Une representation visuelle des classes empietantes :Les pyramides

Edwin Diday

INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P 105
78153 Le Chesnay Cedex
France
Tel 954 90 20

Rapports de Recherche

N° 291

UNE REPRÉSENTATION VISUELLE DES CLASSES EMPIETANTES : LES PYRAMIDES

Edwin DIDAY

Avril 1984

UNE REPRESENTATION VISUELLE DES CLASSES EMPIETANTES : LES PYRAMIDES.

E. DIDAY

Résumé

Le problème de la recherche de classes empiétantes plutôt que de partitions se pose fréquemment dans la pratique. Les pyramides permettent une représentation visuelle de telles classes et constituent une extension naturelle des hiérarchies; elles sont plus riches que les hiérarchies du point de vue des informations fournies en faisant apparaître des recouvrements emboîtés au lieu de partitions ;elles induisent un indice de dissimilarité particulier appelé "indice pyramidal" plus proche des données initiales qu'une ultramétrique. L'ensemble de ces indices (qui font intervenir la notion d'ordre sur les singletons) est en bijection avec l'ensemble des pyramides et contient l'ensemble des ultramétriques. Des procédés constructifs pour la représentation d'une pyramide sont donnés ; on étudie le problème du choix optimal d'une pyramide et plus particulièrement la notion de sur-dominante et sous-dominante pyramidale. On montre comment enrichir une hiérarchie en la "pyramidisant" ou comment "hiérarchiser" partiellement une pyramide. On propose enfin d'autres formes de représentations d'un indice pyramidal (planes, polygonales, curvilignes, arbres "épais" ou "guirlandes").

Mots clés : Analyse des données, reconnaissance des formes, classification automatique, hiérarchies, ultramétriques, ordres.

Abstract

The problem of overlapping clusters is frequently encountered in practice. Pyramids allow a visual representation of such clusters and constitute a natural extension of hierarchies ; they give more information than hierarchies. They induce a special dissimilarity index called "pyramidal index" closer to the data than ultrametrics. The set of indices (which use the notion of order on the single objects) is in bijection with the set of pyramids and contain the set of ultrametrics. Contructive procedures for the representation of a pyramid are given ; the problem of optimal choice of a pyramid is discussed and more specifically the concept of "dominant" and "sub-dominant" of pyramids. It is shown how hierarchies may be enriched by "pyramization" and how it is possible to "hierarchize" a pyramid. Among other results, we show that it is possible to obtain a computer program which gives a representation which is more or less a hierarchy or a pyramid depending on a given threshold. We give different other types of representations of a pyramidal index (planer, polygonal, "thick tree" and "guirlande").

Key-words : Data Analysis, Pattern recognition, clustering, hierarchies, ultrametrics, orders.

INTRODUCTION.

Confronté à la réalité multimensionnelle on a cherché de tout temps des représentations visuelles permettant d'appréhender plus simplement cette réalité ; la classification hiérarchique est un mode de représentation visuelle utilisé de longue date ; depuis Aristote et son arbre de vie, jusqu'aux travaux récents de Sokal et Sneath (1973) et de Benzecri et collaborateurs (1973), en passant par Adanson (1757) et son algorithme (de construction hiérarchique) pour agglomérer des plantes, les hiérarchies ont constitué un outil utile pour la représentation de données multidimensionnelles.

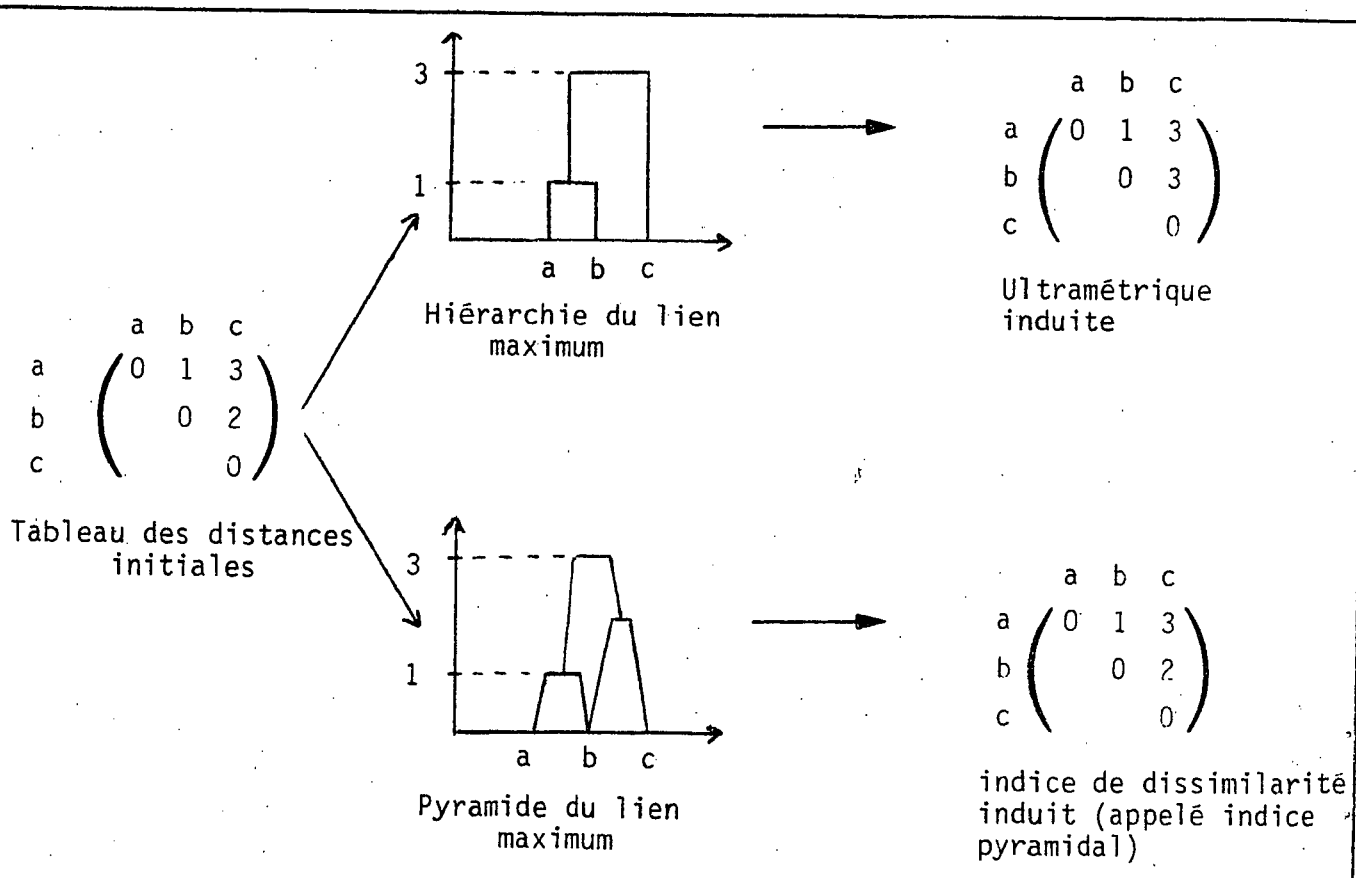
Etant donné l'importance et la variété des retombées pratiques du au développement rapide de l'outil informatique, de nombreux auteurs se sont intéressés dans ces dernières années à la recherche de classes recouvrantes, quatre types d'approches ont été principalement étudiés : 1) utilisation de la théorie des graphes avec Hubert (1974) par exemple* 2) par l'approche probabiliste : décomposition de mélanges de lois (Hartigan (1975 et 1977), Diday et coll.(1979)). 3) l'optimisation d'un critère par l'approche "multidimensionnal Scaling" avec Shepard et Arabie (1978) et Arabie et Caroll (1980) ou par l'approche nuées dynamiques (Diday (1982) : en utilisant des noyaux qui forment un recouvrement ou les formes faibles). 4) l'approche hiérarchique avec Jardine et Sibson (1971) et des améliorations du type de celles de Rohlf (1975) (il s'agit en fait de techniques permettant d'améliorer les hiérarchies obtenues avec l'indice du saut minimum qui sont sujettes à l'effet de chaînes).

Une hiérarchie est formée d'une suite de partitions emboîtées ; nous montrons dans ce travail que l'ensemble des hiérarchies peut être plongé dans un ensemble plus vaste donnant des représentations (appelées pyramides), plus riches d'informations, plus proches des données initiales, faisant apparaître des recouvrements emboîtés au lieu de partitions.

On sait qu'il existe une bijection entre l'ensemble des hiérarchies indicées et l'ensemble des ultramétriques (voir [8] par exemple). L'intérêt de ce résultat vient du fait que la recherche d'une hiérarchie ayant une bonne adéquation aux données initiales peut être remplacée par la recherche d'une distance

* Voir aussi B. Monjardet (1981).

ultramétrique. Ce résultat classique a été étendu de la façon suivante : on a montré qu'il existe un ordre qui rende "Robinson" une matrice de distance ultramétrique ; cette matrice doit tenir compte de l'inégalité ultramétrique ; en supprimant les contraintes dues à cette inégalité, on débouche sur une classe d'indices de dissimilarité particuliers appelés "indices pyramidaux", qui contient l'ensemble des ultramétriques (voir [12]) ; à ces indices, on peut associer une nouvelle forme de représentation visuelle qui contient la représentation hiérarchique et donne des informations plus riches à l'utilisateur ; l'objet essentiel de ce travail est l'étude de cette nouvelle forme de représentation qui a l'allure de "pyramides". (voir la figure 1).



On voit sur cet exemple que :

1) - la hiérarchie est incluse dans la "pyramide"(chaque palier de la hiérarchie est un palier de la pyramide),

2) - la représentation pyramidale est plus fidèle au tableau de distance que la représentation hiérarchique.

Figure 1.

On fait d'abord quelques rappels concernant les différentes formes de compatibilité entre un ordre et une distance en s'appuyant essentiellement sur les résultats obtenus dans [12] ; il en ressort, que les représentations hiérarchiques expriment de façon visuelle la compatibilité entre une ultramétrique et un ordre et que les pyramides en sont une extension naturelle puisqu'elles permettent la représentation visuelle de la compatibilité entre un indice pyramidal et un ordre : en effet, on montre en 5.2. qu'une ultramétrique est un cas particulier d'indice pyramidal ; le tableau 1 montre clairement que les pyramides permettent de combler des cases qui seraient restées vides.

En 3) on donne la définition axiomatique de la notion de pyramide ; il résulte de cette définition que les hiérarchies sont des pyramides particulières : on démontre notamment en 4) que contrairement aux paliers d'une hiérarchie qui ne peuvent avoir qu'un seul prédécesseur, les paliers d'une pyramide peuvent en avoir deux : en 4) on donne également un algorithme permettant de construire un ordre "compatible" avec une pyramide si l'on connait uniquement les paliers qui la constituent.

La connaissance du paragraphe 4) n'est pas indispensable pour la compréhension de la suite de l'article.

En 5) on définit la notion d'indiçage d'une pyramide qui permet d'associer une hauteur à chaque palier d'une pyramide afin de la visualiser.

En 6) on propose un algorithme de classification ascendante pyramidal (CAP) ainsi qu'un algorithme accéléré basé sur la notion de voisin réciproque. Comme pour les hiérarchies on peut utiliser une formule de type Lance et Williams et faire des épurations pour éliminer des paliers inutiles dans la visualisation d'une pyramide. L'algorithme de CAP induit un indice pyramidal d'où l'on peut déduire un ordre "compatible" avec la pyramide construite grâce à un algorithme donné en 6.5.

Nous donnons en 7) un résultat fondamental : il existe une bijection entre l'ensemble des indices pyramidaux et l'ensemble des pyramides indicées "au sens large".

Pour définir cette bijection on utilise deux applications : ϕ et ψ; l'application ϕ permet d'associer simplement à une pyramide un indice pyramidal : l'application ψ associe une pyramide à un indice pyramidal ; elle est plus compliquée c'est pourquoi nous donnons en 7.2 un exemple permettant de l'illustrer.

On montre en 7.3 qu'il y a équivalence entre les notions de croisement et de compatibilité entre un ordre et une pyramide et entre un ordre et un indice pyramidal ; toutes ces notions peuvent s'exprimer sous forme matricielle en terme de matrice Robinson.

L'existence d'une bijection entre les pyramides et les indices pyramidaux, permet de poser le problème de la recherche d'une pyramide en terme d'optimisation : nous abordons ces problèmes en 8) en nous interessant surtout à la recherche de la sur-dominante et sous-dominante pyramidale : contrairement à ce qui se passe pour les ultramétriques la sous-dominante pyramidale n'est pas atteinte, autrement dit ce n'est pas nécessairement un indice pyramidal (on sait que la sous-dominante ultramétrique est une ultramétrique (voir [8])). On montre entre autres que l'algorithme de CAP muni de l'indice d'agragation du saut maximum construit à partir d'un indice pyramidal une pyramide qui induit par ϕ cet indice.

Une pyramide peut devenir difficile à interpréter, quand elle a beaucoup de paliers (elle peut en avoir jusqu'à $\frac{n(n-1)}{2}$ (si n est le nombre d'individus) alors que ce nombre est réduit à n-1 dans le cas d'une hiérarchie) ; différentes techniques peuvent être imaginées pour réduire le nombre de paliers (entre autres celles qui sont utilisées en classification hiérarchique) ; profitant du fait que les pyramides constituent une extension des hiérarchies, nous proposons une technique de "hiérarchisation" qui permet de réduire le nombre de paliers tout en déformant le moins possible l'indice pyramidal associé. Dans le cas où l'on désire enrichir une hiérarchie déjà construite pour faire apparaître des classes recouvrantes, nous proposons une technique de "pyramidisation" d'une hiérarchie.

Dans la dernière partie (§ 10), on s'intéresse à d'autres formes de représentation d'un indice pyramidal : on remarque d'abord que comme pour tout indice

de dissimilarité on peut associer à un indice pyramidal une distance respectant le même ordre sur les couples. Si cette distance est euclidienne, il est facile de voir que l'on peut représenter les n individus par n points du plan respectant 2n-3 distances ; il en résulte qu'il est possible de représenter dans le plan au moins les 2n-3 plus petites distances pyramidales autrement dit les distances respectant les hauteurs des 2n-3 plus bas paliers de la pyramide associée.

En utilisant la propriété de compatibilité entre un ordre et un indice pyramidal on s'intéresse à la représentation d'un tel indice (et donc de la pyramide associée) par un polygone, dont les côtés et diagonales satisfont des propriétés particulières.

Par analogie à la représentation de hiérarchies sous forme d'arbres, il est tentant de chercher ce qui correspond à des arbres dans le cas de pyramides; nous proposons la notion "d'arbres épais" ou "guirlandes" (pour simplifier on peut dire que cela revient à remplacer chaque sommet d'un arbre par un polygone).

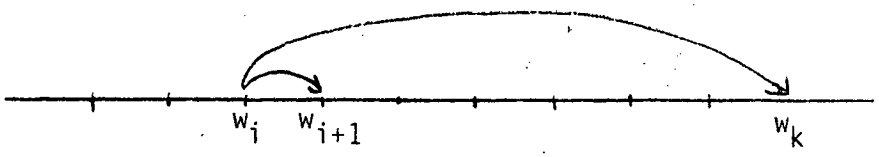
On sait qu'il existe toujours un ordre compatible avec un indice pyramidal d'où l'idée d'utiliser cet ordre pour représenter cet indice par une courbe telle que la position des points qui représentent les individus sur cette courbe ordonnés selon cet ordre ,respectent au mieux l'indice.

2. LA REPRESENTATION VISUELLE DES DIFFERENTS TYPES DE COMPATIBILITE ENTRE ORDRES ET INDICES DE DISSIMILARITE

2.1. *Quelques rappels concernant différents types de compatibilité.*

(Pour plus de détails concernant ce paragraphe on pourra se reporter à [12].)

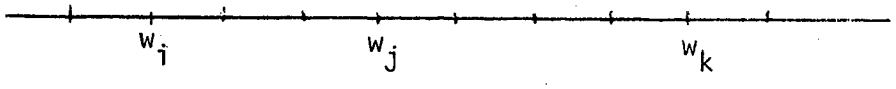
. On dit qu'un indice de dissimilarité d et un ordre Θ sont faiblement compatibles si et seulement si pour tout triplet ordonné selon Θ et ayant deux éléments consécutifs, la distance des éléments consécutifs est inférieure à la distance des éléments extrêmes (voir figure 1)



$$d(w_i, w_{i+1}) \leq d(w_i, w_k)$$

Figure 1 : compatibilité faible.

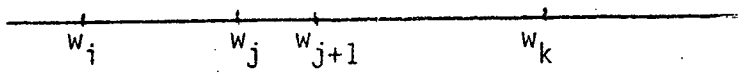
. On dit que Θ et d sont compatibles si et seulement si pour tout triplet ordonné selon Θ, la distance entre les éléments extrêmes est supérieure à la distance entre une extrémité quelconque et l'élement intermédiaire.



$$d(w_i, w_k) \geq \text{Max}(d(w_i, w_j),\ d(w_j, w_k))$$

Figure 2 : compatibilité.

. On dit que Θ et d sont semi-compatibles si et seulement si pour tout quadruplet ordonné selon Θ dont les éléments intermédiaires sont consécutifs, la distance des éléments extrêmes est supérieure à la distance des éléments consécutifs.



$$d(w_i, w_k) \geq d(w_j, w_{j+1})$$

Figure 3 : semi-compatibilité.

2.2. *Aspects matriciels : matrices de Robinson et matrices SDR et SDD*

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 4).

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 4.

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 4, 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 4, 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 4).

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 4).

- Dans toute la suite on note $M(d,\Theta)$ la matrice de dissimilarité associée à d et dont les lignes et colonnes sont rangées selon l'ordre Θ.

