graphe quelconque, un problème NP-complet (cf. Garey et Johnson 1979).

Enfin, on peut élargir le problème en cherchant un arbre minimum commun aussi proche que possible d'une chaîne, par exemple ayant un nombre minimum de sommets terminaux, ou ayant une chaîne de longueur maximum.

3.4. Cas où il n'y a pas d'arbre minimum commun

Les procédures A , B et C (du paragraphe 3.13 ci-dessus) donnent toujours un arbre A (aux étapes A3, B2 et C2 respectivement). Il peut ensuite apparaître que cet arbre n'est pas un arbre minimum commun. Dans ce cas, on a noté que c'est toujours un arbre minimal pour R , c'est-à-dire qu'il ne peut y avoir un autre arbre A' tel que $(A',A) \in DR_i$ pour tout i et $(A,A') \notin DR_j$ pour un certain $j \in \{1,\ldots,v\}$.

De plus, lorsqu'il est issu des procédures B et C, l'arbre A obtenu minimise la quantité $\ell(A) - \sum_{i=1}^{v} \ell_i(a_i)$. L'intérêt de l'arbre A est donc lié à l'interprétation de cette quantité. Ainsi, dans le cas d'une famille de parties de X , reprenons les notations du paragraphe 3.2 : la quantité $\ell - \Sigma(n_i-n)$ est le nombre de ruptures de connexité des classes F_i , $i = 1,\ldots,v$, sur l'arbre A . Celui-ci a la propriété d'être, parmi tous les arbres sur X , celui ou un de ceux qui minimisent ce nombre.

Par contre, dans le cas général, la quantité $\ell(A) - \sum_{i=1}^{v} \ell_i(A_i)$ dépend des dissimilarités d_i particulières choisies, et il en est de même de l'arbre A obtenu.

BIBLIOGRAPHIE

ACHARYA B.D., LAS VERGNAS M., "Hypergraphs with cyclomatic number zero, triangulated graphs, and an inequality", *J. Combinatorial Theory B*, 33 (1982), 52-56.

BARTHELEMY J.P., LECLERC B., MONJARDET B., "Ensembles ordonnés et taxonomie mathématique", in : M. POUZET, D. RICHARD, eds. *Orders : descriptions and roles*, *Annals of Discrete Mathematics* 23, Amsterdam, North-Holland, 1984a.

BARTHELEMY J.P., LECLERC B., MONJARDET B., "Quelques aspects du consensus en classification", in : E. DIDAY et al. eds. *Data Analysis and Informatics III*, Amsterdam, North-Holland, 1984b.

BARTHELEMY J.P., LECLERC B., MONJARDET B., "On the use of ordered sets in problems of comparison and consensus of classification",*J. of Classification* 3, (1986), 185-222.

BATBEDAT A., *Comment reconnaître une prépyramide*, Cahier S, UER de Mathématiques, Montpellier, Université des Sciences et Techniques du Languedoc, 1986.

BENZECRI J.P., "Description mathématique des classifications"(1967) , in : *L'analyse des données I. La taxinomie*, Paris, Dunod, 1973.

BERTRAND P., *Etude de la représentation pyramidale*, thèse de 3ème cycle, Université de Paris-Dauphine et INRIA Rocquencourt, 1986.

BOOTH K.S., LEUKER G.S., "Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms", *J. Comput. Syst. Sci.*, 13 (1976), 335-379.

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

BRUALDI R.A., "Comments on bases in dependance structures",*Bull. Austral. Math. Soc.*, 2 (1969), 161-167.

CAYLEY A., "On the theory of the analytic forms called trees", *Phil. Magazine* XIII (1857), 172-176, *Collected mathematical papers*, vol. 3, p. 242, Cambridge (RU), Cambridge University Press.

DIDAY E., *Croisements, ordres et ultramétriques : application à la recherche de consensus*, Rapport de recherches n° 144, Rocquencourt, INRIA, 1982.

DIDAY E., "Croisements, ordres et ultramétriques", *Math. Sci. hum.*, 83 (1983), 31-54.

DIDAY E., *Une représentation visuelle des classes empiétantes : les pyramides*, Rapport de recherches n° 291, Rocquencourt, INRIA, 1984.

DIDAY E., *Compatibility and consensus in numerical taxonomy*, Rocquencourt, INRIA, 1986.

DOIGNON J.P., MONJARDET B., ROUBENS M., VINCKE Ph., "Biorders families, valued relations and preference modelling, *J. of Math. Psychology*, 30, (1986) à paraître.

DUCHET P., "Propriété de Helly et problèmes de représentation, in : *Problèmes combinatoires et théorie des graphes*, Paris, Editions du CNRS, 1978.

FLAMENT C., "Arêtes maximales des cocycles d'un graphe préordonné",*Math. Sci. hum* ., 51, (1975), 5-12.

FLAMENT C., "Hypergraphes arborés", *Discrete Math.*, 21, (1978), 223-227.

FLAMENT C., LECLERC B., "Arbres minimaux d'un graphe préordonné", *Discrete Math.*, 46, (1983), 159-171.

GAREY M.R., JOHNSON D.S., *Computers and Intractability*, San Francisco, Freeman, 1979.

GIRAUDET M., *Formules, chaînes et ultramétriques*, non publié, 1982.

GOLUMBIC M.C., *Algorithmic graph theory and perfect graphs*, New York, Academic Press, 1980.

GOWER J.C., ROSS G.J.S., "Minimum spanning tree and single linkage cluster analysis", *Applied Statistics*, 18, (1969), 54-64.

HARTIGAN J.A., *Clustering algorithms*, New York, Wiley, 1975.

HUBERT L., "Some applications of graph theory and related non-metric techniques to problems of approximate seriation", *British J. of Math. and Statist. Psychology*, 27, (1974), 133-153.

HUBERT L., "Data analysis implications of some concepts related to the cuts of a graph", *J. of Math. Psychology*, 15, (1977), 199-208.

KRUSKAL J.B., "On the shortest spanning tree of a graph and the traveling salesman problem", *Proc. Amer. Math. Soc.*, 7, (1956), 48-50.

LECLERC B., "An application of combinatorial theory to hierarchical classification", in : *Recent Developments in Statistics*, J.R. BARRA et al. eds, Amsterdam, North-Holland, 1977, 783-786.

LECLERC B., "Semi-modularité des treillis d'ultramétriques", *C.R. Acad. Sci. Paris*, A-288, (1979), 575-577.

LECLERC B., "Description combinatoire des ultramétriques", *Math. Sci. hum.*, 73, (1981a), 5-37.

LECLERC B., "Sur le nombre d'arbres minimums d'une ultramétrique", non publié, 1981b.

LECLERC B., *Comment reconnaître un hypergraphe arboré*, rapport CMS-P.009, Paris, CAMS, 1984.

LECLERC B., "Les hiérarchies de parties et leur demi-treillis", *Math. Sci. hum.*, 89, (1985), 5-34.

LECLERC B., "Caractérisation, construction et dénombrement des ultramétriques supérieures minimales", *Statistique et Analyse des données*, à paraître (1986).

MOON J.W., "Enumerating labelled trees", in : F. HARARY ed. *Graph Theory and Theoretical Physics*, London, Academic Press, 1967.

READ R.C., TARJAN R.E., "Bounds on backtrack algorithms for listing cycles, paths, and spanning trees", *Networks* 5, (1975), 237-252.

ROSENSTIEHL P., "L'arbre minimum d'un graphe", in : P. ROSENSTIEHL, ed. *Théorie des graphes (Rome 1966)*, Paris, Dunod, 1967.