Définition d'une matrice ultramétrique

Dans [12] on démontre que si δ est une ultramétrique, il existe toujours un ordre Θ (non unique) tel que δ et Θ soient faiblement compatibles ; la matrice $M(\delta,\Theta)$ est alors dite "ultramétrique" ; on montre qu'elle est Robinson ; elle n'est pas seulement Robinson car elle satisfait de plus l'inégalité ultramétrique :

$$\forall(i,j,k) : \delta(w_i,w_k) \leq \text{Max}(\delta(w_i,w_j), \delta(w_j,w_k)).$$

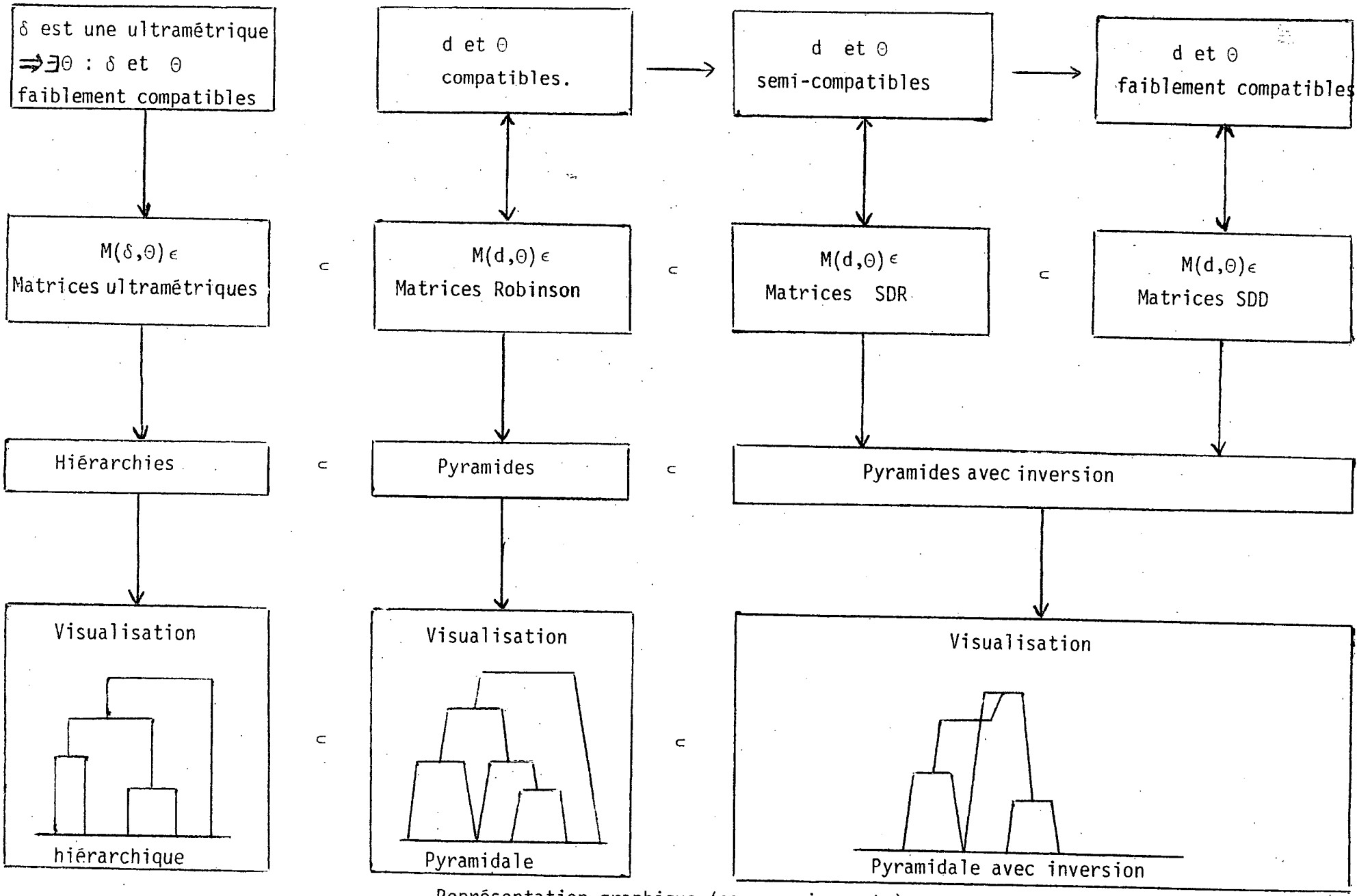
Le lecteur désireux de mieux situer les différents types de compatibilité et leurs propriétés pourra se reporter à l'annexe.

2.3. *La représentation visuelle des différents types de compatibilité.*

Remarquant que la représentation hiérarchique exprime en fait la compatibilité entre un ordre et une ultramétrique en donnant une image visuelle au contenu d'une matrice ultramétrique, il vient de façon naturelle l'idée de donner une représentation visuelle aux autres formes de compatibilité et plus précisement aux matrices Robinson, SDR et SDD.

Ces compatibilités sont une extension de la notion de compatibilité entre un ordre et une ultramétrique comme le montre le tableau 1, leur représentation visuelle étend également la représentation hiérarchique. On aboutit ainsi à une nouvelle forme de représentation visuelle qui a l'allure de "pyramide" ; nous verrons (en 9) que plus la matrice de dissimilarité représentée se rapproche d'une matrice ultramétrique, plus la forme de la pyramide se rapproche d'une hiérarchie.

TABLEAU 1



Représentation graphique (sans croisements)

A → B : A implique B ; [x] ⊂ [y] : l'ensemble des x est inclus dans l'ensemble des y ;
A ↔ B : A implique B et B implique A ; d est un indice de dissimilarité quelconque.

3. LES PYRAMIDES.

Avant d'énoncer leur définition, nous avons besoin d'introduire la notion de "compatibilité" entre un ordre Θ sur Ω et un ensemble P de parties de Ω. Rappelons d'abord qu'une partie h est connexe selon Θ si w' et w" étant les bornes (i.e. le plus petit et le plus grand élément) de h selon Θ, on a la condition :

$$\{w \text{ compris entre } w' \text{ et } w'' \text{ selon } \Theta\} \Longleftrightarrow \{w \in h\}.$$

- Définition de la compatibilité entre Θ et P.

Un ordre Θ est compatible avec un ensemble P de parties de Ω si tout élément $h \in P$ est convexe selon Θ.

- Définition d'une pyramide.

Soit Ω un ensemble fini, P un ensemble de parties non vides (appelées paliers) sur Ω, P est une pyramide si :

1° $\Omega \in P$ (le plus grand palier contient tous les individus)

2° $\forall w \in \Omega$, $\{w\} \in P$ (les points terminaux)

3° $\forall (h,h') \in P^2$ on a $h \cap h' = \emptyset$ ou $h \cap h' \in P$

4° Il existe un ordre Θ compatible avec P.

Exemples

Soit $\Omega = \{w_1, w_2, w_3\}$ et

$$P_1 = \{\{w_1\}, \{w_2\}, \{w_3\}, \{w_1 w_2\}, \{w_2 w_3\},$$

$$P_2 = P_1 \cup \{w_1 w_3\}$$

On vérifie facilement que P_1 est une pyramide alors que P_2 qui ne vérifie pas la quatrième condition n'en est pas une.

3.2. *Pyramides et hiérarchies*

Proposition 1

L'ensemble des hiérarchies est inclus dans l'ensemble des pyramides.

Démonstration.

Une hiérarchie H satisfait aux quatre conditions données dans la définition d'une pyramide ; les 2 premières sont identiques à celles données dans la définition d'une hiérarchie ; pour que H soit une hiérarchie il faut de plus que $\forall(h,h') \in H \times H$, on ait $h \cap h' = \emptyset$ ou $h \cap h'$ identique à h ou h', ce qui implique que $h \cap h' = \emptyset$ ou $h \cap h' \in H$; donc la troisième condition est satisfaite. Pour prouver que la quatrième condition l'est également, nous pouvons d'abord construire un ordre sur Ω induit par H, puis montrer que cet ordre est compatible avec H ; pour construire l'ordre, on utilise le fait que les éléments de H sont, soit emboîtés, soit d'intersection vide : on part de Ω et on choisit un ordre sur les plus grandes parties de H qui soient contenues dans Ω, on recommence le procédé avec chacune de ces parties en choisissant un ordre sur les plus grandes parties qu'elles contiennent et ainsi de suite jusqu'aux singletons qui respectent l'ordre induit par ce procédé et que nous appelons Θ. Cet ordre est connexe pour H ; en effet, soit $h \in H$ et soit (w',w") les extrêmités de h selon Θ, tous les éléments compris entre w' et w" appartiennent par construction même à des parties $h_\ell \in H$ incluses dans h et n'appartiennent qu'à celles-ci donc h est connexe selon Θ.

□

4. VISUALISATION D'UNE PYRAMIDE.

4.1. Notion de successeurs, prédecesseurs et niveau.

Etant donné une pyramide P, on dira que $h \in P$ est successeur de $h' \in P$ si $h \subset h'$ au sens strict et sauf si h' est un singleton ou si $h \equiv \Omega$, s'il n'existe pas h" différent de h et h' tel que $h \subset h'' \subset h'$ au sens strict ; on dira aussi que h' est prédecesseur de h.

On sait que $\Omega \in P$, l'ensemble des successeurs de Ω forme un recouvrement de Ω puisque P contient les singletons ; un tel recouvrement est appelé niveau de la pyramide ; l'ensemble des successeurs des paliers qui forment ce recouvrement forme un nouveau niveau qui est également un recouvrement ; on peut passer ainsi d'un niveau au suivant jusqu'au niveau formé uniquement de singletons puisque d'un niveau au suivant la taille des paliers va en se réduisant.

4.2. Les quatre aspects d'une visualisation

Pour visualiser une pyramide il faut :

a) préciser comment les paliers s'imbriquent les uns dans les autres.
b) Construire un ordre sur les singletons qui soit compatible avec la pyramide.
c) Indicer la pyramide, c'est à dire associer une hauteur à chaque palier.
d) Dire comment se fait la représentation graphique.

En 4.3) nous répondons à la première question en montrant que dans une pyramide chaque palier peut avoir jusqu'à deux prédécesseurs alors que dans une hiérarchie chaque palier a au maximum un seul prédécesseur (voir figure 5).

En 4.4) nous donnons quelques résultats qui débouchent sur un algorithme permettant (en utilisant uniquement les propriétés ensemblistes des pyramides) de construire un ordre compatible avec une pyramide.

En 5) nous définissons la notion d'indiçage d'une pyramide ; la notion de compatibilité entre un ordre et une pyramide s'exprime en terme de "croisement" ;

il en résulte un algorithme simple permettant d'obtenir un ordre sans croisement (i.e. compatible) avec une pyramide indicée (voir 5.4.).

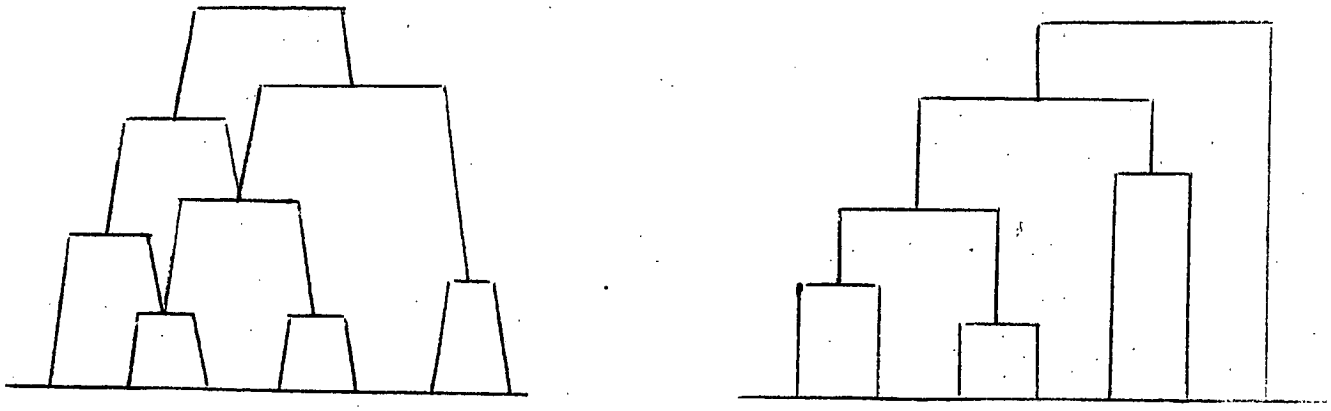
Enfin, en ce qui concerne la représentation graphique, on utilise les trois caractéristiques suivantes des pyramides :

i) deux paliers peuvent être d'intersection non vide.

ii) il existe au moins un ordre pour lequel chaque palier est connexe.

iii) chaque palier a au maximum deux ascendants.

Chaque palier est représenté par un segment horizontal , il est relié à ses ascendants par une ligne oblique ; chaque ligne oblique relie le milieu du segment associé à un palier à l'extrêmité du segment associé à son ou (ses) prédécesseur(s).



Pyramide Hiérarchie

Figure 5

4.3. *Nombre maximum de prédécesseur d'un palier d'une pyramide.*

Deux paliers h et h' sont dits connexes s'il existe une suite de paliers $h_1, \ldots, h_q$ telle que $h_1 = h$, $h_q = h'$ et $h_i \cap h_{i+1} \neq \emptyset$ pour $i=1,\ldots,q-1$. Une partie connexe est un ensemble de paliers connexes entre eux : l'ensemble des plus grandes parties connexes, appelées classes connexes associées à chaque niveau forme une partition de Ω (parfois réduite à un seul élément si tous les paliers du niveau sont connexes entre eux) ; en effet, la relation R :
h R h' $\Longleftrightarrow$ {h et h' sont connexes} est réflexive, symétrique et transitive.

Etant donné un ordre Θ compatible avec P ; un palier h est dit à "gauche" (resp. à "droite") d'un palier h' si parmi les paliers qui contiennent le plus petit (resp. le plus grand) élément de h' selon Θ, h a une intersection de plus grande taille avec h'.

Proposition 2

Soient h et h' deux paliers d'une même classe connexe si h est à gauche (resp. à droite) de h' alors h' est l'unique élément à droite (resp. à gauche)de h.

Démonstration.

Tous les paliers de P sont connexes selon un ordre Θ compatible avec P. Soient w_1 et w_2 (resp. w'_1 et w'_2) l'élément à gauche et à droite de h (resp. de h') selon Θ ; si h est à gauche de h' la seule configuration possible est celle de la figure 6 ; en effet, w_2 ne peut être à droite de w'_2 car alors h' serait inclus dans h et ne pourrait être successeur du niveau précédent ; si w_2 était à gauche de w'_1 l'intersection serait vide ; w_1 ne peut être compris entre w'_1 et w_2 car alors h serait inclus dans h' et ne pourrait être à son tour successeur du niveau précédent.

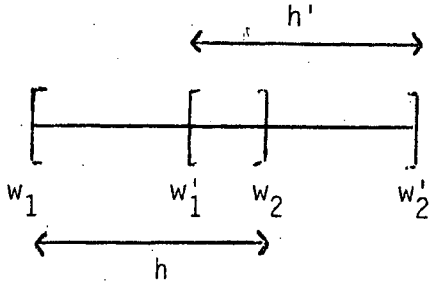


Figure 6

On peut maintenant voir facilement que h' est à droite de h ; en effet, s'il existait un palier h" contenant w_2 et d'intersection plus grande que h' $\cap$ h avec h, son plus petit élément w''_1 devrait être supérieur ou égal à w'_1 car s'il était inférieur son plus grand élément devrait être inférieur ou égal à w_2 (par définition de h qui est de plus grande intersection avec h') et donc égal à w_2 ; h" serait donc inclus dans h et ne pourrait être successeur du niveau précédent ; h' est donc une partie de plus grande intersection avec h.

Il reste à montrer que c'est la seule ; en effet, s'il en existait une autre h", son plus petit élément devrait être w'_1 et son plus grand élément devrait être soit strictement plus petit soit strictement plus grand que w'_2 ; dans le premier cas on aurait h" $\subset$ h' strictement et h" ne pourrait être un successeur du niveau précédent dans le second cas c'est h' qui ne pourrait être un successeur car on aurait h' $\subset$ h". □

Proposition 3.

Chaque palier d'une pyramide a au maximum deux prédécesseurs.

Démonstration.

Tout palier h de P a au moins un prédécesseur ; nous allons montrer que le nombre de prédécesseur est au maximum égal à 2 et que h est identique à leur intersection ; [soit h $\subset$ h' $\cap$ h" où h' et h" sont deux prédécesseurs de h, h ne peut être strictement inclus dans h' $\cap$ h" car par définition d'une pyramide h' $\cap$ h" $\in$ P et h' $\cap$ h" étant successeur du niveau contenant h' et h" on aurait h $\subset$ h' $\cap$ h" $\subset$ h' strictement, h ne pourrait être successeur de h' ; donc h $\equiv$ h' $\cap$ h" si h" contient le plus grand élément w' (on peut faire un raisonnement analogue si c'est le plus petit) de h' au sens de Θ, h" est à droite de h' ; en effet, s'il existait une intersection de h' strictement plus grande que h' $\cap$ h" contenant w' elle contiendrait strictement h = h' $\cap$ h" et h ne pourrait être un successeur de h' et h" ; h' est donc à gauche de h" d'après la proposition précédente qui prouve également qu'aucun autre palier ne peut contenir leur intersection h sinon il n'y aurait pas unicité du palier à gauche et à droite. Il en résulte que les seuls prédécesseurs de h sont h' et h".]

Une autre démonstration plus simple serait :

[Si le palier h est dans plus d'un palier prédécesseur, il est identique à leur plus grande intersection sinon il serait inclus dans cette intersection et ne pourrait être un successeur (l'intersection est un successeur car elle appartient

à P par définition d'une pyramide et elle n'appartient pas au niveau précédent puisqu'elle est strictement incluse dans les paliers successeurs qui la créent); la plus grande intersection est obtenue par deux paliers qui sont uniques d'après la proposition précédente puisqu'ils sont par définition à gauche et à droite l'un de l'autre.] □

4.4. Construction d'un ordre compatible avec une pyramide.

4.4.1. Une condition nécessaire et suffisante pour qu'un ordre Θ soit compatible avec une pyramide P.

Avant d'énoncer un algorithme constructif, nous avons besoin du résultat suivant

Proposition 4

Une condition nécessaire et suffisante pour qu'un ordre Θ sur Ω soit compatible avec une pyramide P est que toute classe connexe C soit ordonnée selon une suite unique $h_1 \ldots h_q$ caractérisée par les 2 propriétés suivantes : pour tout $i=1,\ldots,q$:

a) h_{i-1} et h_{i+1} sont parmi les paliers de C, ceux qui sont de plus grande intersection avec h_i.

b) $h_i \cap h_{i+1}$ est différent de h_i et de h_{i+1}.

Démonstration

Soit C une classe connexe d'une pyramide P et soit Θ un ordre compatible avec P ; soit w_1 le plus petit élément (au sens de Θ) de C ; le palier h_1 de C qui contient $w_1 \in \Omega$ ne peut avoir de palier à gauche car alors w_1 ne serait plus le plus petit élément de C : le palier h_2 qui se trouve à droite de h_1 est unique d'après la proposition 2 ; on peut définir ainsi à partir de h_1 une suite finie h_i pour $i=1,\ldots,\ell$ telle que h_{i+1} est l'unique élément qui se trouve à droite de h_i et h_ℓ est le palier qui contient $w_\ell \in \Omega$ le plus grand élément de C.

Cette suite contient tous les paliers de C ; en effet, soit h'_p un palier quelconque de C ; il existe certainement une suite de palier h'_i qui relie h_1 à h'_p et satisfait à la condition $h'_i \cap h'_{i+1} \neq \emptyset$ sinon h'_p et h_1 ne seraient pas connexes ; parmi les paliers qui précédent h'_p dans cette suite, il en existe certainement au moins un qui contient w'_p le plus petit élément de h'_p (dans le cas contraire si h'_{p-1} ne contient pas w'_p, comme $h'_{p-1} \cap h'_p \neq \emptyset$, h'_{p-1} doit contenir le plus grand élément de h'_p et ne peut contenir d'éléments inférieurs à w'_p pour rester connexe, il en est de même pour h'_{p-2} puisque $h'_{p-2} \cap h'_{p-1} \neq \emptyset$ et ainsi de suite, donc w_1 ne pourrait être atteint) ; notons h'_{p-1} le palier à gauche de h'_p (il existe certainement d'après ce qui précède) ; en reprenant le même raisonnement on peut montrer que h'_{p-1} a un élément à gauche et ainsi de suite jusqu'au palier h'_j d'intersection non vide avec h_1 ; si h_1 n'est pas à gauche de h'_j on poursuit la suite avec h'_{j-1} qui est à gauche de h'_j, h'_{j-2} à gauche de h'_{j-1} etc... tant que h_1 n'est pas atteint (tous ces h'_{j-q} ont un élément à gauche puisque leur intersection avec h_1 n'est pas vide) ; par ce procédé, on atteint certainement h_1 car on finit par tomber sur le palier h'_k qui contient w_1 et qui ne peut être strictement plus petit que h_1 car il ne pourrait être successeur du niveau précédent, ni plus grand que h_1 car alors c'est h_1 qui ne pourrait être un successeur. On a ainsi défini une suite $h'_k = h_1, h'_{k+1}, \ldots, h'_i$ qui relie h_1 à h_i avec h'_{j+1} à droite de h'_j pour j=k à j=i-1 ; comme chaque élément à droite est unique d'après la proposition 2, les termes des uites h'_i et h_i sont identiques à partir de i=k pour la suite h'_i et pour la suite h_i. Il en résulte finalement que h'_p appartient bien à la suite h_i.

On voit enfin que la suite h_i satisfait bien aux conditions a) et b) de la proposition puisque h_{i-1} est à gauche de h_i et h_{i+1} à droite.

Reste à montrer la condition suffisante ; autrement dit, que si les éléments de toute classe connexe d'une pyramide P, sont ordonnés selon une suite de paliers connexes satisfaisant aux conditions a) et b) alors Θ est compatible avec P. Ceci se démontre facilement ; en effet : soit $h \in P$, il existe certainement un niveau où h apparaît ; les classes connexes de ce niveau forment une partition des paliers qui le composent donc h appartient à l'une de ces classes; chaque classe connexe étant formée d'une suite de paliers connexes selon Θ, il en résulte que h est connexe pour Θ.

□

4.4.2. Un algorithme constructif d'un ordre compatible avec une pyramide.

On part du palier Ω ; l'ensemble des successeurs de Ω forme le premier niveau de la pyramide. On note N_1 ce niveau. On considère h un palier quelconque de N_1 et on construit la suite unique (d'après la proposition 4) associée à la classe connexe de N_1 qui contient h ; cette suite notée $h_1 \ldots h_q$ est facile à construire en utilisant les propriétés a) et b) de la proposition 4 : on part de $h = h_i$, on construit h_{i-1} et h_{i+1} puis h_{i-2} et h_{i+2} etc...

On choisit un ordre sur les éléments de cette classe ; on recommence le procédé avec d'autres paliers du niveau N_1 tant qu'ils ne sont pas tous atteints. On choisit un ordre entre les classes connexes ainsi obtenues ; cet ordre est arbitraire pour le niveau N_1 car si un ordre est compatible avec P, en ordonnant autrement les parties connexes du niveau N_1, on obtient un nouvel ordre qui reste compatible avec P puisque tous les paliers restent connexes (cette propriété n'est généralement pas vraie pour les niveaux suivants).

Ayant ainsi défini un ordre sur chaque classe connexe et entre ces parties, on a un ordre O_1 sur les paliers de N_1.

On dit qu'un ordre Θ sur Ω est cohérent avec O_1 si

i) chaque palier de N_1 est connexe pour Θ

ii) si h_i précède h_j selon O_1 alors tout élément de h_i qui n'est pas dans h_j précède tout élément de h_j selon Θ.

Soit $O(N_1)$ l'ensemble des ordres sur Ω qui sont "cohérents" avec O_1 ; à l'étape i on ordonne les successeurs des paliers de N_i de façon à obtenir un ordre O_{i+1} qui soit cohérent avec des ordres de $\Theta(N_i)$; il suffit pour cela d'ordonner chaque classe connexe de N_{i+1} comme précédemment et ces classes entre elles de façon que le dernier palier de chaque classe connexe et le premier palier de la classe connexe qui suit selon O_{i+1} aient même ascendant ; On obtient ainsi un ordre O_{i+1} sur les paliers du niveau N_{i+1} ; soit $\Theta(N_{i+1})$ l'ensemble des ordres sur Ω cohérents avec O_{i+1} et appartenant à $\Theta(N_i)$. On recommence le procédé jusqu'à parvenir au niveau N_ℓ dont tous les paliers sont des singletons ; $\Theta(N_\ell)$ n'est pas vide (et à forciori aucun des $\Theta(N_i)$ n'est vide) car s'il l'était, il n'existerait pas d'ordre compatible avec P et P ne serait

pas une pyramide. Pour avoir un ordre compatible avec P il suffit d'en choisir un quelconque dans $\Theta(N_\ell)$; en effet, d'après la condition suffisante de la proposition 4 $\Theta \in \Theta(N_\ell)$ est compatible avec P puisque par construction toutes les classes connexes de P sont ordonnées par Θ selon une suite de paliers connexes selon Θ et caractérisée par les conditions a) et b) de cette proposition.

En 7.4. nous construisons la représentation visuelle d'une pyramide à l'aide de l'algorithme que nous venons de décrire.

En 6.5 nous verrons un algorithme rapide (utilisant la notion d'indiçage) pour donner un ordre compatible avec P.

5. INDICAGE D'UNE PYRAMIDE

5.1. *Pyramides indicées*

- Définition d'une pyramide indicée

.Une pyramide indicée est un couple (P,f) où P est une pyramide et f une application de P dans $\mathbb{R}^+$ telle que :

1° f(h) = 0 si et seulement si h ne contient qu'un seul élément.

2° Pour tout h et h' dans P, $h \subset h'$ (inclusion stricte) implique $f(h) \leq f(h')$.

-Définition d'une pyramide indicée au sens large et au sens stricte.

. Une pyramide indicée est indicée au sens large si $h \subset h'$ strictement et $f(h) = f(h')$ impliquent l'existence de h_1 et h_2 dans P différents de h : $h = h_1 \cap h_2$. (si h pouvait être égal à h_1 ou h_2, la condition serait toujours satisfaite avec $h = h' \cap h$).

. Une pyramide indicée est indicée au sens stricte si { $h \subset h'$ strictement $\Rightarrow f(h) < f(h')$}.

. La quantité f(h) est appelée hauteur du palier h.

Exemples $\Omega = \{a,b,c,d\}$ (voir figure 7)

$P_1 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$

$f(P_1) = \{0, 0, 0, 0, 2, 1, 2, 3\}$

P_1 est une pyramide indicée au sens strict.

$P_2 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$

$f(P_2) = \{0, 0, 0, 0, 1, 1, 2, 3\}$ voir figure 7.b

P_2 est une pyramide indicée au sens large puisque

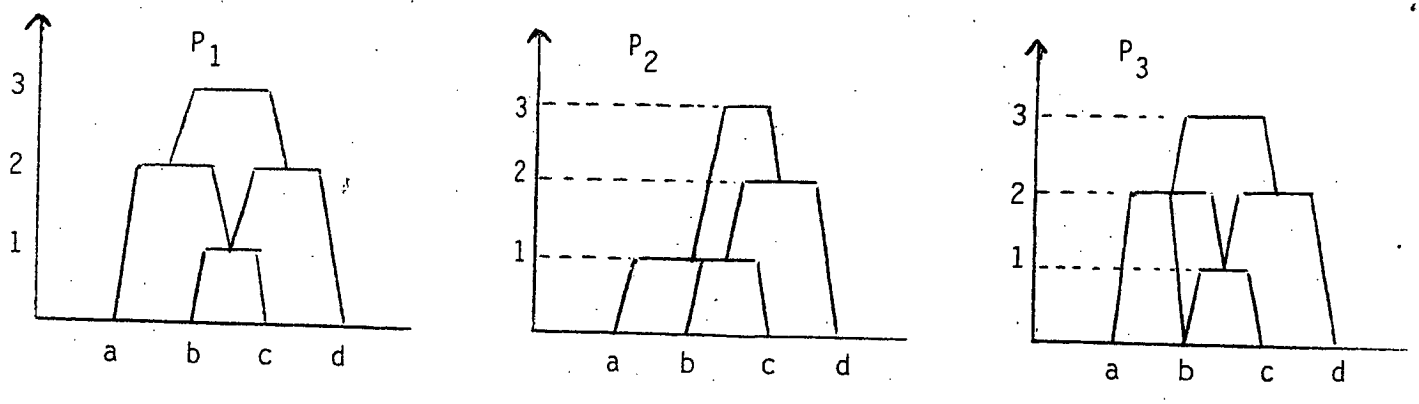
$h_1 = \{bc\} \subset h_2 = \{abc\}$ et $f(h_1) = f(h_2)$.

$P_3 = P_1 \cup \{ab\}$

$P_3 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{ab\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$

$f(P_3) = \{0, 0, 0, 0, 2, 2, 1, 2, 3\}$ voir figure 7.c

P_3 est une pyramide qui n'est ni indicée au sens large ni indicée au sens strict car {ab} n'est pas l'intersection de deux paliers dont l'un est différent de {ab}.



Pyramide indicée au sens strict (a)

Pyramide indicée au sens large (b)

Pyramide indicée (c)

Figure 7

- Pyramide binaire

C'est une pyramide dont chaque palier, non réduit à un singleton est formé de la réunion de deux paliers distincts.

- Remarquons qu'une pyramide peut être indicée au sens large et être binaire (voir par exemple P_2 dans l'exemple précédent) et que toute pyramide peut être rendue binaire.

Définition d'un indice de dissimilarité

Un indice de dissimilarité est une application de $\Omega \times \Omega \to \mathbb{R}^+$ qui satisfait aux deux propriétés suivantes :

1° s est symétrique : $\forall (w,w') \in \Omega \times \Omega \quad s(w,w') = s(w',w)$

2° $s(w,w') = 0$ si $w \equiv w'$.

5.2. *Indices pyramidaux.*

5.2.1. *Définition d'un indice pyramidal.*

C'est un indice de dissimilarité qui vérifie en plus les deux conditions suivantes :

1° $s(w,w') = 0 \iff w \equiv w'$.

2° Il existe un ordre Θ sur Ω tel que tout triplet w, w', w" avec w' compris entre w et w" selon Θ satisfait à l'inégalité suivante :

$$s(w,w") \geq \text{Max } \{s(w,w'), s(w',w")\} \quad \text{(inégalité pyramidale).}$$

On voit donc (d'après 2.1) que tout ordre qui satisfait à l'inégalité pyramidale est compatible avec S.

5.2.2. *Propriétés des indices pyramidaux.*

On a les résultats suivants :

Proposition 5

L'ensemble des ultramétriques est inclus dans l'ensemble des indices pyramidaux.

Démonstration

Une ultramétrique étant une distance, la première condition est satisfaite; d'autre part on sait qu'une ultramétrique δ étant donnée on peut lui associer des ordres Θ tels que $M(\delta,\Theta)$ soit une matrice ultramétrique (et donc Robinson), la $2^{ème}$ condition est donc satisfaite (voir 2) et [10], [12]).

Rappelons que ces ordres sont ceux qui sont sans croisement dans la visualisation de la hiérarchie induite par δ.

□

Proposition 6

Les conditions suivantes sont équivalentes si s est un indice pyramidal.

1) Θ est compatible avec s.

2) $M(s,\Theta)$ est Robinson

3) Tout couple d'éléments (w,w') compris (au sens large) selon Θ entre deux éléments w_i et w_j est tel que $s(w_i,w_j) \geq s(w,w')$.

Démonstration

1) => 2)

Nous savons (voir le §2) que si Θ est compatible avec un indice de dissimilarité alors la matrice $M(s,\Theta)$ est Robinson.

2) => 3)

Nous savons (voir le § 2) que si $M(s,\Theta)$ est Robinson alors $M(s,\Theta)$ est SDR ce qui implique que s et Θ sont semi-compatibles d'où la condition 3).

3) => 1)

Il suffit de prendre $w_i = w$ ou $w' = w_j$ pour retrouver la condition 1).

□

6. CONSTRUCTION D'UNE PYRAMIDE.

6.1. *Algorithme de classification ascendante pyramidale (CAP)*

Par analogie à la classification ascendante hiérarchique, après avoir choisi un indice d'agrégation (c'est un indice de dissimilarité entre classes, voir par exemple dans [8] une liste de choix possibles) on peut construire un algorithme de classification ascendante de la façon suivante :

a) chaque élément de Ω est appelé "groupe".

b) on agrège les 2 groupes les plus proches parmi les groupes qui n'ont pas été agrégés 2 fois.

c) on recommence b) jusqu'à ce qu'un groupe qui contient Ω soit formé.

Si l'on ne veut pas produire des configurations du type indiqué figure 8 et 9 ; il faut ajouter les conditions suivantes :

d) chaque fois qu'un groupe est formé on lui associe un ordre sur les 2 groupes qu'il réunit.

e) Deux groupes ne peuvent être réunis s'ils ne sont pas connexes.

f) Soient i et j les éléments extrêmes de la partie connexe de Ω associée à un groupe h ; aucun groupe ne peut se connecter à un groupe inclus dans h qui ne contient ni i ni j.

Remarque :

La condition d) induit un ordre Θ sur les éléments de Ω ; la condition e) implique que tous les groupes formés par l'algorithme sont formés d'une partie connexe selon Θ de Ω ; la condition f) en s'ajoutant à e) permet d'éviter les croisements au niveau de la représentation visuelle.

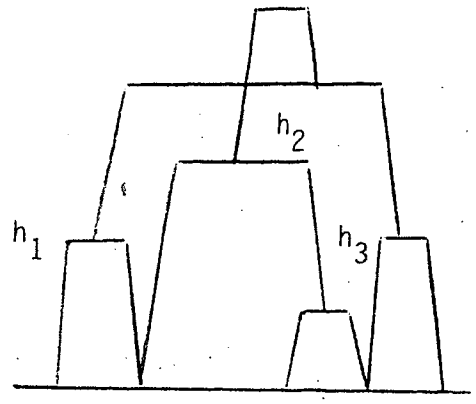


Figure 8
f ≠> e

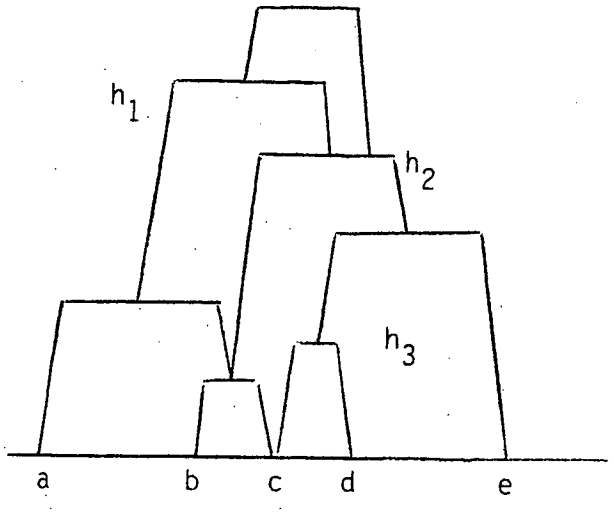


Figure 9
e ≠> f

On voit sur la figure 8 que la condition e) n'est pas satisfaite (h_1 et h_3 ne sont pas connexes à cause de h_2) alors que la condition f) l'est. Par contre, sur la figure 9, on voit que la condition e) est satisfaite mais pas la condition f) puisque h_1 contient $h_3 \subset h_2$ et h_3 ne contient ni b ni e les extrêmités de la partie connexe associée à h_2.

- Les étapes a) b) c) et les conditions d) e) f) définissent un algorithme dit de "classification ascendante pyramidale" (notée CAP) ; on a le résultat suivant :

Proposition 7

L'algorithme de CAP construit une pyramide.

Démonstration

Soit P l'ensemble des groupes construits par l'algorithme de CAP ; montrons que P forme une pyramide.

On voit d'abord que par construction même les singletons et l'ensemble des individus Ω sont dans P.

La condition d) construit un ordre Θ qui est compatible avec P d'après la condition e).

Il reste donc à montrer que si h_1 et h_2 sont deux groupes construits par l'algorithme tels que $h_1 \cap h_2 \neq \emptyset$ alors $h_1 \cap h_2 = h$ est dans P. Si h_1 et h_2 sont d'intersection non vide, leur partie commune est formée d'un ensemble de groupes notés $E \subset P$; soient $H_1, \ldots, H_n$ les plus grands groupes de E (ce sont des groupes qui ne sont strictement contenus dans aucun groupe de E) ; nous allons montrer que ces groupes sont en fait identiques.

Soient i et j les extrêmités de l'ordre O associé à h_1 par l'algorithme (condition d)) ; les H_ℓ étant des groupes de h_1 qui sont dans h_2 doivent contenir i ou j (d'après la condition f).

Si l'une des extrêmités (par exemple j) n'apparaît dans aucun de ces groupes tous les H_ℓ contiennent i ; soit i' l'élément le plus éloigné de i selon O et contenu dans l'un de ces groupes ; il existe donc un groupe $H \in \{H_1,\ldots,H_n\}$ qui contient i et i' ; d'après la condition e) ce groupe contient tous les éléments intermédiaires entre i et i' ; il contient donc tous les H_ℓ puisque les éléments des H_ℓ sont par définition de i' compris entre i et i' selon l'ordre O_1. Comme par définition, les H_ℓ ne sont strictement contenus dans aucun groupe de E, ils sont tous identiques à H ; d'où $h_1 \cap h_2 = H \in P$.

Si les éléments i et j apparaissent dans la réunion des groupes H_ℓ, il en résulte que h_2 contient (d'après la condition e)) tous les éléments compris entre i et j selon Θ et donc que $h_1 \subset h_2$; d'où dans ce cas aussi $h_1 \cap h_2 \in P$ puisque $h_1 \cap h_2 = h_1$.

□

6.2. Algorithme des voisins réciproques.

On peut définir, pour les pyramides, un algorithme de voisins réciproques analogue à celui qui est utilisé pour les hiérarchies (voir [20]) ; il suffit de remplacer dans l'algorithme CAP l'étape b) par :

b') on agrège les groupes qui sont voisins réciproques parmi les groupes qui ne sont pas agrégés deux fois.

Il serait intéressant d'étudier les conditions nécessaires pour que la pyramide ainsi obtenue soit la même que celle donnée par l'algorithme de CAP ; ces conditions sont-elles les mêmes que celles obtenues dans le cas des hiérarchies (voir [20], [11]) ?

6.3. *Utilisation d'une formule de récurrence.*

Pour construire une pyramide par l'algorithme de CAP on peut utiliser tous les indices d'agrégation classiques (voir [8]) et se servir de la formule de récurrence générale de Lance-Williams-Jambu.

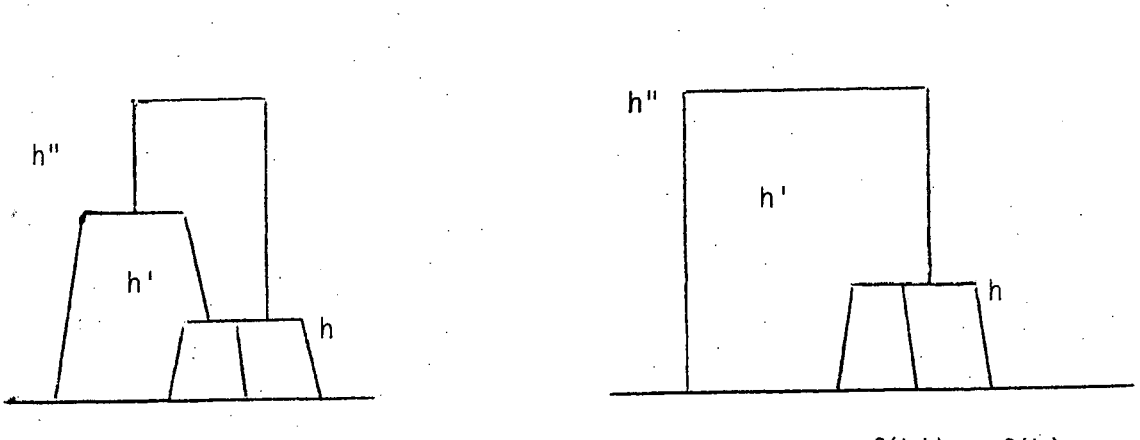
$$\delta(h,h_1 \cup h_2) = a_1\,\delta(h,h_1) + a_2\,\delta(h,h_2) + a_3\,\delta(h_1,h_2) + a_4\,f(h) +$$

$$+ a_5\,f(h_1) + a_6\,f(h_2) + a_7|\delta(h,h_2) - \delta(h,h_1)|$$

Dans le cas où l'indice d'agrégation choisi est celui du minimum ou du maximum on obtient des propriétés particulières qui sont énoncées en 8.

6.4. *Epuration d'une pyramide.*

Comme cela se fait pour les hiérarchies, l'épuration porte d'abord sur les paliers redondants donnés par l'algorithme de construction pyramidal (CAP par exemple) ; on élimine tous les paliers qui sont inutiles pour la représentation visuelle de la pyramide ; autrement dit tous les paliers $h \in P$ tels qu'il existe $h' \in P$ avec $f(h) = f(h'), h' \subset h$ et qu'il n'existe pas $h'' \in P$ différent de h tel que $h' \subset h''$ (voir figure 10) où f définit l'indiçage choisi pour la pyramide (par exemple $f(h) = \delta(h_1,h_2)$ si h est formé de la réunion de h_1 et h_2). De cette façon on aboutit toujours à une pyramide indicée au sens large.



h' = h ∩ h"
=> h' n'est pas supprimé

h' ⊂ h f(h') = f(h)
h' est supprimé.

Figure 10
Epuration

Une autre forme d'épuration qui n'existe pas pour les hiérarchies concerne les "crêtes" de la pyramide. Une arête est le segment de droite qui relie un palier aux plus petits paliers qui le contiennent (2 au maximum).

En effet, une fois l'épuration des paliers réalisée, on s'aperçoit que certains d'entre eux sont reliés par une arête, à plusieurs paliers qu'ils contiennent (par exemple, dans la figure 10 on voit que le palier h est relié par 4 arêtes notées a, b, c, d aux paliers h_1, h_2, h_3, h_4) ; ces arêtes servent à montrer l'inclusion d'un palier dans un autre (par exemple l'arête 2 sert à montrer l'inclusion du palier h_2 dans le palier h); les arêtes "extrêmes" (comme 1 et 4 dans la figure 11) sont indispensables mais certaines arêtes "intérieures" (comme 2 ou 3) deviennent inutiles si l'inclusion est montrée par un palier intermédiaire (l'arête 3 peut être supprimée car elle est redondante avec l'arête 4 qui relie h_3 à h par le palier intermédiaire h_4).

La règle générale de suppression est la suivante : étant donné un palier h, on supprime toutes les arêtes "intérieures" qui relient plus d'une fois un palier avec les paliers qu'il contient.

Dans l'exemple de la figure 11, les arêtes 1 et 4 qui sont extrêmes doivent être conservées ; l'arête 3 doit être supprimée et l'arête 2 qui n'est relié qu'une fois au palier h doit être conservée.

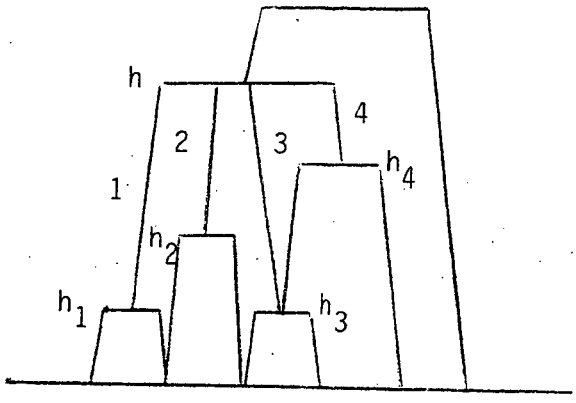


Figure 11

6.5. *Indiçage d'une pyramide obtenue par l'algorithme de CAP et construction d'un ordre compatible.*

Pour ne pas avoir d'inversion (i.e. un palier plus haut que le palier qui le contient) dans la représentation visuelle de la pyramide obtenue par l'algorithme de CAP, on peut toujours définir la hauteur de chaque palier de la façon suivante :

$f_1(h \cup h') = \text{Max}\ (\delta(h, h'), f(h), f(h'))$ où $h \cup h'$ est l'ascendant de h et h'.

Cependant, la représentation visuelle d'une pyramide est plus facilement interprétable si elle est indicée de façon à ce que la hauteur d'un palier corresponde à la valeur prise par l'indice d'agrégation pour les deux paliers qui l'ont formé. Autrement dit : $f_2(h) = \delta(h_1, h_2)$ si h est l'ascendant de h_1 et h_2.

Avec un tel choix il peut apparaître des inversions. Dans [9] nous avons largement étudié ce qui se passe en liaison avec la formule de Lance et Williams.

Ainsi la pyramide est indicée au sens large par f_1 mais pas toujours par f_2.

Nous verrons en 7) (proposition 8) que si l'on associe à chaque couple $(w,w') \in \Omega \times \Omega$ la hauteur du plus bas palier d'une pyramide P indicée au sens large, on définit un indice pyramidal $s(w,w')$; on montrera de plus (proposition 9) que si Θ est compatible avec s alors il est compatible avec P ; d'où l'intérêt du paragraphe suivant car ayant obtenu les paliers d'une pyramide par l'algorithme de CAP et ayant définit par f_1 ou f_2 leur hauteur il est nécessaire de connaître un ordre sur les singletons qui soit compatible avec P.

6.6. Construction d'un ordre Θ compatible avec un indice pyramidal s.

Connaissant un indice pyramidal s, on peut construire un ordre compatible Θ, de la façon suivante : on part d'un élément quelconque w et on construit successivement la suite obtenue en passant de l'étape $w_i \ldots w_j$ à l'étape suivante en ajoutant l'élément w_ℓ à gauche ou à droite suivant que c'est parmi les éléments de Ω non encore placés, celui qui est le plus proche de w_i ou de w_j ; si w_ℓ est unique on a nécessairement $s(w_\ell,w_i) \geq s(w_i,w_j)$ si w_ℓ est à droite de w_j et $s(w_\ell,w_j) \geq s(w_i,w_j)$ si w_ℓ est à gauche de w_i ; si w_ℓ n'est pas unique on peut toujours en choisir un qui satisfait à cette inégalité sinon il n'existerait pas d'ordre Θ compatible avec s, ce qui serait contraire à la définition d'un indice pyramidal.

7. EXISTENCE D'UNE BIJECTION ENTRE LES INDICES PYRAMIDAUX ET LES PYRAMIDES INDICEES.

L'existence d'une telle bijection est un résultat important qui étend le théorème de bijection entre hiérarchies et ultramétriques (voir par exemple [8]) ; l'énoncé de ce résultat est court mais sa démonstration est beaucoup plus longue.

Proposition 8

Il existe une bijection entre l'ensemble des pyramides indicées au sens large, et l'ensemble des indices pyramidaux $\mathcal{S}$

Démonstration

Nous allons montrer qu'il existe une application ϕ de Π dans $\mathcal{S}$ et une application ψ de $\mathcal{S}$ dans Π puisque ϕ et ψ sont inverses l'une de l'autre.

Construction d'une application ϕ de Π dans $\mathcal{S}$.

Soit $\phi : \Pi \rightarrow \mathcal{S}$ telle que

$$\phi((P,f)) = s \text{ avec } s(k,\ell) = \inf \{f(h)/h \in P, (k,\ell) \in h \times h\}$$

où (P,f) est une pyramide notée P indicée au sens large par f.

Montrons que s est une dissimilarité pyramidale.

. $s(\ell,k) = 0 \iff \ell = k$; en effet $\ell = k \Rightarrow s(\ell,k) = 0$ car $s(\ell,\ell) = f(h)$ avec $h = \{\ell\}$ d'où $f(h) = 0$ par définition de f ; $s(\ell,k) = 0 \Rightarrow \ell = k$ puisque le plus bas des paliers contenant ℓ et k étant de hauteur nulle ne peut contenir qu'un seul élément par définition de f.

. $s(\ell,k) = s(k,\ell)$ car évidemment $(k,\ell) \in h \times h \Rightarrow (\ell,k) \in (h \times h)$

Reste à démontrer que s satisfait à l'inégalité pyramidale ou autrement dit, qu'il existe un ordre compatible avec s.

Soit Θ un ordre compatible avec la pyramide P ; nous allons montrer que Θ est compatible avec s. Considérons un triplet quelconque i,j,ℓ de Ω tel que j soit compris entre i et ℓ selon Θ : le plus bas palier noté $h_{i\ell}$ qui contient i et ℓ contient i et j puisqu'il doit être connexe, donc le plus bas palier noté h_{ij} qui contient i et j est au maximum à la hauteur de $h_{i\ell}$ puisque P est indicé au sens large par f donc $s(i,\ell) \leq s(i,j)$; on démontre de même que $s(\ell,j) \leq s(i,j)$, d'où finalement $s(i,j) \geq \text{Max}\ \ s(i,\ell),\ s(\ell,j)$. Donc Θ est bien compatible avec s et s est bien un indice pyramidal.

Construction d'une application ψ de $\mathcal{S}$ dans Π .

Soit $\psi : \mathcal{S} \longrightarrow \Pi$ telle que $\psi(s) = (P,f)$ où

. P est l'ensemble des parties h de Ω qui satisfont à la condition :

$$\begin{cases} \exists\, \alpha : h = \{x \in \Omega \ /\ \forall y \in h \quad s(x,y) \leq \alpha\} \qquad (1) \\ \text{ou} \quad \{ \quad h_1 \text{ et } h_2 \text{ d'intersection non vide dans P tels que} \\ \qquad h \neq h_1,\ h \neq h_2 \text{ et } h = h_1 \cap h_2\} \qquad (2) \end{cases}$$

L'ensemble des parties h de Ω qui satisfont à la condition (1) est noté $P(\alpha)$

. f est l'application $P \to \mathbb{R}^+$:

$$f(h) = \text{Min}\ \{\alpha / h \in P(\alpha)\}$$

Remarquons que $P(\alpha)$ est généralement un recouvrement et non une partition car la relation $x\ R\ y \iff s(x,y) \leq \alpha$ bien que réflexive et symétrique n'est pas transitive ; remarquons aussi que l'ensemble P' formé de tous les éléments des $P(\alpha)$ pour α quelconque dans $\mathbb{R}^+$ est inclus et non nécessairement identique à P car des intersections de classes de P' peuvent ne pas être dans P' mais sont dans P.

En conséquence P' n'est pas toujours une pyramide.

Exemple. $\Omega = \{a, b, c, d\}$

$P = \{\{a\}, \{b\}, \{c\}, \{d\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$

$f(P) = \{0, 0, 0, 0, 1, 1, 2, 3\}$ (voir figure 12)

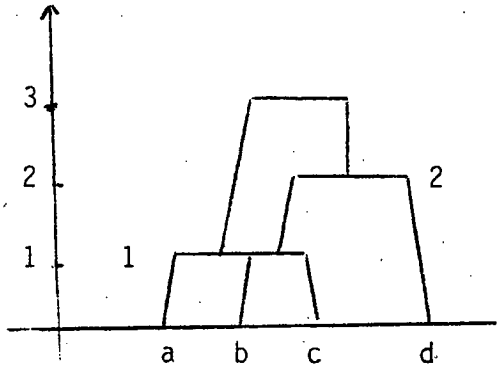


Figure 12

On voit que $P' = P - \{bc\}$ puisque $\{bc\} \subset \{abc\}$ ne satisfait pas à la condition (1) qui est satisfaite par le palier $\{abc\}$: on remarque que $\{bc\} = \{abc\} \cap \{bcd\}$ satisfait bien à la condition (2) et que plus généralement toutes les intersections non vides d'éléments de P sont dans P ; P qui contient Ω et les parties réduites à des singletons est donc une pyramide ; ce qui n'est pas le cas de P'.

Il faut montrer que (P,f) est une pyramide indicée.

Montrons d'abord deux résultats qui nous serviront par la suite :

1° $f(h) = \text{Min } \{\alpha/\alpha \in \mathbb{R}^+, h \in P(\alpha)\}$ autrement dit : $h \in P(f(h))$.

2° $f(h) = \text{Max } \{s(\ell,k)/(\ell,k) \in h \times h\}$

Montrons le 1° :

Si h est réduit à un seul élément, on a $f(h) = 0$ et $h \in P(0)$ donc $h \in P(f(h))$. Si h contient plus d'un élément, soit $\alpha_0 = \inf \{\alpha/\alpha \in \mathbb{R}^+, h \in P(\alpha)\}$; considérons une suite décroissante $\alpha_n \to \alpha_0$ et soit ℓ un élément extérieur à h tel que $s(\ell,j) = \text{Min } \{s(\ell,i)/\ell \notin h,\ i \in h, s(\ell,i) > \alpha_0\}$, (pour tout $\ell \notin h$, il existe au moins un $i \in h$: $s(\ell,i) > \alpha_0$ sinon il serait dans h). Il existe N suffisament grand tel que pour tout $n > N$ on ait $s(\ell,j) > \alpha_n$, $s(i,j) \leq \alpha_n$ pour tout $(i,j) \in h \times h$; en passant à la limite on a donc $s(\ell,j) > \alpha_0$ et $\forall (i,j) \in h \times h \quad s(i,j) \leq \alpha_0$; donc tous les éléments de h sont dans une classe de $P(\alpha_0)$ et cette classe ne contient aucun élément qui n'est pas dans h ; donc $h \in P(\alpha_0)$ d'où $f(h) = \alpha_0$.

Montrons le 2° :

Soit $s(x,y) = \text{Max } \{s(i,j)/(i,j) \in h \times h\}$; si $s(x,y) < \alpha_0$, il existerait α tel que $s(x',y') < \alpha < \alpha_0 < s(\ell,y)$ où $\ell \notin h$ est défini comme ci-dessus ; on aurait donc $h \in P(\alpha)$ et α_0 ne serait pas la plus petite valeur telle que $h \in P(\alpha_0)$; donc $s(x,y) \geq \alpha_0$; on a vu ci-dessus que $s(x,y) \leq \alpha_0$ d'où finalement $\alpha_0 = f(h) = s(x,y)$.

Montrons maintenant que le couple (P,f) définit bien une pyramide indicée.

P satisfait bien aux quatre conditions qui définissent une pyramide :

1° $\Omega \in P$; en effet, prenons $\alpha = \text{Max } \{s(x,y)/(x,y) \in \Omega \times \Omega\}$; la partie $h = \{x \in \Omega/\forall y \in \Omega\ s(x,y) \leq \alpha\}$ est par définition un élément de P et elle est identique à Ω.

2° $\forall w \in \Omega$, $\{w\} \in P$; en effet, soit $h = \{w\}$, on a $h = \{w/\forall y \in h\ s(w,y) = 0\}$ puisque $s(w,y) = 0 \iff w \equiv y$ par définition de s.

3° $h_1 \in P$, $h_2 \in P$ et $h = h_1 \cap h_2 \neq \emptyset$ implique $h \in P$; en effet, si $h = h_1 \cap h_2$ satisfait à la condition (1) ou si $h = h_1$ ou $h = h_2$, h est dans P ; dans le cas contraire $h \in P$ d'après (2).

4° Il existe un ordre Θ compatible avec P.

Soit Θ un ordre compatible avec s, nous allons montrer qu'il est compatible avec P.

Soit $h \in P$, il faut montrer que h est connexe selon Θ : autrement dit, que si w_1 et w_2 sont les bornes de h selon Θ on a : i) tout élément w compris au sens de Θ entre w_1 et w_2 est dans h et ii) tout élément extérieur à l'intervalle $[w_1,w_2]$ est hors de h.

i) Soit $w \in h$ tel que $w_1 \leq w \leq w_2$ d'où d'après la proposition 6 $\forall w' \in h$ on a $s(w,w') \leq s(w_1,w_2) = \text{Max } \{s(x,y)/(x,y) \in h \times h\}$ d'où $s(w,w') \leq f(h)$; or

$$h = \{x \in \Omega/(w_1,w_2) \in h \times h,\ \forall w' \in h.\ s(w,w') \leq f(h)\}$$

d'où $w \in h$.

ii) Si $w \notin [w_1,w_2]$, w est soit à gauche de w_1 soit à droite de w_2 selon Θ; dans le premier cas on a : $s(w,w_2) > s(w_1,w_2) = f(h)$; dans le second cas : $s(w,w_1) > s(w_1,w_2) = f(h)$ donc dans tous les cas w n'appartient pas à h.

Montrons maintenant que la pyramide est indicée au sens large ; autrement dit, que f satisfait aux trois conditions de la définition :

1° $f(h) = 0 \iff \{h \text{ est un singleton}\}$; en effet,
$f(h) = 0 \iff \min \{\alpha/h \in P(\alpha)\} = 0$
$\iff \{h \in P(0)\} \iff \{\forall y \in h\ \ s(x,y) = 0\}$
$\iff \{\forall y \in h\ \ x \equiv y\} \iff \{h \text{ ne contient qu'un seul élément}\}$.

2° $(h,h') \in P \times P$ et $h \subset h'$ strictement $\Rightarrow f(h) \leq f(h')$; en effet, soit $w' \in \Omega$ contenu dans le complémentaire de h dans h' ; on a nécessairement $s(w,w') \geq f(h')$ pour au moins un élément $w \in h$ sinon w' serait dans h, d'autre part $s(w,w') \leq f(h') = \text{Max } \{s(\ell,k)/(\ell,k) \in h \times h\}$ puisque $(w,w') \in h' \times h'$ d'où $f(h) \leq s(w,w') \leq f(h')$; on a bien l'inégalité cherchée.

3° $f(h) = f(h')$ et $h \subset h'$ strictement impliquent l'existence de h_1 et h_2 dans P tels que $h = h_1 \cap h_2$ et $h \neq h_1$, $h \neq h_2$ car h n'étant pas identique au plus grand palier de hauteur $f(h)$ ne peut satisfaire la condition (1).

. Les applications ψ et ϕ sont inverses l'une de l'autre.

Il faut démontrer que $\phi \circ \psi(s) = s$ puis que $\psi \circ \phi((P,f)) = (P,f)$; montrons d'abord que $\phi \circ \psi(s) = s$, autrement dit que si $\phi((P,f)) = \sigma$ et $\psi(s) = (P,f)$ alors $\sigma = s$; soit h un palier de plus petite hauteur qui contient x et y, par définition de σ, on sait que $\sigma(x,y) = f(h)$.

Soit $f(h) = \alpha$, on a $s(x,y) \leq \alpha$ puisque par définition de ψ, h est formé d'un ensemble d'éléments dont la distance est inférieure à $f(h)=\text{Min}\{\alpha'/h\in P(\alpha')\}$; d'autre part $s(x,y)$ ne peut être strictement inférieur à α car il existerait h' contenant x et y, plus bas que $f(h)$ ce qui est contraire à la définition de h ; d'où $s(x,y) = \alpha = \sigma(x,y)$, comme ce résultat peut être prouvé pour tout x et y dans Ω : on a donc bien $\phi \circ \psi(s) = s$.

- Montrons maintenant que $\psi \circ \phi((P,f)) = (P,f)$: soit $\phi((P,f)) = \sigma$ et $\psi(\sigma) = (P',f')$, il s'agit de montrer que $P \equiv P'$ et $f = f'$.

L'identité entre P et P' se déduit de la suite d'équivalences suivantes :

(a) $\left\{ h \in P \right\} \Leftrightarrow \left\{ \exists (i,j) \in h \times h : h = \{x \in \Omega / \forall y \in h\ \sigma(x,y) \leq \sigma(i,j)\} \right.$ (1)

ou $\left. \{\exists h' \text{ et } h'' \text{ dans P avec } h' \neq h \text{ et } h'' \neq h : h = h' \cap h''\} \right\}$ (2) (b)

$\Leftrightarrow h \in P'$ (c)

Montrons que (a) => (b)

$\phi((P,f)) = \sigma$ et $h \in P$ impliquent l'existence de i et j dans h tels que $\forall (x,y) \in h \times h$, $\sigma(x,y) \leq \sigma(i,j)$; d'après la proposition 2, on peut choisir i et j de façon qu'ils constituent les éléments extrêmes de la partie connexe

associée à h selon un ordre Θ compatible avec σ ; puisque σ est pyramidale on a bien $\sigma(i,j) \geq \sigma(x,y)$ $\forall x,y$ compris entre i et j selon l'ordre Θ : si h contient tous les éléments w de Ω tels que $\forall y \in h$ $\sigma(w,y) \leq \sigma(i,j)$ alors la condition (1) est satisfaite ; sinon, soit $w \notin h$ tel que $\sigma(w,y) \leq \sigma(i,j)$ $\forall y \in h$; si i se situe entre w et j au sens de Θ on a $\sigma(w,j) \geq \sigma(i,j)$ puisque σ est pyramidale ; or par définition de w on a : $\sigma(w,j) \leq \sigma(i,j)$ d'où $\sigma(w,j) = \sigma(i,j)$ (on fait bien sur le même raisonnement si c'est j qui se situe entre w et i) ; donc le palier h' de plus basse hauteur qui contient w et j est à la même hauteur que h ; comme il contient w et j, il contient tous les éléments intermédiarires, donc ceux compris entre i et j, donc h ; il en résulte que $h \subset h'$ et que $f(h) = f(h')$; par définition d'une hiérarchie indicée au sens large, il existe donc h_1 et h_2 distincts dans P : $h \neq h_1$ et $h = h_1 \cap h_2$.

Montrons maintenant que (b) => (a).

Il faut montrer que si (1) ou (2) est satisfait alors $h \in P$.

Supposons (1) vrai ; soit h' un palier de P de plus basse hauteur qui contient i et j (les 2 points les plus éloignés de h selon Θ) ; d'où $f(h') = \sigma(i,j)$ et $h \subset h'$; h' ne contient que des éléments de h car tout élément w extérieur à h est tel qu'il existe $y \in h$: $\sigma(w,y) > \sigma(i,j)$ sinon il serait dans h ; donc w et y ne peuvent être simultanément dans h' puisque $f(h') = \sigma(i,j)$; d'où $h' \subset h$, donc $h' \equiv h$ qui prouve que h est bien un élément de P. Si (2) est vrai h est dans P par définition d'une pyramide.

(b) <=> (c)

car par définition de $P(\alpha)$ on a :

$$h = \{x \in \Omega / \quad (i,j) \in h \times h : \forall y \in h \quad \sigma(x,y) \leq \sigma(i,j)\}$$

$$\Leftrightarrow h = \{x \in \Omega / \ \forall y \in h \ \sigma(x,y) \leq \alpha\} \quad \text{si } \alpha = \sigma(i,j)$$

(b) devient alors équivalent à (c) par définition de P'.

Reste à montrer que $f \equiv f''$; autrement dit, que $f'(h) = f(h)\ \forall h \in P$; en effet, pour tout h dans $P' \equiv P$ on sait que

$$f'(h) = \text{Min}\ \{\alpha \in P^{+}/h \in P(\alpha) \Rightarrow f'(h) = \sigma(i,j)$$

où $\sigma(i,j) = \text{Max}\ \{\sigma(\ell,k)/(\ell,k) \in h \times h\}$; d'autre part $\sigma(i,j)$ est la hauteur du plus bas palier de $P \equiv P'$ qui contient i et j ; comme i et j sont dans h on a donc $\sigma(i,j) \leq f(h)$; soient x et y les deux éléments de h les plus éloignés selon l'ordre Θ : comme σ est pyramidal on a $\sigma(x,y) \geq \sigma(i,j)$ d'où $\sigma(i,j) = \sigma(x,y)$; montrons que $\sigma(x,y) = f(h)$; si on avait $\sigma(x,y) \leq f(h)$, par définitions de σ, il existerait h' plus bas que h (i.e. $f(h') \leq f(h)$) tel que $\sigma(x,y) = f(h')$; or $h \subsetneq h'$ puisque (par définition de θ) h et h' contiennent tous les éléments compris entre x et y et que h ne contient que ceux-ci ; d'où $f(h) \leq f(h')$; il en résulte que $f(h) = f(h') = \sigma(x,y) = \sigma(i,j)$, d'où finalement $\forall h \in P\ \ f(h) = \sigma(i,j) = f'(h)$.

□

7.2. *Exemple de construction d'une pyramide à partir d'un indice pyramidal à l'aide de l'application ψ.*

L'indice pyramidal s est défini figure 13

a	b	c	d	e	
0	3	3	6	7	a
	0	2	5	5	b
		0	2,5	4	c
			0	4	d
				0	e

Figure 13

Les paliers de la pyramide P sont définis à l'aide de ψ ; on range les valeurs du tableau de la figure 13 par ordre décroissant et on associe à chaque valeur α le palier h de P formé de l'ensemble des éléments de Ω qui sont entre eux à une distance inférieure à α et dont les éléments les plus éloignés sont une dissimilarité égale à $f(h) = \alpha$. On obtient ainsi :

$\alpha = 7 \longrightarrow h_1 = \Omega$ car $f(h_1) = s(a,e) = 7$

$\alpha = 6 \longrightarrow h_2 = \{a\ b\ c\ d\}$ car $f(h_2) = s(a,d) = 6$

$\alpha = 5 \longrightarrow h_3 = \{b\ c\ d\ e\}$ car $f(h_3) = s(b,e) = 5$

$\alpha = 4 \longrightarrow h_4 = \{c\ d\ e\}$ car $f(h_4) = s(c,e)$

$\alpha = 3 \longrightarrow h_5 = \{a\ b\ c\}$ car $f(h_5) = s(a,c)$

$\alpha = 2{,}5 \longrightarrow h_6 = \{c\ d\}$

$\alpha = 2 \longrightarrow h_7 = \{b\ c\}$

A ces paliers, il faut ajouter le palier $h_8 = h_3 \cap h_2 = \{b\ c\ d\}$ de hauteur $f(h_8) = s(b,d) = 4$.

Pour trouver un ordre Θ compatible avec P, on utilise l'algorithme indiqué en 4.4 (on aurait pu aussi appliquer l'algorithme défini en 6.5).

Le premier niveau N_1 est formé d'une seule classe connexe $h_1\ h_2$; on a ensuite :

$N_2 \longrightarrow h_5\ h_8\ h_4$, $\quad N_3 \longrightarrow \{a\}\ h_7\ h_6\ \{e\}$

$N_4 \longrightarrow \{a\}\ \{b\}\ \{c\}\ \{d\}\ \{c\}$.

On retrouve l'ordre du tableau initial qui est généralement inconnu (par exemple, dans le cas d'un indice pyramidal induit par la pyramide indicée obtenue par un algorithme de CAP).

Ayant ainsi obtenu l'ordre Θ et les hauteurs de chaque palier on obtient la visualisation pyramidale indiquée figure 14 :

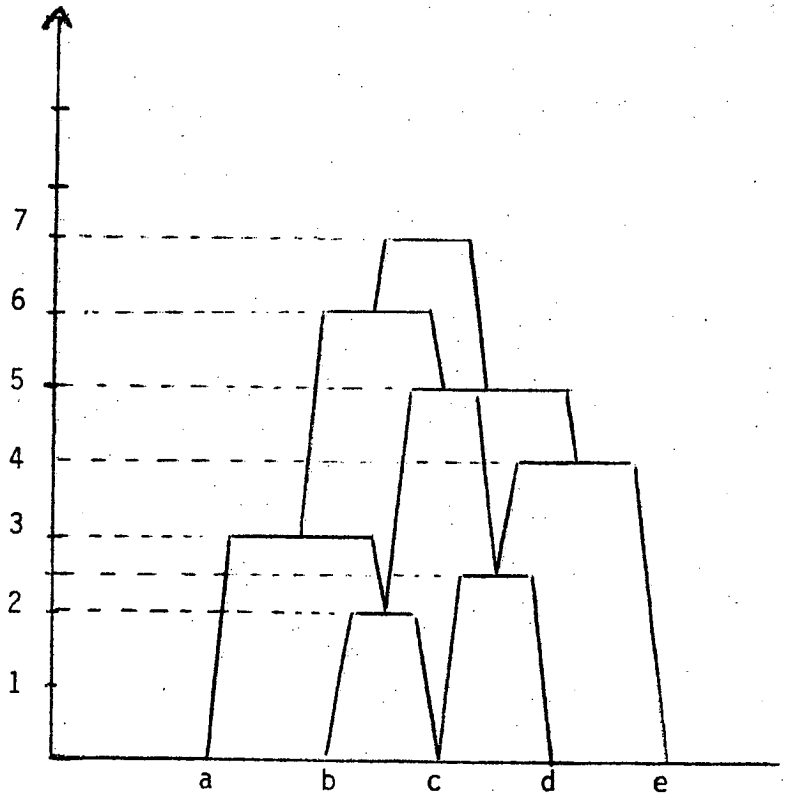


Figure 14

Nous verrons en 8.2 et 9.2 qu'il est également possible d'obtenir cette pyramide avec l'algorithme de CAP muni de l'indice d'agrégation du maximum.

7.3. Croisements, ordres et pyramides.

Définition

On dit qu'un ordre Θ donne lieu à un croisement pour une pyramide P s'il existe un palier de P qui n'est pas connexe selon Θ.

Il résulte facilement de cette définition, de la proposition 6 et des résultats obtenus au cours de la démonstration de la proposition 8, le résultat suivant où $\phi(P) = s$.

Proposition 9

Les propriétés suivantes sont équivalentes

. Θ est sans croisement pour la pyramide P.

. Θ est compatible avec P.

. Θ est compatible avec s.

. s est Robinson.

8. OPTIMISATION D'UNE CLASSIFICATION PYRAMIDALE.

8.1. *Quelques problèmes d'optimisation.*

Sachant qu'il y a équivalence entre une pyramide et un indice pyramidal, il est naturel de se poser le problème du choix de la meilleure pyramide sous la forme de la recherche de l'indice pyramidal ayant la meilleure adéquation avec la distance d donnée par l'utilisateur ; autrement dit : "Optimiser $\{\Delta(d,s)\ s \in \mathbb{S}\}$ où d est la distance donnée par l'utilisateur, $\mathbb{S}$ l'ensemble des indices pyramidaux et Δ une mesure de l'adéquation entre d et s ; dans le cas où $\mathbb{S}$ est réduit aux ultramétriques, différents auteurs ([5], [6], [14]) ont étudiés le critère :

$$\Delta_1(d,s) = \sum_{i,j} p(w_i)p(w_j)(d(w_i,w_j) - s(w_i,w_j))^2$$

Defays (1975) a étudié le critère suivant :

$$\Delta_2(d,s) = \sum_{i,j} |d(w_i,w_j) - s(w_i,w_j)|$$

Il serait intéressant de reprendre ces études et en particulier celle de Fichet (1980) (qui optimise le critère Δ_1 avec des contraintes d'ordre) dans le cas d'indices pyramidaux.

On peut remarquer que le problème de l'indice pyramidal le plus proche de d peut se ramener à celui de la recherche de la matrice Robinson la plus proche de la matrice de dissimilarité associée à d.

Hubert propose dans [18] et [19] plusieurs critères mesurant l'adéquation entre une matrice Robinson et une matrice de dissimilarité et des algorithmes permettant d'obtenir des solutions approchées.

Un problème d'optimisation classique en classification hiérarchique est celui de l'enveloppe supérieure des ultramétriques inférieures à un indice de dissimilarité d. On montre que cette enveloppe est une ultramétrique (la "sous-dominante"). Par contre l'enveloppe inférieure des ultramétriques supérieures

à d (la "sur-dominante") n'est pas nécessairement une ultramétrique.

On démontre également que la hiérarchie du saut minimum (i.e. celle qui utilise l'indice d'agrégation du saut minimum) est unique, ce qui n'est pas toujours le cas de la hiérarchie du saut maximum.

Le paragraphe suivant est consacré à l'étude des enveloppes supérieures et inférieures à d dans le cas des indices pyramidaux.

8.2. Sous-dominante et sur-dominante pyramidale.

La sous-dominante pyramidale d'un indice de dissimilarité d est l'enveloppe supérieure de l'ensemble des indices pyramidaux inférieurs à d ; on la note s_o; $\mathcal{S}$ étant l'ensemble des indices pyramidaux, on a donc :

$$\forall (w,w') \in \Omega^2, s_o(w,w') = \sup \{s(w,w')/s(w,w') \leq d(w,w'),\ s \in \mathcal{S}\}$$

$$\text{ou encore } s_o = \sup \{s/s \leq d,\ s \in \mathcal{S}\}$$

On définit de façon analogue la sur-dominante notée s_u et donc :

$$s_u = \inf \{s/s \geq d,\ s \in \mathcal{S}\}$$

Une pyramide notée P_{Min} est dite du "saut minimum" si elle est construite par l'algorithme de CAP muni de l'indice d'agrégation dit du "saut minimum pyramidal" :

$$\delta_{Min}(h_1,h_2) = \text{Min } \{d(x,y)/x \in h_1, y \in h_2, x \text{ et } y \notin h_1 \cap h_2\}$$

et indicée par $f(h) = \delta_{Min}(h_1,h_2)$ où h est l'ascendant de h_1 et h_2.

On note de même P_{Max} la pyramide dite du "saut maximum" construite avec l'indice du "saut maximum pyramidal" :

$$\delta_{Max}(h_1,h_2) = Max \{d(w,w')/w \in h_1,\ w' \in h_2\}$$

$$(\neq Max \{d(w,w')/w \in h_1,\ w' \in h_2,\ w \text{ et } w' \notin h_1 \cap h_2\})$$

On note $\phi(P,f)$ l'indice induit par une pyramide P indicée au sens large par f.

On note $s_{Min}(P)$ l'indice induit par une pyramide P indicée par $f_{Min}(h) = \delta_{Min}(h_1,h_2)$ si h est l'ascendant de h_1 et h_2 dans P.

De même $s_{Max}(P)$ est l'indice induit par une pyramide P indicée par $f_{Max}(h) = \delta_{Max}(h_1,h_2)$.

Nous allons démontrer l'ensemble des résultats suivants : (on utilise l'application ϕ qui a été définie dans la proposition 8)

- Une pyramide P étant donnée l'enveloppe supérieure de l'ensemble des indices pyramidaux $s = \phi(P,f)$ quand f est un indiçage de P qui varie de façon que s reste inférieur à d est s_{Min}.

Autrement dit : si $\phi(P,f) \leq d$ alors $\phi(P,f) \leq \phi((P,f_{Min}))$.

- On a un résultat analogue avec l'enveloppe supérieure :

$$\phi(P,f) \geq \phi(P,f_{Max})$$

si $\phi(P,f) \leq d$.

- L'enveloppe supérieure (resp. inférieure) de l'ensemble des indices pyramidaux inférieurs (resp. supérieurs) à d est d, ce n'est donc généralement pas un indice pyramidal.

- La hiérarchie du saut minimum notée H_{Min} (resp. du saut maximum notée H_{Max}) est contenue dans la pyramide du saut minimum P_{Min} (resp. du saut maximum P_{Max}).

L'ensemble de ces résultats peut être résumé dans la proposition suivante (où rappelons-le, $\mathcal{S}$ est l'ensemble des indices pyramidaux et d un indice de dissimilarité quelconque) :

Proposition 10

1) $\phi(P,f) \leq d \Rightarrow \phi(P,f) \leq \phi(P,f_{Min}) \leq d$

$\phi(P,f) \geq d \Rightarrow \phi(P,f) \geq \phi(P,f_{Max}) \geq d$

2) $\sup \{s \in \mathcal{S} / s \leq d\} = \inf \{s \in \mathcal{S} / s \geq d\} = d.$

3) Si d est un indice pyramidal alors $\phi(P_{Max},f_{Max}) = d$; si de plus P_{Min} est une pyramide indicée au sens strict alors $\phi(P_{Min},f_{Min}) = d$.

Démonstration

1) Il faut vérifier que $\phi(P,f) \leq d \Rightarrow \phi(P,f) \leq \phi(P,f_{Min}) \leq d$.

Montrons d'abord que $s_{Min} = \phi(P,f_{Min}) \leq d$.

Soit P une pyramide indicée par f_{Min} et h un palier de P admettant deux successeurs h_1 et h_2 ; par définition de l'indiçage de P on a $\forall (w_1,w_2) \in h_1 \times h_2$ avec w_1 et w_2 hors de $h_1 \cap h_2$: $s(w_1,w_2) = f_{Min}(h)$; en effet s'il existait un palier h' ne contenant que des éléments de $h_1' = h_1 - h_1 \cap h_2$ et $h_2' = h_2 - h_1 \cap h_2$ il serait à une hauteur supérieure à $f_{Min}(h) = \delta_{Min}(h_1,h_2)$ (strictement supérieure si h' ne contient pas de couple $(w_1',w_2') \in h_1' \times h_2'$: $s(w_1',w_2') = \delta_{Min}(h_1,h_2)$, égale sinon ; si h' contient des éléments différents de h_1' et h_2' compris entre w_1' et w_2', f(h') peut être inférieur à f(h) mais alors $f(h' \cup h_i) \leq f(h)$ (avec i=1 ou 2) par définition de s, et on ne peut avoir $f(h' \cup h_i) < f(h)$ car alors l'ascendant de h_i serait $h' \cup h_i$ et non h). Il en résulte que $\forall (w_1,w_2) \in h_1' \times h_2'$ on a :

$$s(w_1,w_2) = f_{Min}(h) = \delta_{Min}(h_1,h_2)$$

$$= \text{Min } \{d(x,y)/x \in h_1, y \in h_2, \text{ x et y } \notin h_1 \cap h_2\} \leq d(w_1,w_2)$$

d'où $s = \phi(P,f_{Min}) \leq d$.

Soit d'autre part w et w' :

$$d(w,w') = \text{Min } \{d(x,y)/(x,y) \in h_1' \times h_2'\} = \delta_{Min}(h_1,h_2).$$

Si $s = \phi(P,f) \leq d$, on a $\forall (w_1',w_2') \in h_1' \times h_2' \quad s(w_1',w_2') \leq s_{Min}(w_1,w_2) = d(w,w')$ car sinon on aurait $s(w,w') > d(w,w')$ (puisque $(w,w') \in h_1' \times h_2'$ et que s est égal à f(h) pour tous les couples qui sont dans $h_1' \times h_2'$) ce qui est contraire à $s \leq d$. Il en résulte que $s \leq s_{Min}$ et donc que $\phi(P,f) \leq \phi(P,f_{Min})$.

L'implication $\phi(P,f) \geq d \Rightarrow \phi(P,f) \geq \phi(P,f_{Max}) \geq d$ se démontre de façon tout à fait analogue.

2) Il faut vérifier que

$$\sup \{s \in \mathbb{S}/s \leq d\} = \inf \{s \in \mathbb{S}/s \geq d\} = d.$$

Montrons d'abord que $\sup \{s \in \mathbb{S}/s \leq d\} = d$.

Soit $s^* = \sup \{s \in \mathbb{S}/s \leq d\}$, il faut montrer que $\forall (w,w') \in \Omega \times \Omega$ $s^*(w,w') = d(w,w')$.

Soit (w,w') un couple quelconque de $\Omega \times \Omega$ on peut toujours construire une pyramide P munie d'un ordre compatible $\Theta = w_1, \ldots, w_n$ avec $w_1 = w$ et $w_n = w'$ et indicée par f de la façon suivante : la hauteur du plus haut palier $f(\Omega) = d(w,w')$, la hauteur des autres paliers inférieure à $d(w_i,w_j) = \text{Min } \{d(x,y)/(x,y) \in \Omega \times \Omega\}$: il suffit pour cela de construire le plus bas palier h qui contient w_2 et w_{n-1} à une hauteur $f(h) \leq d(w_2,w_{n-1})$. La pyramide indicée au sens large (P,f) ainsi obtenue induit par ϕ un indice s tel que $s(w,w') = d(w,w')$; de plus par construction $\forall (w_\ell,w_k) \in \Omega \times \Omega$ avec $(w_e,w_k) \neq (w_1,w_2)$ on a : $s(w_\ell,w_k) \leq d(w_i,w_j) \leq d(w_\ell,w_k)$ et $s(w_1,w_2) = d(w_1,w_2)$ d'où $s \leq d$.

On a donc $s^*(w,w') = \sup \{s(w_1,w_2)/s \in \mathbb{S}, s \leq d\} = d(w,w')$. Comme on peut faire le même raisonnement pour tous les couples (w,w') de Ω on a le résultat.

Pour démontrer que $s^* = \inf \{s \in \mathbb{S}/s \geq d\} = d$, on utilise un raisonnement analogue en choisissant cette fois-ci w et w' consécutifs selon l'ordre Θ; l'indice f est construit de la façon suivante : $f(h) = d(w,w')$ ou h est le plus bas palier de la pyramide P et contient w et w' ; les autres paliers sont situés à une hauteur supérieure à la plus grande distance notée $d(w_i,w_j)$ qui sépare un couple quelconque d'éléments de Ω. L'indice s induit par la pyramide indicée au sens large (P,f) est tel que $s(w,w') = d(w,w')$ et $\forall (w_\ell,w_k) \in \Omega \times \Omega$ avec $(w_\ell,w_k) \neq (w,w')$ on a $s(w_\ell,w_k) \geq d(w_i,w_j) \geq d(w_\ell,w_k)$ d'où finalement $s \geq d$ et $s(w,w') = d(w,w')$ impliquent que $s^*(w,w') = d(w,w')$ pour tous les couples de Ω.

3) Il faut montrer que $\phi(P_{Min},f_{Min}) = \phi(P_{Max},f_{Max}) = d$ si d est indice pyramidal.

Nous avons vu en 6.4 qu'après épuration, l'algorithme de la CAH donne une pyramide indicée au sens large donc $s = \phi(P_{Min},f_{Min})$ est un indice pyramidal si P_{Min} est épuré. Pour montrer que s=d, il suffit de vérifier que $\psi(d)=(P_{Min},f_{Min})$ puisque d'après la proposition 8 on aura $\phi \circ \psi(d) = d = \phi(P_{Min},f_{Min})$.

Soit $(P,f) = \psi(d)$; soient h_1 et h_2 deux paliers quelconques de P ayant le même ascendant h ; nous allons montrer que $f(h) = \delta_{Max}(h_1,h_2)$; en effet, soit le couple (w_i,w_j) tel que $d(w_i,w_j) = \text{Max } \{d(x,y)/(x,y) \in h_1 \times h_2\}$; la hauteur de h est égale à $d(w_i,w_j)$; en effet, elle ne peut être inférieure puisque nous savons que $f(h) = \text{Max } \{d(x,y)/(x,y) \in h \times h\}$ (page 34) ; elle ne peut être supérieure sinon il existerait par définition de ψ au niveau $\alpha = d(w_i,w_j)$ une partie connexe contenant h_1 et h_2 à une hauteur inférieure à celle de h et donc h ne serait pas l'ascentdant de h_1 et h_2. D'où finalement

$$f(h) = d(w_i,w_j) = \text{Max } \{d(x,y)/(x,y) \in h_1 \times h_2\} = \delta_{Max}(h_1,h_2)$$

Supposons que la pyramide $P = \psi(d)$ soit indicée au sens strict ; on a $\forall(w,w') \in h_1' \times h_2' = (h_1 - h_1 \cap h_2) \times (h_2 - h_1 \cap h_2)$ $f(h) = d(w,w')$ sinon il existerait un palier h' contenant des éléments de h_1' et h_2' à une hauteur inférieure à celle de h (si $f(h') \geq f(h)$, h' contiendrait h par définition de $P = \Psi(d)$) ; $h' \cap h_1'$ donnerait un palier à la même hauteur que h_1 et donc la pyramide P ne serait pas indicée au sens strict. Or si $\forall(w,w') \in h_1' \times h_2'$ $f(h) = d(w,w')$ on a en particulier

$$f(h) = \text{Min } \{d(w,w')/w \in h_1, w' \in h_2, w \text{ et } w' \notin h_1 \cap h_2\}$$

d'où $f(h) = \delta_{Min}(h_1,h_2)$.

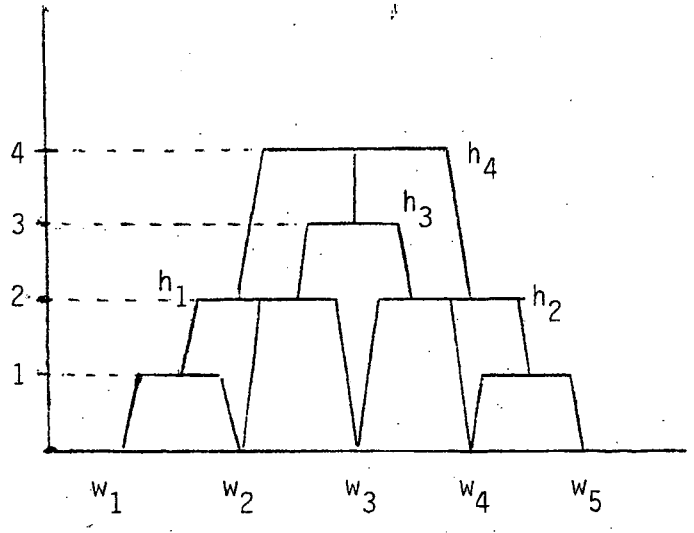
Remarques :

. $\psi(d) = (P,f)$ est certainement une pyramide indicée au sens strict si toutes les distances $d(w,w')$ $\forall(w,w') \in \Omega \times \Omega$ sont distinctes.

. Nous donnons figure 15 un exemple d'indice pyramidal d qui n'induit pas par ψ une pyramide du saut minimum puisque $h_4 = h_1 \cup h_2$ (avec $h_1=\{w_1\ w_2\ w_3\}$ et $h_2=\{w_3\ w_4\ w_5\}$) est tel que $4 = f(h_4) \neq \delta_{Min}(h_1,h_2) = d(w_2,w_4) = 3$

	w_1	w_2	w_3	w_4	w_5
w_1	0	1	2	4	4
w_2		0	2	3	4
w_3			0	2	2
w_4				0	1
w_5					0

indice pyramidal d



(P,f) = ψ(d)

Figure 15

. Remarquons enfin que H_{Min} (la hiérarchie du saut minimum) n'est pas nécessairement incluse dans P_{Min} (une pyramide du saut minimum) ; en effet, si l'indice de dissimilarité est

$$d = \begin{array}{c} \begin{matrix} a & b & c & d \end{matrix} \\ \begin{pmatrix} 0 & 4 & 3 & 5 \\ & 0 & 2 & 6 \\ & & 0 & 1 \\ & & & 0 \end{pmatrix} \end{array} \begin{matrix} a \\ b \\ c \\ d \end{matrix}$$

on obtient la hiérarchie H_{Min} et la pyramide P_{Min} données figure 16 ; on voit alors que le palier h = b, c, d appartient à H_{Min} mais pas à P_{Min}.

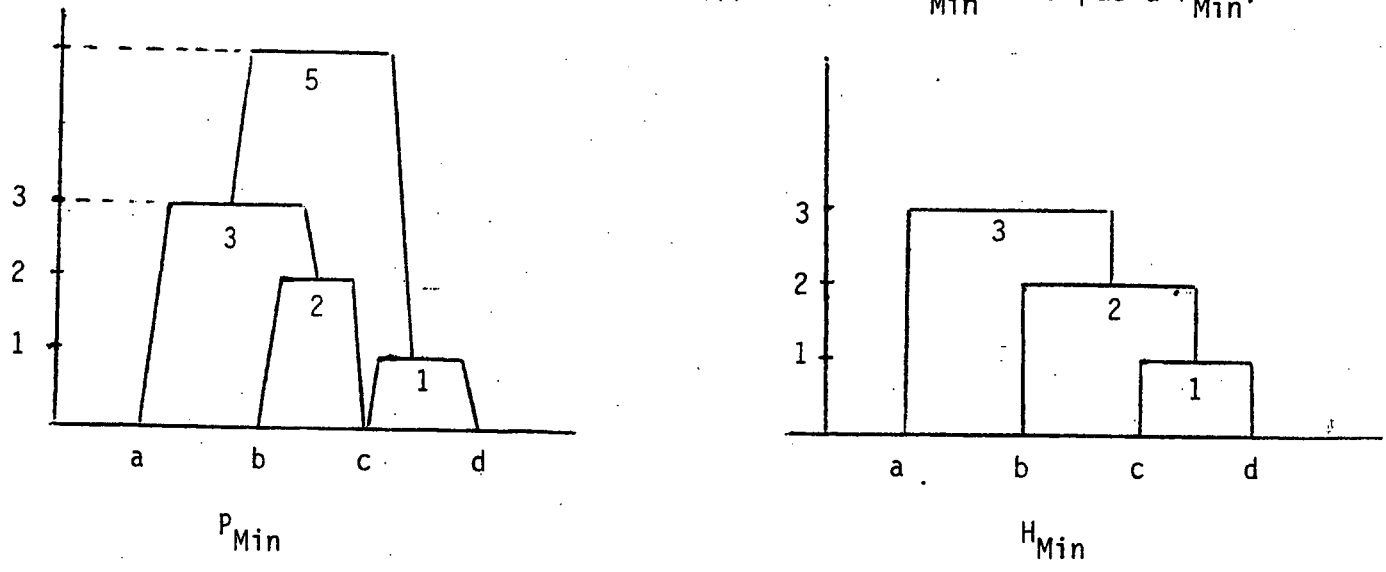


Figure 16

. On peut montrer aussi que H_{Max} (hiérarchie du saut maximum) n'est pas toujours contenue dans P_{Max} (la pyramide du saut maximum) ; cela se voit en prenant par exemple les 4 sommets d'un carré de côtés de longueur égale à l'unité.

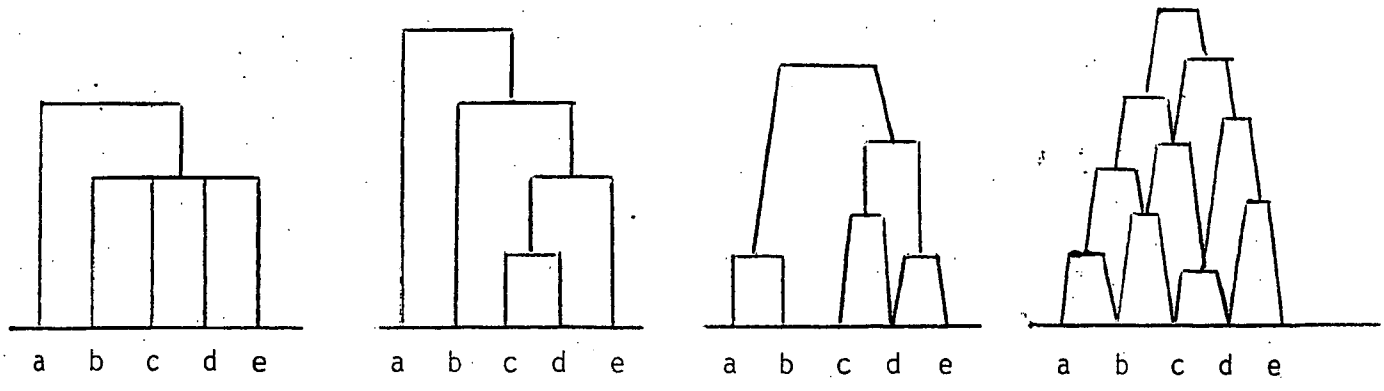
9. HIERARCHIES ET PYRAMIDES.

9.1. *Hiérarchies et pyramides saturées.*

Définition

Une hiérarchie ou une pyramide est saturée quand le nombre de ses paliers est maximum.

On donne figure 17 des exemples de hiérarchies et pyramides saturées ou non. Il résulte de cette définition qu'une hiérarchie sur n objets est saturée quand le nombre de ses paliers est égal à n-1 ; une pyramide est saturée quand le nombre de paliers est égal au nombre maximum de distances différentes entre objets, c'est à dire $\frac{n(n-1)}{2}$. Il y a donc $\frac{n}{2}$ fois plus de paliers dans une pyramide saturée que dans une hiérarchies saturée ; ainsi pour n = 200 il y a 100 fois plus de paliers dans la pyramide saturée et l'utilisateur aura du mal à interpréter une pyramide à $\frac{200 \times 199}{2}$ = 19.900 paliers ! Afin d'éviter cette difficulté on est conduit à se ramener à des pyramides non saturées.



Hiérarchie non saturée — Hiérarchie saturée — Pyramide non saturée — Pyramide saturée

Figure 17

9.2. *Construction de pyramides non saturées.*

On peut imaginer deux stratégies : i) partir d'une hiérarchie pour l'enrichir par des paliers qui la rende pyramidale sans créer d'inversions (i.e. un palier ne doit pas être plus haut qu'un palier qui le contient) ; ii) partir d'une pyramide saturée et supprimer des paliers inutiles parce que trop voisins ; voyons comment procéder dans ces deux cas.

i) "Pyramidisation" d'une hiérarchie :

Elle peut se faire en trois étapes : 1) on construit une hiérarchie à l'aide d'un indice d'agrégation δ : 2) on choisit un ordre Θ sans croisement (voir en 7.3 la définition de cette notion) pour cette hiérarchie ; 3) on considère toutes les classes consécutives selon Θ ; en commençant par les plus basses, on les réunit pour former un nouveau palier chaque fois que sa hauteur est inférieure à la hauteur du plus bas palier qui le contient ; si les paliers ainsi obtenus sont trop nombreux on utilise la stratégie ii).

Exemple

Considérons la matrice de dissimilarité de la figure 18.

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

Figure 18

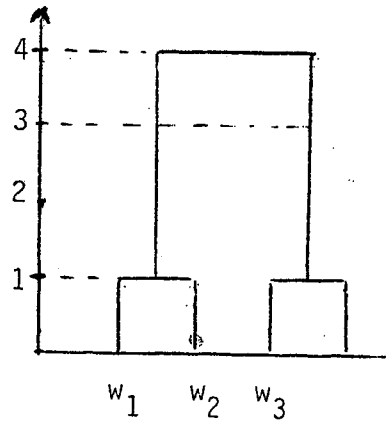


Figure 19

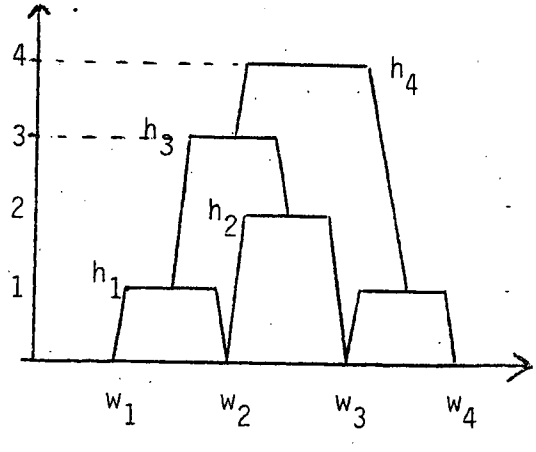


Figure 20

Si on choisit la distance du maximum, on obtient la hiérarchie de la figure 19. La pyramidisation de cette hiérarchie donne la pyramide de la figure 20. La pyramide ainsi obtenue contient deux paliers de plus que la hiérarchie : les paliers h_2 et h_3 qui peuvent successivement apparaître puisque leur hauteur est plus basse que celle du plus bas palier qui les contient : h_4.

ii) Hiérachisation d'une pyramide.

Une pyramide (construite par l'algorithme de CAP, par exemple) induit un indice pyramidal s et donc une matrice de dissimilarité $M(s,\Theta)$ où Θ est un ordre associé à s. En utilisant s on peut rendre chaque triangle i j k isocèle avec la base plus petite que les côtés en utilisant un nouvel indice de dissimilarité s' tel que

$$s'(w_i,w_k) = \text{Max}(s(w_i,w_j),\ s(w_j,w_k))$$

si $s(w_i,w_k)$ n'est pas le plus petit côté du triangle $w_i\ w_j\ w_k$ (au lieu de Max, on pourrait prendre le Min). Ainsi on transforme le triangle $w_1\ w_2\ w_3$ défini par la matrice de dissimilarité de la figure 18 de la façon suivante $s(w_1,w_2) = 1 \rightarrow s'(w_1,w_2) = 1$ $s(w_2,w_3) = 2 \rightarrow s'(w_2,w_3) = 3$ et $s(w_1,w_3) = 3 \rightarrow s'(w_1,w_3) = 3$.

Chaque fois qu'un triangle est rendu isocèle, on rapproche la pyramide d'une hiérarchie ; à la limite quand les $C_3^n = \frac{n!}{3!(n-3)!}$ triangles sont rendus isocèles avec la base plus petite que les côtés l'indice s' est une ultramétrique. Pratiquement on peut utiliser l'algorithme suivant :

- Algorithme de Hiérarchisation d'une Pyramide. (HDP)

(1) Comparer toutes les valeurs consécutives sur les lignes et les colonnes de la matrice $M(s,\Theta)$ et retenir le couple de valeurs dont l'écart est minimum.

(2) Remplacer ces deux valeurs par une valeur unique (la plus petite m, la plus grande M, ou toute valeur comprise entre m et M, $\frac{m+M}{2}$ par exemple), on

obtient ainsi un nouvel indice de dissimilarité noté s'.

③ Recommencer en ① en remplaçant s par s' tant que le pourcentage p de saturation de la pyramide définit par l'utilisateur n'est pas atteint.

Le nombre de valeurs distinctes contenues dans la matrice $M(s,\Theta)$ peut être généralement confondu avec le nombre de paliers de la pyramide. (Ce n'est pas vrai pour l'indice de proximité du minimum mais c'est vrai pour celui du maximum et la plupart des autres).

. Le pourcentage p peut donc être défini de la façon suivante : s'il y a $\frac{n(n-1)}{2}$ valeurs distinctes, la pyramide est saturée à 100% ; s'il y en a x la pyramide est saturée à $p = \frac{200\ x}{n(n-1)}\%$; remarquons qu'une hiérarchie est alors une pyramide saturée à $p = \frac{200(n-1)}{n(n-1)} = \frac{200}{n}$ %. Le choix de p par l'utilisateur lui permet de définir le "taux de saturation" désiré de la pyramide (compris entre 200/n et 100 s'il désire obtenir une pyramide qui ne soit pas réduite à une hiérarchie).

Proposition 11.

A chaque étape de l'algorithme HDP le nouvel indice créé s' est un indice pyramidal et le nouveau triangle formé est isocèle avec la base plus petite que les côtés.

Démonstration.

Remarquons d'abord que les longueurs des côtés de tout triangle sont des valeurs de la matrice $M(s,\Theta)$ qui se trouvent deux à deux sur la même ligne ou la même colonne (l'une d'elle se trouve nécessairement à l'intersection de cette ligne et de cette colonne) ; d'autre part, comme $M(s,\Theta)$ est Robinson, les lignes et les colonnes sont croissantes à partir de la diagonale principale dans la matrice triangulaire supérieure ; on est ainsi assuré qu'en remplaçant 2 valeurs : m et M consécutives (sur une même ligne ou une même colonne) par une valeur comprise entre m et M, la nouvelle matrice $M(s',\Theta)$ reste Robinson et donc que s' est pyramidale.

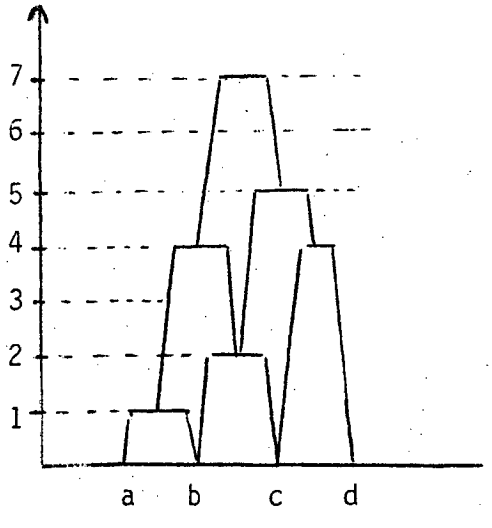
Chaque fois que l'on réalise une nouvelle égalité dans la matrice $M(s,\Theta)$, on crée un nouveau triangle isocèle avec la base plus petite (au sens large) que les côtés ; en effet, deux valeurs* consécutives s_{ij} s_{ij+1} sur une ligne (resp. s_{ij} s_{i+1j} sur une colonne) caractérise un triangle (i,j,j+1) (resp. (i,i+1,j)) dont le troisième côté (j,j+1) (resp. (i,i+1)) est de longueur $s_{jj+1} \leq s_{ij+1}$ (resp. $s_{ii+1} \leq s_{ij}$) puisque dans la matrice triangulaire supérieure $i < i+1 \leq j$ les colonnes (resp. les lignes) sont croissantes à partir de la diagonale principale. D'autre part $s_{jj+1} \leq s_{ij}$ (resp. $s_{ii+1} \leq s_{i+1j}$) puisque s_{ij} et s_{ij+1} (resp. s_{ij} et s_{i+1j}) sont les 2 valeurs (consécutives et même non consécutives puisque $M(s,\Theta)$ est Robinson) les plus proches par construction même de l'algorithme. Finalement, le plus petit côté du triangle (i,j,j+1) (resp.(i,i+1,j) est (j,j+1) (resp.(i,i+1)) ; il en ressort que l'égalisation effectuée donne aux deux plus grands côtés du triangle (i,j,j+1) (resp.(i,i+1,j) une même valeur supérieure au plus petit côté on obtient donc bien un triangle isocèle avec la base plus petite.

□

Remarque : Chaque fois que l'on crée une nouvelle égalité on n'augmente pas nécessairement le nombre de triangles isocèles en raison des triangles équilatéraux.

Exemple

On part de la pyramide saturée donnée figure 21 et de la matrice Robinson qui lui est associée.

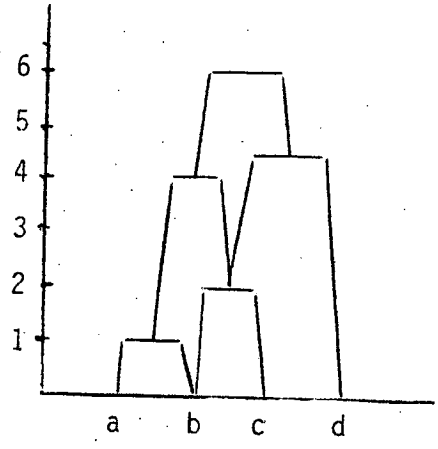


	a	b	c	d
a	0	1	4	7
b		0	2	5
c			0	4
d				0

Figure 21

* On note pour simplifier $s_{ij} = s(w_i,w_j)$

L'égalisation la plus avantageuse est obtenue dans le triangle b c d puisque s(d,b) - s(d,c) = 1 est le plus petit écart. On obtient ainsi voir figure 22 une nouvelle pyramide où les paliers (b c d) et (c d) sont réunis en un seul à la hauteur 4.5 ; on remarque alors que l'arête qui relie c au palier b c d devient inutile et est supprimée.



	a	b	c	d
a	0	1	4	7
b		0	2	4.5
c			0	4.5
d				0

Figure 22

9.3. *Une règle générale pour la suppression d'arêtes inutiles dans une pyramide.*

Quand des égalités apparaissent de façon consécutive dans une matrice de dissimilarité pyramidale (i.e. une matrice Robinson) que l'on désire représenter par une pyramide, il apparaît dans cette pyramide des arêtes inutiles ; la règle de suppression générale est décrite par les trois schémas de la figure 23.

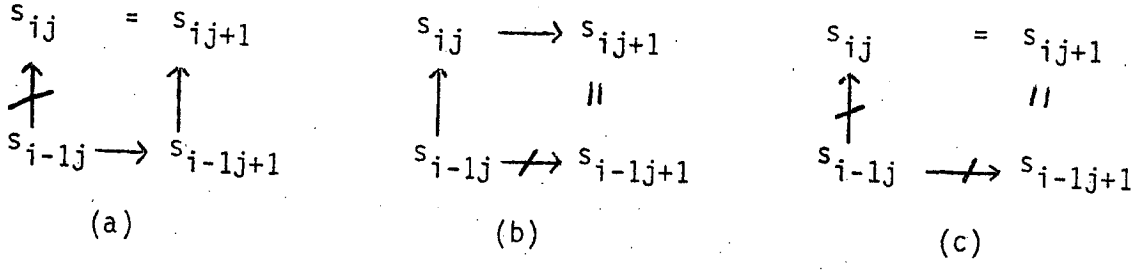


Figure 23

où la flèche barrée indique l'arête à supprimer ; ainsi, dans le schéma (a) où $s_{ij} = s_{ij+1}$ il faut supprimer l'arête qui relie le plus bas palier noté h_{i-1j}

contenant w_{i-1} et w_j au plus bas palier contenant w_i et w_j ; de même dans le cas (b) où $s_{i-1j+1} = s_{ij+1}$, c'est l'arête qui relie le palier h_{i-1j+1} au palier h_{ij+1} qu'il faut supprimer.

Enfin, dans le cas (c) on supprime l'une des deux arêtes qui relient h_{i-1j} à h_{ij} et h_{i-1j} à h_{i-1j+1}.

Exemple

L'indice pyramidal et la pyramide associée sont données figure 24

$$s = \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 0 & 1 & 2 & 2 & 4 \\ & 0 & 1 & 2 & 4 \\ & & 0 & 1 & 4 \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix} \end{array} \begin{array}{c} a \\ b \\ c \\ d \\ e \end{array}$$

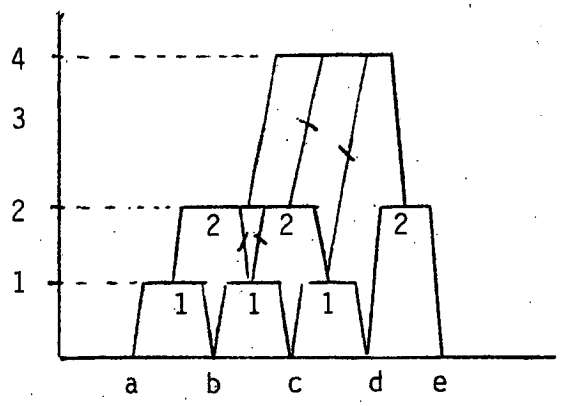


Figure 24

Après suppression des arêtes inutiles, on obtient la pyramide de la figure 25.

on supprime les arêtes reliant

h_{bc} à h_{bd}

h_{bd} à h_{bc}

h_{cd} à h_{ce}

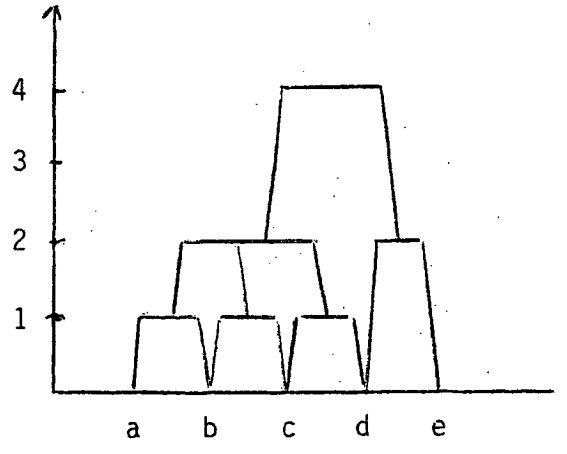


Figure 25

Remarque : Un indice pyramidal s étant donné, l'algorithme de CAP du lien maximum permet de construire la pyramide qui induit (par ϕ) justement cet indice pyramidal (voir en 8.2, la proposition 10) ; il en résulte que dans le cas où s est ultramétrique, en appliquant la règle de suppression, la représentation visuelle de la pyramide est une hiérarchie.

10. DIVERSES REPRESENTATIONS D'UN INDICE PYRAMIDAL.

10.1. Indices pyramidaux et distances :

On sait qu'il est toujours possible d'associer à tout indice de dissimilarité une distance respectant le même ordre sur les couples (voir par exemple [8], chap. 1) ; un indice pyramidal étant un indice de dissimilarité particulier, on a le résultat suivant :

Proposition 12

A tout indice pyramidal on peut associer une distance qui satisfait au même ordre sur les couples.

10.2. Représentation de 2n-3 distances exactes, respectant le même ordre qu'un indice pyramidal.

Soit $w_1 \ldots w_n$ un ordre Θ sur Ω compatible avec un indice pyramidal s, et soit d une distance associée satisfaisant au même ordre sur les couples que s; on peut toujours associer à w_1 les sommets w_2 et w_3 en les représentant dans le plan de façon à respecter exactement les distances $d(w_1,w_2)$ et $d(w_1,w_3)$. De même à w_i on peut associer les distances exactes $d(w_i,w_{i+1})$ et $d(w_i,w_i+2)$; ayant représenté dans le plan successivement les triangles $w_1\ w_2\ w_3$, $w_2\ w_3\ w_4 \ldots w_{n-2}\ w_{n-1}\ w_n$ on s'aperçoit que l'on a respecté de façon exacte 2n-3 distances (à chaque élément on peut associer 2 distances exactes sauf à l'avant dernier (1 seul) et au dernier (aucun)). On trouvera plus de détails concernant cet algorithme dans [22].

Remarquons que les hauteurs des paliers d'une pyramide saturée permettent de respecter $\frac{n(n-1)}{2}$ dissimilarités pyramidales.

10.3. Représentation polygonale.

Nous avons montré en 5.2 (Proposition 6) que si s est un indice pyramidal et Θ un ordre compatible avec s alors tout couple d'éléments (w,w') compris

(au sens large) selon Θ entre w_i et w_j est tel que $s(w_i,w_j) \geq s(w,w')$. Il en résulte que tous les polygones dont l'ordre des sommets sur le périmètre respecte Θ ont leurs diagonales de longueur supérieure aux côtés et diagonales intermédiaires (i.e. aux côtés et diagonales qui réunissent des sommets compris selon Θ entre ceux qui sont sous-tendus par la diagonale).

Exemple

$\Omega = \{w_1,w_2,w_3,w_4,w_5,w_6\}$: Θ est l'ordre $w_1\ w_2\ w_3\ w_4\ w_5\ w_6$.

On suppose que les w_i sont des points du plan répartis selon la figure 26.

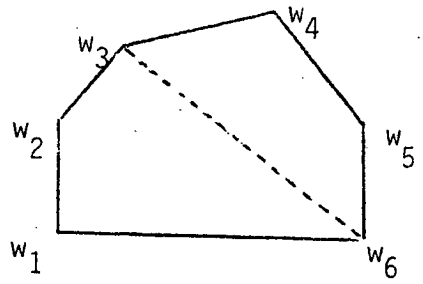


Figure 26

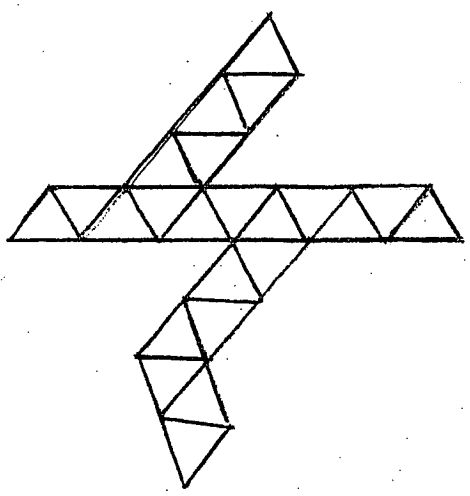
La diagonale $w_3\ w_6$ a une longueur supérieure aux côtés intermédiaires : $w_3\ w_4$, $w_4\ w_5$, $w_5\ w_6$ et aux diagonales intermédiaires $w_3\ w_5$, $w_6\ w_4$.

On peut toujours représenter de façon exacte les longueurs de n-1 côtés du polygone et on ne peut généralement pas respecter les longueurs de plus de n-2 diagonales.

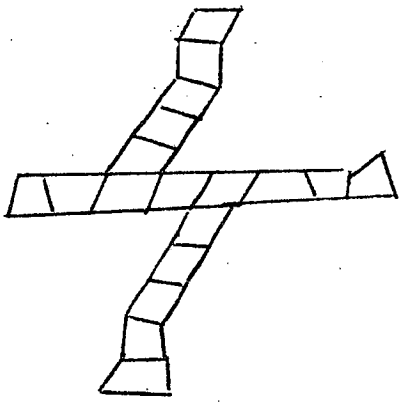
10.4. Représentation par des "arbres épais" ou "guirlandes".

Nous savons que l'on peut associer la notion d'abre de longueur minimum à celle de hiérarchie du saut minimum (voir [8] chapitre 2 § 2) ; il est donc naturel de chercher à représenter sous forme d'un graphe une pyramide du saut minimum.

On appelle "guirlande" un ensemble de polygones "connexes" et sans "cycles"; "connexe" signifie que deux polygones quelconques peuvent être reliés par une suite (appelée "chaîne") de polygones ayant successivement un côté commun et un seul ; un "cycle" est une chaîne de polygones dont le premier et le dernier élément sont identiques. Si tous les polygones sont de même type on dit que la guirlande est de ce type. On obtient ainsi des guirlandes triangulaires, quadrilatères etc... voir figure 27.



Guirlande triangulaire



Guirlande quadrilatère

Figure 27

Une guirlande triangulaire est obtenue en utilisant les deux premiers niveaux d'une pyramide P ; si Θ défini par la suite $w_1 \ldots w_n$ est un ordre compatible avec P, la guirlande triangulaire induite est une "chaîne de triangles" définie par la suite de triangles suivants :

$$w_1\ w_2\ w_3,\ w_2\ w_3\ w_4,\ \ldots,\ w_{n-3}\ w_{n-2}\ w_{n-1},\ w_{n-2}\ w_{n-1}\ w_n.$$

Les "greffes" voir figure 27,peuvent être obtenues du fait que la pyramide est séparée en plusieurs parties connexes ; le premier sommet d'une partie connexe pouvant être plus proche (au sens de l'indice de dissimilarité initial) des sommets d'une arête qui n'est pas nécessairement la dernière arête de la partie connexe qui précède.

En utilisant en 10.1 et 10.2 on peut remplacer l'indice de dissimilarité par une distance euclidienne ; il est alors possible de représenter dans le plan la guirlande triangulaire en respectant de façon exacte les longueurs des 2n-3 côtés des triangles.

On peut obtenir de même une guirlande quadrilatère en utilisant l'ordre Θ induit par P ; on définit ainsi la suite de quadrilatères (qui satisfont à la propriété indiquée en 10.3) suivante :

$$w_1\ w_2\ w_3\ w_4,\ w_3\ w_4\ w_5\ w_6,\ \ldots,\ w_{n-5}\ w_{n-4}\ w_{n-3}\ w_{n-2},\ w_{n-3}\ w_{n-2}\ w_{n-1}\ w_n.$$

(il peut se produire que le dernier quadrilatère soit en fait réduit à un triangle). Les greffes s'obtiennent de façon analogue en cas des guirlandes triangulaires.

Le nombre de distances exactes dans le cas d'une guirlande de quadrilatères est de l'ordre de $4 + \frac{(n-4)}{2} \times 3 = \frac{3}{2}n - 3$; si les longueurs des diagonales sont respectées il est de l'ordre de $\frac{5n}{2} - 4$.

Il serait intéressant d'étendre la théorie des arbres de longueurs minimum concernant leur lien avec les hiérarchies (voir [8]) à l'étude de guirlandes de longueur minimum et de leur lien avec les pyramides.

10.5. Représentation d'un indice pyramidal par une courbe.

Nous avons vu en 10.3 que l'on peut associer à un indice pyramidal s un polygone P particulier puisque chacune de ses diagonales $w_i\ w_j$ est de taille supérieure aux diagonales et côtés qui relient des sommets compris entre w_i et w_j selon un ordre Θ associé à s ; d'où l'idée de chercher à donner une représentation visuelle approchée de ce polygone, sous forme d'une courbe plane.

Le problème général peut s'énoncer sous la forme suivante :

Trouver une courbe C du plan (voir figure 28), représentée par une application f :

$$D(f(w_i),\ f(w_j)) = s(w_i,\ w_j)$$

où $f(w_i) = z_i = (x_i,\ y_i)$ est un point du plan associé à l'individu $w_i \in \Omega$; f est une application de $\mathbb{R}^p$ dans $\mathbb{R}^2$ qui dépend de m coefficients (f peut être polynomiale par exemple) ; s est l'indice de dissimilarité donné ; D est une distance à choisir (euclidienne, L_1, ...)

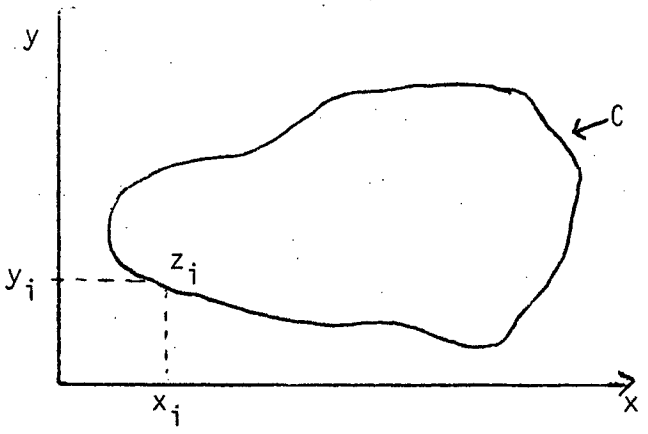


Figure 28

Comme s est connu et card $\Omega = n$, on a un système non-linéaire de $\frac{n(n-1)}{2}$ équations à $2n + m$ inconnues à résoudre (car il faut trouver les 2n éléments $(x_i,\ y_i)$ et les m coefficients qui définissent f).

Il y a donc plus d'équations que d'inconnues, on se ramène donc à résoudre le problème d'optimisation non linéaire suivant :

$$\text{Min}\{\sum_{i,j} (s(w_i,w_j) - F(w_i,w_j,a))^2 / a \in \mathbb{R}^m\}$$

où $F(w_i,w_j,a) = D(f(w_i),f(w_j))$ et a représente les m coefficients dont f dépend.

On peut simplifier le problème en cherchant la courbe plane C représentant l'application f :

(1) $\quad |f(x_i) - f(x_j)| = s(w_i,w_j) \qquad \forall w_i,w_j \in \Omega$

où f est une application $\mathbb{R} \to \mathbb{R}$, les x_i représentant les individus $w_i \in \Omega$. (voir figure 29) ; ils sont ordonnés à intervalles égaux sur l'axe des abcisses en respectant un ordre Θ compatible avec s. Si f est une application dépendant de m paramètres (polynome de degrè m-1, par exemple) avec $m < \frac{n(n-1)}{2}$ on cherche d'abord les $y_i = f(x_i)$ qui satisfont au mieux la relation (1) ; f est ensuite calculé par interpollation.

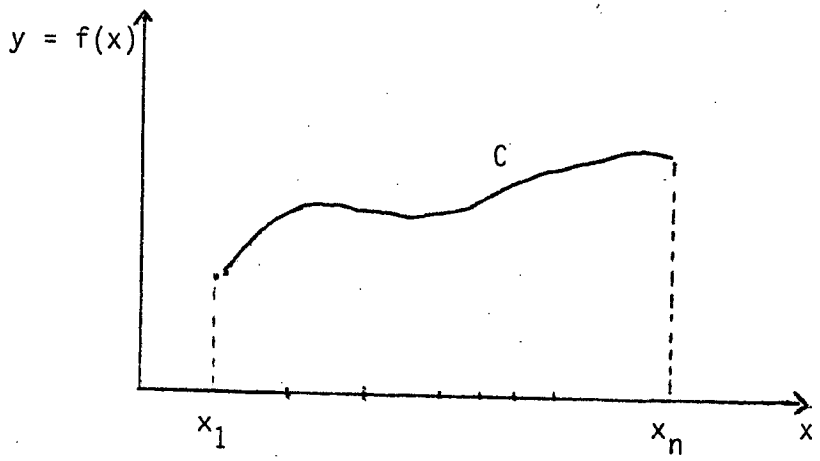


Figure 29

Dans le cas où l'on dispose de plusieurs tableaux de données que l'on désire comparer, on peut utiliser les courbes C mais cela pose le problème intéressant d'un ordre "consensus"* compatible avec plusieurs dissimilarités pyramidales.

FIN

* Pour plus de précision concernant la notion de consensus, le lecteur pourra se reporter à [10].

CONCLUSION

En approfondissant les liens existants entre ordres et ultramétriques on a débouché sur une théorie qui généralise celle des hiérarchies et permet de représenter de façon visuelle des classes recouvrantes ; les pyramides induisent des ordres sur les individus en nombre plus restreints que les hiérarchies (2 pour une pyramide binaire saturée, 2^{n-1} pour une hiérarchie binaire). Pour construire une pyramide on peut utiliser un algorithme de CAP qui induit un ordre sur Ω : on peut aussi chercher l'indice pyramidal s le plus proche de l'indice de dissimilarité d et en déduire la pyramide de la CAP avec l'indice d'agrégation du maximum (qui induit justement cet indice s) ; on peut enfin "pyramidiser" une hiérarchie.

La difficulté due au risque d'un trop grand nombre de paliers dans la pyramide est levée grâce à la possibilité de hiérarchisation et a la proposition 11 qui permet à l'utilisateur de préciser le taux de saturation désiré. En "pyramidisant" une hiérarchie on peut aussi obtenir un nombre raisonnable de paliers dans la pyramide associée.

De nombreuses directions de recherche restent ouvertes : étude des structures pyramidales dans le cas où l'existence d'un ordre compatible n'est pas imposé ; l'étude du problème des inversions dans le cas des pyramides et ses conséquences en "inférence pyramidale" devrait déboucher sur des indices d'agrégation plus souples et plus précis que ceux obtenus en inférence hiérarchique*. Il serait intéressant de vérifier que les conditions nécessaires pour obtenir une pyramide sans déformation en utilisant l'algorithme accéléré des voisins réciproques sont les mêmes que dans le cas hiérarchique . Il y a aussi beaucoup à faire pour la recherche d'une pyramide optimale au sens d'un critère donné (i.e. chercher l'indice pyramidal le plus "proche" d'un indice de dissimilarité donné).

Les liens avec l'analyse factorielle sont à approfondir : on peut bien sûr représenter les classes empiétantes (obtenues en coupant une pyramide à un niveau "significatif") sur le plan d'une analyse factorielle ; mais on peut aussi poser le problème suivant : le 1er axe d'une analyse factorielle induit un indice pyramidal, un ordre compatible étant défini par la position des projections des individus sur cet axe et l'indice pyramidal par la distance deux à deux de ces

projections. Quel est le degré de proximité de cet indice avec la dissimilarité initiale ?

Les arbres épais et les guirlandes posent également de nombreux problèmes ouverts : comment obtenir une guirlande de longueur minimale ? Comment la représenter ? Peut-on étendre à ce cas les algorithmes de Prim et Kruskal ? Ayant plusieurs tableaux de données caractéristiques du même ensemble d'individus, nous avons vu que par l'intermédiaire d'indices pyramidaux on peut leur associer une courbe du plan ; mais cela pose tout un autre champ de recherche : celui des consensus entre pyramides, etc...

Prolongement naturel de mes travaux sur les ordres et ultramétriques (voir [10] et [12]), le tableau 1 qui contient l'idée de base de ce travail a constitué le dernier transparent que j'ai présenté pour la première fois lors de ma conférence devant l'auditoire international réuni par les professeurs J.Geoffroy et M. Tenenhaus du petit séminaire qu'ils ont organisé à l'ISUP (juillet 83), et qui a précédé les journées internationales de classification automatique et de Psychométrie à Jouy en Josas.

En ce qui concerne l'avenir, P. Bertrand qui a eu la gentillesse de relire cet article lors de sa rédaction définitive (en permettant ainsi d'épargner au lecteur un nombre appréciable de coquilles !) va poursuivre ce travail dans le cadre d'une thèse de 3° cycle ; sur le plan théorique, il va entre autres étudier les pyramides plus générales qui ne satisfont pas nécessairement à la condition d'ordre (4° condition de la définition d'une pyramide) ; sur le plan pratique il va réaliser un programme général permettant d'obtenir une représentation visuelle, plus ou moins pyramidal ou hiérarchique, suivant le taux de saturation désiré par l'utilisateur.

BIBLIOGRAPHIE

[1] ADANSON M. (1757). "Histoire naturelle du Sénégal". Bauche ; Paris.

[2] ARABIE P. et CAROLL J.D. (1980). "MAPCLUS : a mathematical programming approach to fitting the ADCLUS model".

[3] BENZECRI J.P. et coll. (1973). "L'analyse des Données, Tome 1 La Taxonomie", Dunod.

[4] CAROLL J.D. et PRUZANSKY S. (1975). "Fitting of hierarchical tree structure", US-Japan Seminar on Multidimensional Scaling, University of California at San Diego.

[5] CHANDON J.L, LEMAIRE J., POUGET J. (1980) "Construction de l'ultramétrique la plus proche d'une dissimilarité", RAIRO, 14, 2 pp. 157-170.

[6] DEFAYS D.(1975) "Recherche des ultramétriques à distance minimum d'une dissimilarité donnée". Bull de la soc. Royale des sciences de Liège, 44, 5-6, pp. 330-343.

[7] DIDAY E. et coll. (1979). "Optimisation en classification automatique", INRIA, Rocquencourt 78150 (France).

[8] DIDAY E., LEMAIRE J., POUGET J., TESTU F. (1982). "Eléments d'analyse des données", Dunod.

[9] DIDAY E. (1982). "Problèmes d'inversions en classification hiérarchique, application à la recherche adaptative d'ultramétriques", Revue de statistiques appliquées, vol. 2.

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

[11] DIDAY E. MOREAU J.V. (1984) "Learning hiérachical clustering from examples" Rap. de recherche INRIA n° 289.

[12] DIDAY E. (1983). "Croisements, ordres et ultramétriques", Mathematiques des Sciences humaines ; 21ème année, n°83 pp. 31-54

[13] DIDAY E. (1982). "Crossing order and ultramétrics" Compstat - Proceedings in Computer Statistics - Physica - Verlag - Vienne.

[14] FICHET B. (1981). "Sur des approximations d'indices de dissimilarité via les représentations euclidiennes et hiérarchiques", Revue de l'ASU, statistiques et Analyse des données, Vol. 2.

[15] HARTIGAN J.A. (1975). "Clustering Algorithm", Wiley

[16] HARTIGAN J.A. (1977). "Clustering as modes" First International Symposimum on data analysis and Informatics.

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

[18] HUBERT L. (1982). "Inference procedures for the evaluation and comparing of proximity matrices", Graduate School of Education UCLA.

[19] HUBERT L. (1974). "Some applications of graph theory to clustering", Psychometrica 1974, 39, pp. 283-309.

[20] J. JUAN (1982) "Le programme Hivor de classification ascendante hiérarchique selon les voisins réciproques et le critère de la variance " cahiers d'analyse des données, Vol. 7 n° 4 pp. 5 - 25.

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

[22] LEE R.C.T, SLAGLE J.R., BLUM H. (1977) "A triangulation méthod for sequential mapping of points from N-space to two-space" IEEE Trans. on Computers, Mars 77 pp.288-292.

[23] MONJARDET B. (1980). "Théorie des graphes et Taxonomie mathématique", in Regards sur la théorie des graphes, Presses polytechniques Romandes, LAUSANNE, pp. 111-125.

[24] ROHLF F. (1975). "A new approach to the computation of the Jardine-Sibson B_k clusters", Computer Journal, 18 pp. 164 - 168.

[25] SHEPARD R., ARABIE P. (1979). "Additive clustering : Representation of Similarities as Combinations of Discrete Overlapping Properties", Psychological Review, Vol. 86, N° 2, pp. 87-123.

[26] SNEATH P., SOKAL R. (1973). "Numerical Taxonomy", Freeman.

ANNEXE

Quelques propriétés concernant les différents types de compatibilité.

Dans [12] on démontre entre autres que si δ est une ultramétrique, il existe toujours un ordre (non unique) tel que δ et Θ soient faiblement compatibles ; cet ordre est sans croisement (voir figure 30) pour la hiérarchie induite par l'ultramétrique.

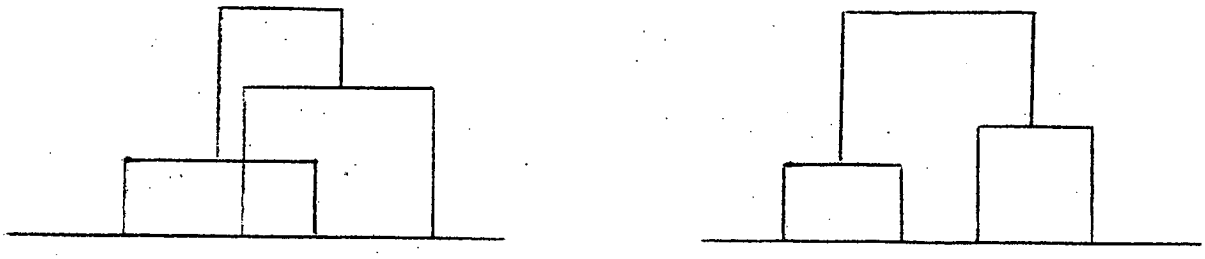
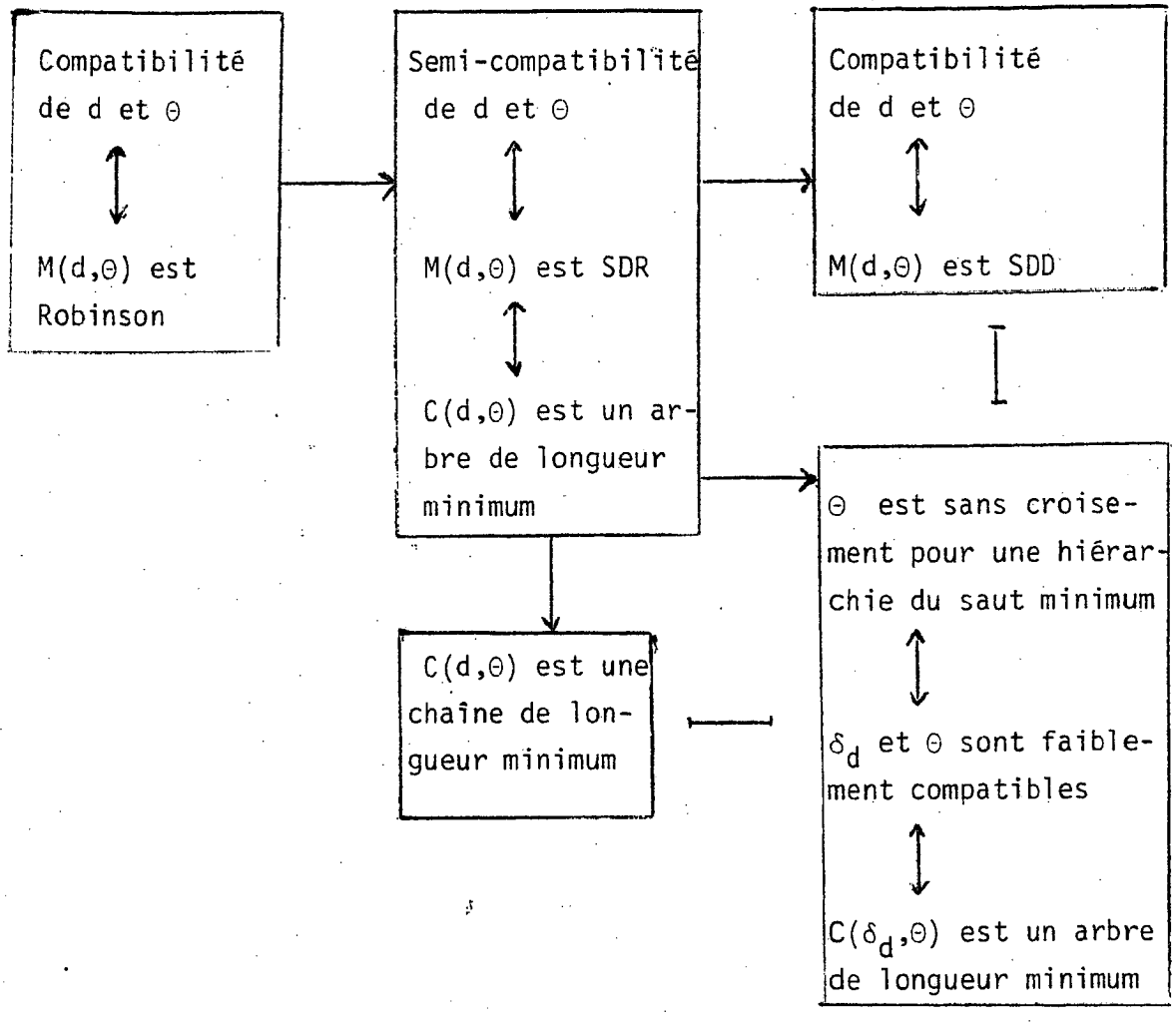


Figure 30

On note $C(d,\Theta)$, la chaîne dont les sommets (les éléments de Ω) sont ordonnés selon un ordre Θ, le poids de chaque arête $w_i\ w_{i+1}$ étant $d(w_i,w_{i+1})$.

Le tableau 2 résume un ensemble de résultats qui sont donnés dans [12].



A → B : A implique B.

A ⟶ B : aucune implication entre A et B

A ↔ B : A implique B et B implique A.

TABLEAU 2

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique