REVUE DE STATISTIQUE APPLIQUÉE

GILDAS BROSSIER

Étude des matrices de proximité rectangulaires en vue de la classification

Revue de statistique appliquée, tome 34, n° 4 (1986), p. 43-68

<http://www.numdam.org/item?id=RSA_1986__34_4_43_0>

ÉTUDE DES MATRICES DE PROXIMITÉ RECTANGULAIRES EN VUE DE LA CLASSIFICATION

Gildas BROSSIER

U.E.R. des Sciences et techniques, Université Rennes 2 Haute Bretagne

RÉSUMÉ

A partir d'un tableau de proximité rectangulaire nous cherchons à représenter simultanément les individus et les variables sur un arbre hiérarchique ou sur un arbre additif.

Pour cela nous étudions les propriétés métriques des tableaux rectangulaires de dissimilarités et nous mettons en évidence les conditions nécessaires et suffisantes pour qu'ils puissent être représentés par un arbre additif.

Mots clés : **Tableaux rectangulaires, Représentation simultanée, Arbres additifs, Classification.**

D'une façon traditionnelle la donnée utilisée en entrée des algorithmes de classification automatique est une matrice de dissimilarité définie sur les éléments d'un ensemble E. C'est-à-dire que l'on considère en donnée une relation de E dans E et on cherche à construire une représentation arborée de cette relation représentant les liens unissant les éléments de E entre eux.

Or, dans certains cas, la donnée se présente sous la forme d'une relation entre deux ensembles, que nous noterons alors A et X (souvent A est l'ensemble des individus et X celui des variables).

La démarche classique consiste à déduire de cette relation sur $A \times X$ une relation sur $A \times A$ ou $X \times X$ et à la représenter. Concrètement on calcule à partir du tableau défini sur $A \times X$ un tableau de dissimilarité ou de distance sur $A \times A$ (ou sur $X \times X$) que l'on cherchera à représenter par un arbre hiérarchique ou un arbre additif.

En procédant ainsi on perd la relation sur $A \times X$ pour ne considérer que deux relations sur $A \times A$ et $X \times X$. On est passé d'une relation sur deux ensembles à deux relations sur un ensemble. Dans le cas où la représentation cherchée est du type euclidien on arrive à reconstruire une représentation simultanée qui peut représenter la relation initiale sur $A \times X$. Mais dans le cas où la représentation cherchée est du type arboré, on ne sait plus faire coïncider les deux représentations, celle obtenue sur A et celle obtenue sur X.

Beaucoup d'études ont déjà été faites sur ce thème proposant soit de croiser les partitions ou les hiérarchies obtenues chacune sur un ensemble, soit de chercher des classifications simultanées. On trouvera une étude bibliographique dans GOVAERT (1983).

L'approche qui est présentée ici découle du « unfolding metric problem » (problème de la métrique dépliée) introduite initialement par

COOMBS (1950) dans le cadre euclidien et reprise récemment par FURNAS (1980) dans le cas ultramétrique. Cette approche, permet de donner une réponse au problème posé , à savoir représenter la relation initiale sur A × X et les relations sur A et sur X.

Dans cette approche on considère que la donnée initiale est une relation sur A × X qui à chaque élément « a » de A et à chaque élément « x » de X, associe une valeur mesurant leur « lien », ressemblance, distance, similarité ou dissimilarité.

Tous les tableaux A × X ne peuvent être vus comme décrivant de cette façon la relation entre A et X. Mais beaucoup de tableaux peuvent l'être ou devraient l'être.

C'est notamment le cas des tableaux de préférence où chaque individu (ensemble A) exprime sa préférence plus ou moins grande pour certains objets (ensemble X). C'est aussi le cas des tableaux de contingences où à chaque modalité « a » de la variable A et à chaque modalité « x » de la variable X on associe le nombre n_{ax} d'individus possédant simultanément ces deux modalités. Alors n_{ax} mesure la proximité entre la modalité « a » et la modalité « x ».

D'une façon générale nous considérons que la donnée initiale s'exprime sous la forme d'un tableau rectangulaire T sur A × X dont les éléments T_{ax} expriment une mesure de dissimilarité entre a et x. Si la donnée est exprimée en terme de proximité ou de similarité, on se remenera au cas précédent en considérant une transformation simple du type $(K - T_{ax})$, K étant une constante adéquate.

Le principe de la méthode est simple : il s'agit de considérer le tableau T défini sur A × X comme étant un sous-tableau extrait d'une matrice carrée D définie sur $(A \cup X \times A \cup X)$ et possédant les propriétés métriques qui nous intéressent : être une distance, une ultramétrique, une quadrangulaire.

Considérant le problème résolu (chercher D possédant certaines propriétés et contenant le tableau T), la représentation arborée de D est une représentation arborée simultanée des deux ensembles A et X d'où l'on peut extraire des sous-arbres représentant la classification sur A et sur X.

Trois types de problèmes sont donc à étudier :

— les conditions nécessaire et suffisantes que doit vérifier une matrice rectangulaire T pour que l'on puisse la considérer comme une partie d'une matrice carrée ayant certaines propriétés métriques,
— l'unicité ou la non unicité des solutions obtenues,
— les algorithmes d'approximation quand la matrice rectangulaire ne vérifie pas les conditions nécessaires et suffisantes.

Si on s'intéresse au cas des ultramétriques ou des distances quadrangulaires, les seuls travaux dans le domaine sont ceux de FURNAS (1980) qui a étudié le cas ultramétrique et abordé l'unicité des solutions dans le cadre quadrangulaire, et ceux de DESOETE, DESARBO, FURNAS et CAROLL (1984) qui donnent une généralisation de leurs algorithmes d'approximation des hiérarchies et des arbres additifs au cas rectangulaire.

Ici nous reprenons dans les § 2 et 3 les résultats de FURNAS sur les distances et les ultramétriques et nous étudierons le cas des distances à centre et des distances quadrangulaires dans les § 4 et 5.

1. NOTATIONS — DÉFINITIONS

Soient A et X deux ensembles finis, on note par a, b, c ... les éléments de A et par x, y, z, t ... ceux de X. On note par E l'ensemble $A \cup X$ et par p, q ... ses éléments.

Si p est un élément de E, on notera par E(p) l'ensemble (A ou X) auquel il appartient et par $\widetilde{E}(p)$ celui auquel il n'appartient pas.

Définition 1. — Dissimilarité rectangulaire (Sur $A \times X$).

T définie sur $A \times X$ est une dissimilarité rectangulaire si T est une application de $A \times X$ dans $\mathbf{R}^+$ qui à (a, x) associe T_{ax}. On prolonge T sur $X \times A$ par symétrie $T_{xa} = T_{ax}$.

Définition 2. — Dissimilarité carrée (sur E).

D est dissimilarité sur E, si D est une application de $E \times E$ dans $\mathbf{R}^+$ vérifiant pour tout p, $D_{pp} = 0$ pour tout p et q, $D_{pq} = D_{qp}$.

Définition 3. — Extension de T.

T étant une dissimilarité sur $A \times X$, D est une extension de T si et seulement si D est une dissimilarité sur E qui coïncide avec T sur $A \times X$: $\forall a \in A, \forall x \in X, D_{ax} = T_{ax}$. On note par $\mathscr{D}_T$ l'ensemble des extensions de T.

Définition 4. — Réduction de D.

D étant une extension de T alors T est une réduction de D à $A \times X$.

Dès lors on peut écrire les trois problèmes :

Pb 1 : Conditions nécessaires et suffisantes (C.N.S.) sur T pour qu'il existe $D \in \mathscr{D}_T$ tel que D soit une distance, une ultramétrique, une distance à centre ou une distance quadrangulaire.

Pb 2 : Unicité de l'extension.

Pb 3 : Soit T ne vérifiant pas les C.N.S. trouver T′ la plus proche de T et vérifiant les conditions.

2. DISTANCES RECTANGULAIRES

On étudie ici les conditions auxquelles doit satisfaire une dissimilarité rectangulaire pour admettre une extension qui soit une matrice de distance.

Propriété 1

Une C.N.S. pour que T, dissimilarité rectangulaire, admette une extension D qui soit une distance est que :

$$(tr)\ \forall\, a, b \in A \quad \forall x, y \in X \quad T_{ax} \leqslant T_{ay} + T_{yb} + T_{bx}$$

Si T satisfait la condition (tr), on parlera alors de distance rectangulaire.

Démonstration

— *condition nécessaire :* soit D une distance appartenant à $\mathscr{D}_T$, alors :

$$\forall a, b \in A \quad \text{et} \quad \forall x \in X \quad D_{ax} \leqslant D_{ab} + D_{bx}$$

$$\forall a, b \in A \quad \text{et} \quad \forall y \in X \quad D_{ab} \leqslant D_{ay} + D_{yb}$$

donc : $\forall a, b \in A \qquad \forall x, y \in X \quad D_{ax} \leqslant D_{ay} + D_{yb} + D_{bx}$

D coïncidant avec T sur $(A \times X)$, T vérifie (tr).

— *condition suffisante :* soit T vérifiant (tr) et soit D^* appartenant à $\mathscr{D}_T$ et définie par :

$$\forall p \in E(p) \quad \forall q \in \tilde{E}(p) \quad D^*_{pq} = T_{pq} \quad D^*_{pp} = 0$$

$$\forall p \in E(p) \quad \forall q \in E(p) \quad D^*_{pq} = \max_{r \in \tilde{E}(p)} | T_{rp} - T_{rq} |$$

alors on vérifie que D^* est une distance.

Pour T fixé, vérifiant (tr) il n'y a pas unicité des extensions vérifiant l'inégalité triangulaire. On peut cependant les caractériser de la façon suivante :

Propriété 2

Soit T vérifiant (tr) et soit $D \in \mathscr{D}_T$, une condition nécessaire pour que D soit une distance est que :

$$\forall p, q \in E(p) \max_{r \in \tilde{E}(p)} | T_{rp} - T_{rq} | \leqslant D_{pq} \leqslant \min_{r \in \tilde{E}(p)} (T_{rp} + T_{rq})$$

Démonstration

1. Supposons qu'il existe p et q dans E(p) tels que :

$$D_{pq} > \min_{r \in \tilde{E}(p)} (T_{rp} + T_{rq})$$

donc il existe ℓ dans $\tilde{E}(p)$ tel que : $D_{pq} > T_{\ell p} + T_{\ell p}$. D coïncidant avec T sur $A \times X$, D n'est pas une distance.

2. Supposons qu'il existe p et q dans E(p) tel que :

$$D_{pq} \leqslant \max_{r \in \tilde{E}(p)} | T_{rp} - T_{rq} |$$

alors il existe ℓ dans $\tilde{E}(p)$ t.q. $D_{pq} \leqslant T_{\ell p} - T_{\ell q}$ (au besoin on intervertit p et q) et donc D n'est pas une distance.

Cependant cette condition n'est pas suffisante pour assurer que D, extension de T satisfaisant (tr), soit une distance. En effet, nous sommes ainsi assurés que l'inégalité triangulaire est satisfaite pour tout triplet dont deux éléments sont dans un ensemble et le troisième dans l'autre, mais nous ne savons rien sur les triplets dont les trois éléments sont dans le même ensemble.

En conclusion une C.N.S. pour que, T vérifiant (tr), D soit une extension de T qui soit une distance est que :

a) D vérifie la condition de la proposition 2.
b) D vérifie l'inégalité triangulaire sur $A \times A$ et $X \times X$.

On vérifie aisément que les deux extensions définies par les limites de la propriété 2 sont des distances.

D_1 défini par $D_{pq} = \max_{r \in \tilde{E}(p)} | T_{rp} - T_{rq} | \quad \forall p, q \in E(p)$

D_2 défini par $D_{pq} = \min_{r \in \tilde{E}(p)} (T_{rp} + T_{rq} \quad \forall p, q \in E(p)$

Nous allons maintenant nous intéresser aux extensions de T qui vérifient l'inégalité ultramétrique.

3. ULTRAMÉTRIQUES RECTANGULAIRES

Définition 5. — Condition (ur) (ultramétrique rectangulaire).

T dissimilarité rectangulaire sur $(A \times X)$ vérifie la condition (ur) si et seulement si :

$$\forall a, b \in A \quad \forall x, y \in X \quad T_{ax} \leqslant \max (T_{ay}, T_{yb}, T_{bx})$$

On parle alors d'une ultramétrique rectangulaire.

Propriété 3

Une condition nécessaire et suffisante pour qu'une dissimilarité rectangulaire T admette une extension qui vérifie l'inégalité ultramétrique est que T vérifie la condition (ur).

Démonstration

La condition nécessaire est évidente. D étant une ultramétrique sur E, alors sa restriction T sur $A \times X$ vérifie (ur). En effet :

$$\forall a, b \in A$$

et

$$\forall x, y \in X$$

on a : $$D_{ax} \leqslant \max (D_{ab}, D_{bx})$$

et $D_{ab} \leqslant \max (D_{ay}, D_{by})$

d'où $D_{ax} \leqslant \max (D_{ay}, D_{by}, D_{bx})$.

comme D coïncide avec T sur $(A \times X)$, T vérifie (ur).

— Condition suffisante : Soit T une ultramétrique rectangulaire et soit D définie par :

$$\forall p, q \in E(p) \quad D_{pq} = \min_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq})), \; D_{pp} = 0$$

$$\forall p \in E(p) \quad \forall q \in \tilde{E}(p) \quad D_{pq} = T_{pq}$$

alors D est une ultramétrique :

1. si $p, q \in E(p)$ et $\ell \in \tilde{E}(p)$

   alors $\min_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq})) \leqslant \max (T_{\ell p}, T_{\ell q})$

   d'où $D_{pq} \leqslant \max (D_{\ell p}, D_{\ell q})$

2. si $p, \ell \in E(p)$ et $q \in \tilde{E}(p)$ on veut $D_{pq} \leqslant \max (D_{\ell p}, (D_{\ell q})$

   or $D_{p\ell} = \min_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{r\ell})) = \max (T_{sp}, T_{s\ell})$

   et comme $T_{pq} \leqslant \max (T_{sp}, T_{s\ell}, T_{\ell q})$

   d'après la condition (ur) on a le résultat voulu.

3. si $p, q, \ell \in E(p)$ il faut $D_{pq} \leqslant \max (D_{p\ell}, (D_{q\ell})$

   on a $D_{pq} = \min_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq})) \leqslant \max (T_{sp}, T_{sq}) \qquad \forall s \in \tilde{E}(p)$

   comme $T_{sq} \leqslant \max (T_{s\ell}, T_{t\ell}, T_{tq}) \qquad \forall t \in \tilde{E}(p)$

   on a $D_{pq} \leqslant \max (T_{sp}, T_{s\ell}), \max (T_{t\ell}, T_{tq})) \qquad \forall s, t \in \tilde{E}(p)$

   soit s tel que : $\max (T_{ps}, T_{\ell s}) = \min_{r \in \tilde{E}(p)} (\max (T_{pr}, T_{\ell r})) = D_{p\ell}$

   et t tel que : $\max (T_{qt}, T_{\ell t}) = \min_{r \in \tilde{E}(p)} (\max (T_{qt}, T_{\ell t})) = D_{q\ell}$

   alors $D_{pq} \leqslant \max (D_{p\ell}, (D_{q\ell})$

On obtient donc, si T satisfait la condition (ur), une extension explicite de T qui est ultramétrique. Cette extension est « presque unique » au sens où s'il n'existe pas 2 lignes identiques ou 2 colonnes identiques dans T, la solution est unique.

Propriété 4

Soit une ultramétrique rectangulaire, si :

$$\forall p \in E, \quad \forall q \in E(p), \quad \exists r \in \tilde{E}(p) \text{ tel que } T_{pr} \neq T_{qr}$$

alors D définie par:

$$\forall p \in E, \quad \forall q \in \tilde{E}(p) \quad D_{pq} = T_{pq}$$

$$\forall p \in E, \quad \forall q \in E(p) \quad D_{pq} = \min_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq})), \quad D_{pp} = 0$$

est la seule extension ultramétrique de T.

Démonstration

Soit, $D' \in \mathscr{D}_T$, D' ultramétrique et $D' \neq D$.

Donc il existe au moins un couple (p, q) dans E(p) tel que l'on ait :

soit (1) $\quad D'_{pq} > D_{pq} = \min\limits_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq}))$

soit (2) $\quad D'_{pq} < D_{pq} = \min\limits_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq}))$

Si on a (1), alors il existe s dans $\tilde{E}(p)$ tel que :

$$D'_{pq} > \max (T_{sp}, T_{sq}) = \max (D'_{sp}, D'_{sq})$$

et D' n'est pas ultramétrique.

Si on a (2), alors pour tous s dans $\tilde{E}(p)$ on a :

$$D'_{pq} < \max (T_{sp}, T_{sq})$$

l'hypothèse nous permet de choisir s tel que $T_{sp} \neq T_{sq}$. Supposons que l'on ait $T_{sp} > T_{sq}$, donc :

$$D'_{pq} < T_{sp} = D'_{sp}$$

Comme D' est ultramétrique on doit avoir simultanément :

$$D'_{sp} \leqslant \max (D'_{sq}, D'_{pq}) \text{ ce qui est impossible.}$$

Dans le cas où la condition n'est pas satisfaite, on a donc un ensemble d'éléments, p, q, r, ... de E(p) qui forment une classe de D. Les distances à l'intérieur de la classe sont alors inférieures à :

$$\min_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq}))$$

pour tout couple p et q, et doivent bien sûr vérifier l'inégalité ultramétrique.

4. DISTANCES À CENTRE RECTANGULAIRES

Rappelons qu'une distance à centre est une distance engendrée par un vecteur X de $\mathbf{R}^n$ à termes positifs, de la façon suivante :

$$\begin{array}{ll} \text{pour tout} \quad i \neq j & D_{ij} = X_i + X_j \\ \text{pour tout} \quad i & D_{ii} = 0 \end{array}$$

Définition 6 : Condition (cr)

Une dissimilarité T sur $(A \times X)$ vérifie la condition (cr) si et seulement si $\forall a \in A$, $\forall x \in X$ il existe un vecteur U de $\mathbf{R}^{n+}$ et un vecteur V de $\mathbf{R}^{p+}$ tel que :

$$T_{ax} = U_a + V_x$$

(n = card A et p = card X).

Propriété 5

Une condition nécessaire et suffisante pour qu'une matrice de dissimilarité admette une extension D qui soit une distance à centre est que T vérifie la condition (cr).

Démonstration

La condition nécessaire est évidente : si D est une distance à centre alors :

$$\forall a \in A,\ \forall x \in X,\ \exists O \in \mathbf{R}^{n+p} \text{ t.q. } D_{ax} = O_a + O_x$$

en appelant U le vecteur formé des n premières composantes et V celui formé des p dernières, on obtient la condition (cr).

Condition suffisante : soit D défini par :

$$\forall a \in A,\ \forall x \in X,\ D_{ax} = T_{ax} = U_a + V_x$$

$$\forall a \in A,\ \forall b \in A,\ D_{ab} = U_a + U_b$$

$$\forall x \in X,\ \forall y \in X,\ D_{xy} = V_x + V_y$$

Alors en notant O le vecteur de $\mathbf{R}^{n+p}$ obtenu en juxtaposant les p composantes de V aux n de U, $O = (V \mid U)$, on obtient une distance à centre engendrée par le vecteur O.

On n'a pas l'unicité de l'extension dans le cas des distances à centre. En effet on a la propriété suivante.

Propriété 6

Soit T une dissimilarité rectangulaire sur $(A \times X)$ vérifiant la condition (cr) et D une extension qui soit une distance à centre engendrée par le vecteur O. Alors la distance D′ définie par :

$$D'_{pq} = O'_p + O'_q \text{ avec } \begin{cases} \text{si } p \in A : O'_p = O_p + K,\ K \in \mathbf{R} \\ \text{si } p \in X : O'_p = O_p - K,\ K \leqslant \min\limits_p (O_p) \end{cases}$$

est une distance à centre extension de T.

La démonstration est évidente.

Pour une distance à centre rectangulaire donnée on engendre ainsi la totalité de ses extensions qui soient des distances à centre comme le montre la propriété 7.

Propriété 7

Soit T une distance à centre rectangulaire, soient D et D′ deux extensions de T qui soient des distances à centre, alors elles sont égales à une constante près.

Démonstration

Soit $D_{ax} = O_a + O_x$

et $$D'_{ax} = O'_a + O'_x$$

comme $$D_{ax} = D'_{ax} = T_{ax} \quad \forall a \in A \quad \forall x \in X$$

on a $$O_a - O'_q = O'_x - O'_x \quad \forall a, \forall x$$

Les distances à centre rectangulaires ne sont pas en soi très intéressantes mais vont nous permettre d'aborder le cas des arbres additifs avec les distances quadrangulaires. En effet l'étude des arbres additifs rectangulaires qui est beaucoup plus compliquée que celle des ultramétriques rectangulaires se fait par le biais de la décomposition de toute distance quadrangulaire en la somme d'une ultramétrique et d'une distance à centre.

La différence vient du fait que l'on a maintenant deux centres, un pour chaque ensemble et donc que la décomposition sera double.

5. ARBRES ADDITIFS RECTANGULAIRES

Rappelons qu'une distance est quadrangulaire si pour tout quadruplet i, j, k, l l'inégalité du même nom est vérifiée :

$$\forall i, j, k, l \quad D_{ij} + D_{kl} \leqslant \max (D_{ik} + D_{jl}, D_{il} + D_{jk})$$

L'intérêt des distances quadrangulaires est quelles sont représentables exactement par un arbre additif. On appelle arbre additif un arbre valué dont les éléments terminaux sont des éléments de E et tels que la distance entre deux éléments soit la longueur de l'unique chemin les reliant : la somme des valeurs des arêtes composant le chemin. On peut voir par exemple [1] et [6] pour une étude sur les propriétés des distances quadrangulaires et des arbres additifs.

Définition 7 : Condition (ar)

Une dissimilarité rectangulaire T définie sur $(A \times X)$ satisfait la condition (ar) si et seulement si :

$$\forall a, b, c \in A, \quad \forall x, y, z \in X$$

on a :

$$T_{ax} + T_{bx} + T_{cz} \leqslant \max (T_{ax} + T_{bz} + T_{cy}, T_{ay} + T_{bx} + T_{cz}, T_{ay} + T_{bz} + T_{cx}, T_{az} + T_{bx} + T_{cy}, T_{az} + T_{by} + T_{cx})$$

Il est bien clair que comme dans le cas classique cette condition est équivalente à la suivante.

Pour tout sextuplet $(a, b, c \in A, x, y, z \in X)$ parmi les six sommes de distances que l'on peut former, les deux plus grandes sont égales.

Afin de montrer que la condition (ar) est une condition nécessaire et suffisante pour qu'une dissimilarité rectangulaire admette une extension qui soit une distance quadrangulaire, nous allons montrer 3 lemmes.

Auparavant nous allons montrer que les ultramétriques rectangulaires vérifient la condition (ar).

Propriété 8

Si T vérifie (ur), T vérifie (ar).

Démonstration

Puisque T vérifie (ur), T admet une extension U qui comme toute ultramétrique vérifie l'inégalité quadrangulaire. Donc :

$$\forall a, b \in A, \quad \forall x, y \in X : U_{ax} + U_{by} \leqslant \max (U_{ab} + U_{xy}, U_{ay} + U_{bx})$$

et donc : $\forall a, b, c \in A, \quad \forall x, y, z \in X$

$$U_{ax} + U_{by} + U_{cz} \leqslant \max (U_{ab} + U_{xy} + U_{cz}, U_{ay} + U_{bx} + U_{cz})$$

Considérons la somme :

$$U_{ab} + U_{xy} + U_{cz}$$

comme : $U_{xy} + U_{cz} \leqslant \max (U_{cx} + U_{yz}, U_{cy} + U_{xz})$

on a :

$$U_{ax} + U_{by} + U_{cz} \leqslant \max (U_{ab} + U_{cx} + U_{yz}, U_{ab} + U_{cy} + U_{xz}, U_{ay} + U_{bx} + U_{cz})$$

D'autre part :

$$U_{ab} + U_{yz} \leqslant \max (U_{ay} + U_{bz}, U_{az} + U_{by})$$

et

$$U_{ab} + U_{xz} \leqslant \max (U_{ax} + U_{bz}, U_{az} + U_{bx})$$

on a :

$$U_{ax} + U_{by} + U_{cz} \leqslant \max (U_{ay} + U_{bz} + U_{cx}, U_{az} + U_{by} + U_{cx}, U_{ax} + U_{bz} + U_{cy}, U_{az} + U_{bx} + U_{cy}, U_{ay} + U_{bx} + U_{cz})$$

pour tout $a, b, c \in A$ et $x, y, z \in X$

comme U coïncide avec T sur $A \times X$, T vérifie (cr).

Le lemme 1 montre que comme dans le cas classique les quadrangulaires ayant des propriétés de sphéricité sont des ultramétriques : les éléments de l'ensemble A sont sur une sphère de centre ω (appartenant à X), de rayon k. De même les éléments de X sont sur une sphère de centre α (appartenant à A) et de rayon k. De plus toutes les distances inter-ensembles sont inférieures au rayon de la sphère.

Lemme 1

Soit T une dissimilarité sur $(A \times X)$ et vérifiant la condition (ar). Si :

$$\exists \alpha \in A \text{ et } \omega \in X$$

tels que :

$$T_{\alpha x} = K \text{ pour tout } x \in X$$

$$T_{a\omega} = K \text{ pour tout } a \in A$$

et $T_{ax} \leqslant K$ pour tout $a \in A$ et $x \in X$

alors T vérifie la condition (ur).

Démonstration

T véfifiant (ar), elle la vérifie en particulier en α et ω. Soit :

$$\forall a, b \in A \quad , \quad \forall x, y \in X$$

$$T_{ax} + T_{\alpha y} + T_{c\omega} \leqslant \max (T_{ax} + T_{\alpha\omega} + T_{cy}, T_{ay} + T_{\alpha x} + T_{c\omega},$$
$$T_{ay} + T_{\alpha\omega} + T_{cx}, T_{a\omega} + T_{\alpha x} + T_{cy}, T_{a\omega} + T_{\alpha y} + T_{cx}$$

ou par hypothèse :

$$T_{ax} + 2k \leqslant \max$$
$$(T_{ax} + T_{cy} + k, T_{ay} + 2k, T_{ay} + T_{cx} + k, T_{cy} + 2k, T_{cx} + 2k)$$
$$T_{ax} \leqslant \max (T_{ay}, T_{cy}, T_{cx}, T_{ax} + T_{cy} - k, T_{ay} + T_{cx} - k)$$

comme on a :

$$T_{cy} \leqslant k \Rightarrow T_{ax} + T_{cy} - k \leqslant T_{ax}$$

et

$$T_{cx} \leqslant k \Rightarrow T_{ay} + T_{cx} - k \leqslant T_{ay}$$

d'où :

$$\forall a, b \in A \quad \forall x, y \in X \quad T_{ax} \leqslant \max (T_{ay}, T_{cy}, T_{cx})$$

Le lemme 2 nous dit que comme dans le cas classique les arbres additifs rectangulaires sont stables par ajout d'une distance à centre rectangulaire.

Lemme 2

Soient T et C deux dissimilarités définies sur $(A \times X)$, T vérifiant la condition (ar) et C la condition (cr), alors $T + \lambda C$ vérifie la condition (ar) pour tout λ réel.

Démonstration

La condition (ar) fait intervenir 6 sommes définies par 6 éléments distincts : a, b, c, x, y, z. Donc rajouter la matrice λC à la matrice T revient à rajouter le terme $\lambda (U_a + U_b + U_c + V_x + V_y + V_z)$ à chacune des 6 sommes, laissant inchangée l'inégalité.

Le lemme 3 nous assure de l'existence des centres α et ω, postulée au niveau du lemme 2 et les caractérisent.

Lemme 3

Soit T une dissimilarité rectangulaire sur $(A \times X)$ vérifiant (ar) alors il existe au moins un couple (α, ω) de $(A \times X)$ tel que :

$$\forall a \in A \quad , \quad \forall x \in X \quad T_{\alpha\omega} \leqslant T_{\alpha x} + T_{a\omega} - T_{ax}$$

Démonstration

Il revient au même de montrer que :

$$T_{\alpha\omega} \leqslant \min_{a,x} (T_{\alpha x} + T_{a\omega} - T_{ax})$$

ou qu'il existe α et ω tels que :

$$F(\alpha, \omega) = \min_{a,x} (T_{\alpha\omega} + T_{ax} - T_{\alpha x} - T_{a\omega}) \leqslant 0$$

posons α et ω des éléments de (A, X) réalisant :

$$\min_{\substack{b \in A \\ y \in X}} F(b, y)$$

et montrons que ce couple satisfait la condition.

On a donc :

$$F(\alpha, \omega) = \min_{b,y} (\min_{a,x} (T_{ax} + T_{by} - (T_{ay} + T_{bx})))$$

soit :

$$\forall a, b \in A \quad \forall x, y \in X \quad F(\alpha, \omega) \leqslant (T_{ax} + T_{by} - (T_{ay} + T_{bx})) \qquad (1)$$

que l'on peut écrire :

$$\forall a, b, c \in A \quad \forall x, y, z \in X$$
$$F(\alpha, \omega) \leqslant (T_{ax} + T_{by} + T_{cz}) - (T_{ay} + T_{bx} + T_{cz}) \quad (2)$$

D'autre part en écrivant l'inégalité (1) pour le quadruplet (c, a, x, z) et en sommant les deux inégalités on obtient :

$$\forall a, b, c \in A \quad \forall x, y, z \in X$$
$$2F(\alpha, \omega) \leqslant (T_{az} + T_{by} + T_{cx}) - (T_{ay} + T_{bx} + T_{cz}) \quad (3)$$

T satisfaisant la condition (ar), les deux plus grandes sommes de 3 distances entre 6 éléments sont égales. Ces deux sommes sont soit du type (2), c'est-à-dire avec une distance commune (ici T_{cz}), soit sous la forme (3), c'est-à-dire qu'aucune distance n'est commune. Comme les inégalités (2) et (3) sont satisfaites pour tout a, b, c et pour tout x, y, z on a toujours, au besoin en renommant les éléments, un des termes à droite de l'inégalité qui est nul. Donc :

$$F(\alpha, \omega) \leqslant 0$$

Nous allons maintenant pouvoir montrer la condition nécessaire et suffisante.

Propriété 9

Soit T une dissimilarité définie sur (A x X), une condition nécessaire et suffisante pour que T admette une extension vérifiant l'inégalité quadrangulaire est que T vérifie la condition (ar).

Démonstration :

— *Condition nécessaire :* soit Q une dissimilarité définie sur E (avec : E = A ∪ X) et vérifiant l'inégalité quadrangulaire, alors Q_R restriction de Q à (A × X) satisfait (ar).

En effet : soit U + C une décomposition de Q en une ultramétrique et une distance à centre. D'après la propriété 3 la restriction U_R de U satisfait (ur) et donc (ar) d'après la propriété 8.

C_R restriction de C à (A × X) satisfait la condition (cr) et donc $U_R + C_R$ satisfait (ar) d'après le lemme 2.

— *Condition suffisante :* soit T une dissimilarité définie sur (A × X) et satisfaisant (ar), montrons qu'elle admet une extension Q définie sur E et vérifiant l'inégalité quadrangulaire.

Soit : $\alpha \in A$ et $\omega \in X$

tels que : $T_{\alpha\omega} \leqslant T_{\alpha x} + T_{a\omega} - T_{ax}$ pour tout a et x (lemme 3)

On pose : $k = T_{\alpha\omega}$ et

$$\forall x \in X \quad O_x = T_{\alpha x} - 1/2\, k$$

$$\forall a \in A \quad O_a = T_{a\omega} - 1/2\, k$$

Posons pour tout $a \in A$ et pour tout $x \in X$

$$D_{ax} = T_{ax} - O_a - O_x + K_0$$

K_0 étant une constante arbitraire choisie telle que D_{ax}, soit partout positif.

D'après le lemme 2, D vérifie la condition (ar) et d'autre part on a :

$$\forall x \in X \quad D_{\alpha x} = K_0$$

$$\forall a \in A \quad D_{a\omega} = K_0$$

et comme : $D_{ax} = T_{ax} - T_{\alpha x} - T_{a\omega} + T_{\alpha\omega} + K_0 \leqslant K_0$ par définition de α et ω, D vérifie (ur) d'après le lemme 1.

Donc T s'écrit comme la somme de D et de O (moins la constante K_0), où D est une ultramétrique rectangulaire et O une distance à centre rectangulaire. Soit Q la somme des extensions de D et de O, Q est une extension de T et est une distance quadrangulaire comme somme d'une ultramétrique et d'une distance à centre.

On en déduit la propriété suivante.

Propriété 10

Toute dissimilarité rectangulaire T vérifiant la condition (ar) peut se décomposer en la somme d'une ultramétrique rectangulaire et d'une distance à centre rectangulaire.

(La démonstration est incluse dans celle de la condition suffisante de la propriété précédente).

Les conditions d'unicité de l'extension sont donc liées à celles des ultramétriques rectangulaires et celles des distances à centre. On peut toujours rallonger d'une longueur donnée les arêtes de l'arbre adjacentes aux éléments d'un ensemble à condition de raccourcir celles adjacentes aux éléments de l'autre ensemble. Cette propriété est liée à l'extension des distances à centre. Pour ce qui est de la partie ultramétrique, elle est unique s'il n'existe pas dans T deux lignes ou deux colonnes identiques à une constante près.

Propriété 11

Soit T une distance sur $(A \times X)$ vérifiant la condition (ar) et D une extension quadrangulaire de T.

Si $\forall p \in E$, $\forall q \in E(p)$ il n'existe pas de constante C telle que

$$T_{pr} = T_{qr} + C \text{ pour tout } r \in \tilde{E}(p)$$

alors toutes les extensions quadrangulaires de T sont de la forme :

$$D'_{pq} = D_{pq} + K_p + K_q \quad \text{avec} \quad K_r = k \quad k \text{ si } r \in A \quad k \in \mathbf{R}$$
$$K_r = -k \text{ si } r \in X$$

Démonstration

D′ est clairement une extension quadrangulaire de T, car D′ coïncide avec T sur (A × X) et est quadrangulaire comme somme d'une quadrangulaire et d'une distance à centre. Soient U_R et C_R une décomposition de T; par hypothèse U_R admet une extension unique et donc D′ diffère de D par ajout ou retrait d'une constante sur les ensembles A et X.

Afin d'illustrer les propriétés précédentes, considérons les exemples suivants :

Soit A = {a, b, c, d} et X = {x, y, z}

et T la matrice

T =

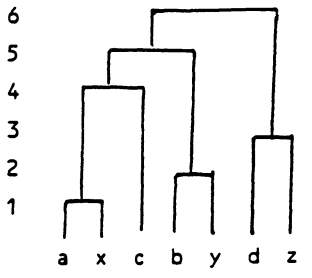
	x	y	z
a	1	5	6
b	5	2	6
c	4	5	6
d	6	6	3

T vérifiant la condition (ur) et la condition de la propriété. 4. admet une extension qui est unique, la matrice D.

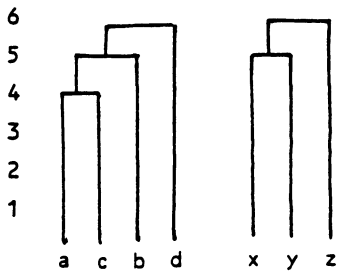
D =

	a	b	c	d	x	y	z
a	0	5	4	6	1	5	6
b		0	5	6	5	2	6
c			0	6	4	5	6
d				0	6	6	3
x					0	5	6
y						0	6
z							0

La représentation hiérarchique de D est l'arbre suivant :



L'extension de A × X définit deux ultramétriques (A × A) et (X × X) qui sont les sous arbres de l'arbre précédent restreint à (A × A) et à (X × X).



On obtient bien ainsi simultanément la classification sur les éléments de A, sur les éléments de X et sur (A × X).

— *cas où l'extension n'est pas unique*

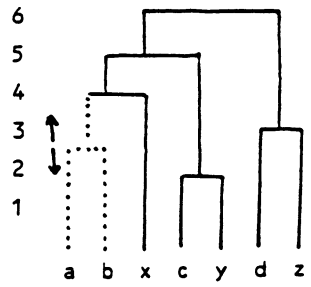
$$A = \{a, b, c, d\} \text{ et } X = \{x, y, z\}$$

et T_2 la matrice

$T_2 =$

	x	y	z
a	4	5	6
b	4	5	6
c	5	2	6
d	6	6	3

Les lignes a et b étant identiques, l'extension ne sera pas unique (propriété 4). Il est en effet impossible de calculer la distance (a b) qui peut être comprise entre 0 et 4 (car ax = bx = 4)



a et b s'agrègent à un niveau indéterminé compris entre 0 et 4.

— *cas d'une distance quadrangulaire*

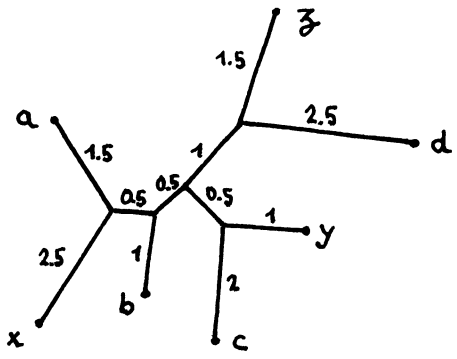
T_3 =

	x	y	z
a	4	4	5
b	4	3	4
c	6	3	5
d	7	5	4

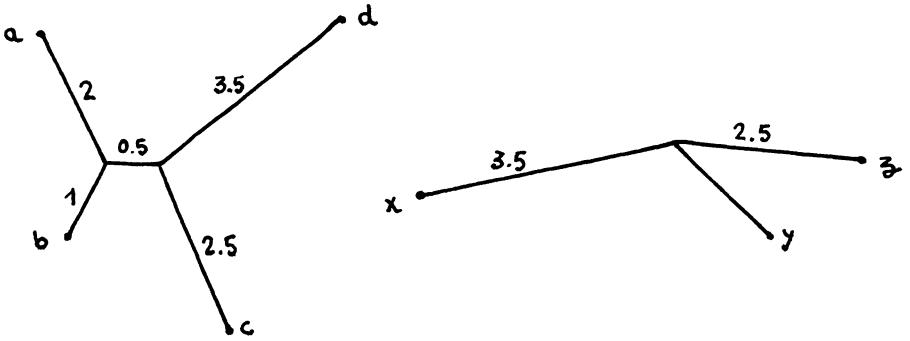
elle admet comme extension Q

	a	b	c	d	x	y	z
a	0	3	5	6	4	4	5
b		0	4	5	4	3	4
c			0	6	6	3	5
d				0	7	5	4
x					0	5	6
y						0	4
z							0

qui se représente sous la forme de l'arbre additif suivant :



d'où l'on peut extraire les sous arbres sur A et sur X.



la décomposition de T est donnée par :

$U_R =$

	x	y	z
a	1	3	3
b	2	3	3
c	3	2	3
d	3	3	1

$C_R =$

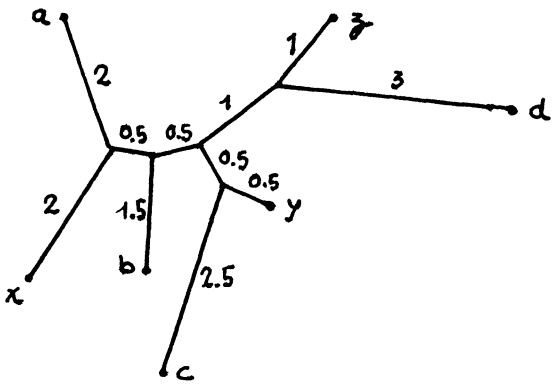
	x	y	z
a	3	1	2
b	2	0	1
c	3	1	2
d	4	2	3

L'extension calculée n'est pas unique, on peut ajouter à toutes les arêtes incidentes aux éléments de A une constante K et la retrancher à ceux de X. Choisissons K = 0.5. On obtient :

$D' =$

	a	b	c	d	x	y	z
a	0	4	6	7	4	4	5
b		0	5	6	4	3	4
c			0	7	6	3	5
d				0	7	5	6
x					0	4	5
y						0	3
z							0

et l'arbre additif correspondant :



6. CALCUL DES EXTENSIONS

Etant donné une matrice T vérifiant une des conditions précédentes ((tr), (ur), (cr) ou (ar)), comment calculer explicitement une extension satisfaisant à la propriété métrique voulue ?

1. Distance rectangulaire

Si T satisfait la condition (tr) les deux extensions, minimales et maximales, sont données par D_1 et D_2.

$$\forall p \in E(p) \quad \forall q \in \tilde{E}(p) \qquad D_{1pq} = D_{2pq} = T_{pq}$$

$$\forall p, q \in E(p) \qquad D_{1pq} = \max_{r \in \tilde{E}(p)} |T_{rp} - T_{rq}|, \; D_{1pp} = 0$$

$$D_{2pq} = \min_{r \in \tilde{E}(p)} (T_{rp} + T_{rq}), \; D_{2pp} = 0$$

2. Distance ultramétrique

Si T satisfait la condition (ur) et la condition de la propriété 4. L'unique extension est donnée par :

$$\forall p \in E(p) \quad \forall q \in \tilde{E}(p) \qquad D_{pq} = T_{pq}$$

$$\forall p, q \in E(p) \qquad D_{pq} = \min_{r \in \tilde{E}(p)} (\max (T_{rp}, T_{rq})), \; D_{pp} = 0$$

Si T ne satisfait pas la condition de la propriété 4, D est l'extension maximale.

3. Distance à centre

On choisit de façon arbitraire a, x, U_x, et V_a de façon à avoir $U_x + V_a = T_{ax}$, puis on calcule :

pour $\quad i \in A \quad V_i = T_{xi} - U_x$

pour $\quad j \in X \quad U_j = T_{ja} - V_a$

en posant $\quad O_i = V_i \quad \text{si} \quad 1 \leqslant i \leqslant n$

et $\quad O_i = U_{i-n} \quad \text{si} \quad n+1 \leqslant i \leqslant n+p$

on calcule $\quad D_{ij} = O_i + O_j$

4. Distances quadrangulaires

1) On cherche un couple (α, ω) de $(A \times X)$ vérifiant :

$$\forall a \in A \quad \forall x \in X \quad T_{\alpha\omega} \leqslant T_{\alpha x} + T_{a\omega} - T_{ax}$$

En général le couple (α, ω) vérifiant $T_{\alpha\omega} = \min_{a,x} T_{ax}$ satisfait la condition.

2) On calcule le vecteur O par :

$$\forall a \in A \quad O_a = T_{a\omega} - 1/2\, T_{\alpha\omega}$$

$$\forall x \in X \quad O_x = T_{\alpha x} - 1/2\, T_{\alpha\omega}$$

3) On en déduit les matrices C_R sur $(A \times X)$ et C sur $(E \times E)$ par :

$\forall i, j \in E \quad C_{ij} = O_i + O_j$ et C_R la restriction de C.

4) On calcule $U_R = T - C_R$. Comme U_R a des termes négatifs, on peut lui ajouter une constante arbitraire K_0.

5) On calcule U l'extension de U_R.

6) On calcule $Q = U + C$. Si on a ajouté une constante K_0 à U_R, il faut la retrancher à Q.

7. ASPECTS ALGORITHMIQUES

Quand la matrice T ne vérifie ni la condition (ur), ni la condition (ar), ce qui est bien sur le cas général, on est amené à rechercher une matrice U_R ou une matrice Q_R qui approche au mieux la matrice T.

Pour ce faire plusieurs approches sont possibles :

— recherche de U_R puis de Q_R par des méthodes numériques de pénalisation cherchant à minimiser :

$$\sum_a \sum_x (T_{ax} - U_{ax})^2 \qquad \text{ou} \qquad \sum_a \sum_x (T_{ax} - Q_{ax})^2$$

voir les travaux de DE SOETE, DE SARBO, FURNAS et CAROLL. Ces méthodes exigent énormément de calculs.

— La deuxième solution consiste à calculer une extension de la matrice T comme si elle vérifiait la condition (ur) puis à appliquer un des algorithmes classiques de construction hiérarchique ou d'arbre additif.

Cette solution, peu élégante, n'est pas très satisfaisante car on a du mal à comprendre le sens d'une telle extension et la signification finale de la figure construite. En outre elle oblige à travailler sur une matrice $(n + p) \times (n + p)$ au lieu de la matrice $(n \times p)$ d'origine.

— La troisième voie consiste à adapter les algorithmes usuels de recherche d'ultramétrique et d'arbres additifs au cas rectangulaire. Pour les algorithmes hiérarchiques, pas de difficulté essentielle. Nous présentons ci-dessous l'algorithme d'agrégation proposé par FURNAS, généralisé à un fonction d'agrégation quelconque.

Pour les arbres additifs, il nous faut éliminer tous les algorithmes qui travaillaient à partir de l'inégalité quadrangulaire. En effet maintenant l'inégalité caractéristique s'écrit à partir de 6 éléments, et les algorithmes d'ajustement seraient donc au minimum en $O(n^6)$ ce qui est sans intérêt. Nous présentons donc ici un algorithme basé sur la décomposition en ultramétrique et distance à centre.

a. Algorithme d'agrégation hiérarchique

1) Recherche (a, x) tel que : $T_{ax} = \min_{b,y} T_{by}$

2) Agréger le couple (a, x) pour ne former qu'un seul élément e. L'élément e n'appartient donc ni à A ni à X, mais est une classe de E. $(E = A \cup X)$.

3) Recalcul des distances $T_{e,p}$

si $p \in A$ $\quad T_{e,p} = T_{x,p}$

si $p \in X$ $\quad T_{e,p} = T_{a,p}$

mais en général e et p sont des classes de E.

Appelons e_A et e_X les éléments de e qui appartiennent à A et à X et de même p_A et p_X les éléments de p appartenant à A et X. Alors $T_{e,p}$ est une fonction de T_{ij} pour $i \in e_A$ et $j \in p_X$ et de $T_{\ell,k}$ pour $\ell \in e_X$ et $k \in p_A$. On peut choisir la fonction min, max, moyenne ou tout autre fonction d'agrégation comme dans le cas classique.

4) On itère les étapes 2 et 3 jusqu'à n'avoir qu'une seule classe.

b. Algorithme de recherche d'abre additif

L'idée est d'enlever une distance à centre rectangulaire à la matrice T pour se ramener à une distance proche d'une ultramétrique rectangulaire, c'est-à-dire sphérique (voir lemme 1). Ensuite d'appliquer l'algorithme précédent pour trouver l'ultramétrique. En sommant la distance à centre et l'ultramétrique, on trouve ainsi l'arbre additif recherché. Cet algorithme est la transcription au cas rectangulaire de l'algorithme présenté dans [1].

1) On choisit 2 centres α et ω par $T_{\alpha\omega} = \min_{a,x} T_{ax}$

2) On calcule le vecteur O par :

$$\forall a \in A \qquad O_a = T_{a\omega} - 1/2\, T_{\alpha\omega}$$

$$\forall x \in X \qquad O_x = T_{\alpha x} - 1/2\, T_{\alpha\omega}$$

3) On en déduit la matrice : $C_R(a, x) = O_a + O_x$.

4) On calcule : $D_{ax} = T_{ax} - C_R(a, x) + K_o$.

K_o est une constante arbitraire pour que D ait tous ses termes positifs ou nuls.

5) On approxime D par U_R, ultramétrique rectangulaire, par l'algorithme du paragraphe précédent.

6) On calcule son extension U.

7) On compose : $Q = U + C - K_o$ pour obtenir la distance quadrangulaire.

On vérifie que la complexité des deux algorithmes est en O (np). D'autre part on montre les propriétés suivantes.

Propriété 12

Si T vérifie la condition (ur) et si on choisit pour fonction d'agrégation les fonctions min, max ou moyenne alors l'algorithme a- reconstruit la matrice T.

Démonstration

Si T vérifie la condition (ur) T admet une extension ultramétrique U. Alors il est clair que l'algorithme « a » coïncide dans ce cas avec l'algorithme de classification ascendante hiérarchique qui laisse invariant la matrice U pour les fonctions d'agrégations min, max et moyenne. ■

Propriété 13

Si T vérifie la condition (ar) et si le couple (α, ω) vérifie la condition du lemme 3, alors l'algorithme b- reconstruit la matrice T.

Démonstration

Si T vérifie (ar) alors l'algorithme b- effectue la décomposition de T en la somme d'une ultramétrique et d'une distance à centre rectangulaire (voir démonstration de la propriété 9). ■

Comme dans le cas classique, on peut à partir de ces algorithmes de bases construire un certain nombre de variantes :

— pour approximer un critère des moindres carrés on peut conserver l'arbre obtenu par l'algorithme b- et réévaluer la longueur des arêtes,

— on peut également chercher la distance à centre rectangulaire la plus proche de la matrice T donnée, puis appliquer l'algorithme a- à la matrice (T − C). On obtient ainsi une ultramétrique rectangulaire qu'il suffit d'ajouter à C pour obtenir la distance Q.

8. UN EXEMPLE

Afin d'illustrer le fonctionnement des algorithmes précédent considérons l'exemple fictif suivant : soit l'ensemble X des 4 disciplines Mathématiques, Français, Histoire et Education Physique, et l'ensemble A de 6 élèves {a, b, c, d, e, f}. Le tableau des notes est le suivant :

	Math.	Fr.	Hist.	Ed. Ph.
a	16	14	12	12
b	15	10	8	5
c	11	12	12	13
d	10	12	10	12
e	8	8	10	13
f	8	13	13	12

Si une note reflète la « proximité » entre un élève et une discipline, son complément à 20 sera une dissimilarité. Nous prenons donc comme tableau T de donnée, le tableau des compléments à 20.

T =

	Math.	Fr.	Hist.	Ed. Ph.
a	4	6	8	8
b	5	10	12	15
c	9	8	8	7
d	10	8	10	8
e	12	16	10	7
f	12	7	7	8

Le couple (α, ω) le plus proche est le couple (a, Mathématique).

On en déduit le vecteur (2, 3, 7, 8, 10, 10) (2, 4, 6, 6), (voir § 7.6) puis suivant l'algorithme précédent on calcule l'ultramétrique rectangulaire U_R (K_o = 10 et fonction d'agrégation = max).

$U_R =$

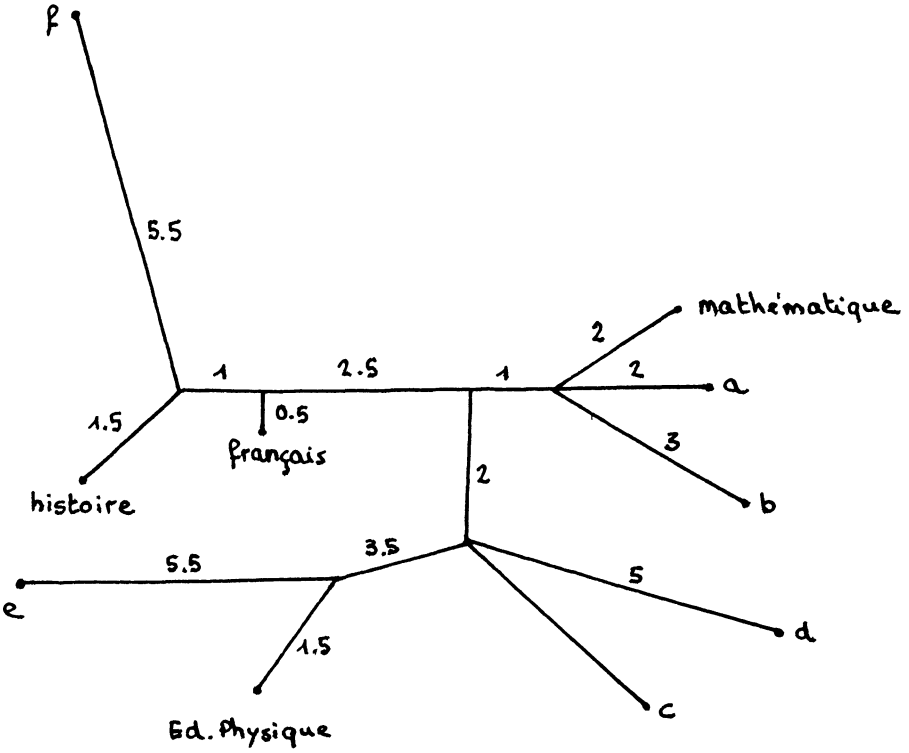
	Math.	Fr.	Hist.	Ed. Ph.
a	10	10	10	10
b	10	10	10	10
c	10	8	8	4
d	10	8	8	4
e	10	8	8	1
f	10	3	1	8

On en déduit la matrice U puis la matrice Q.

$Q =$

	a	b	c	d	e	f	Ma	Fr	Hi	E.P.
a	0	5	9	10	12	12	4	6	8	8
b		0	10	11	13	13	5	7	9	9
c			0	9	11	15	9	9	11	7
d				0	12	16	10	10	12	8
e					0	18	12	12	14	7
f						0	12	7	7	14
Math.							0	6	8	8
Fr.								0	1	8
Hist.									0	10
Ed. P.										0

Que l'on représente par l'arbre additif suivant :

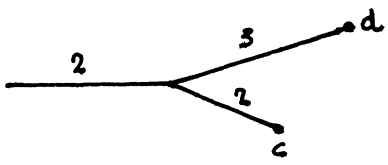


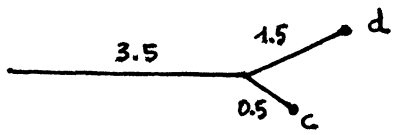
D'où l'on peut extraire les sous arbres sur les élèves ou sur les disciplines.

Remarques sur l'interprétation des résultats

1) On remarque que les arêtes des disciplines sont plutôt courtes alors que celles qui vont aux élèves sont plutôt longues. Ceci est agréable (l'idéal serait d'avoir les disciplines en nœuds de l'arbre) mais ceci est illusoire. En effet l'arbre n'est connu qu'à une constante K près : on peut augmenter les longueurs des arêtes arrivant à un ensemble d'une longueur arbitraire K à condition de réduire les arêtes conduisant à l'autre ensemble d'une longueur K.

2) D'autre part, nous obtenons deux couples de lignes identiques (a et b, c et d) dans l'ultramétrique rectangulaire calculée à l'étape intermédiaire. Sur le graphique on a représenté les couples (c, d) et (a, b) sans nœud intermédiaire propres alors que l'on aurait pu les représenter sous la forme suivante :





Ce qui fait que la distance (c, d) est indéterminée entre 1 et 9, et la distance (a, b) entre 1 et 5.

3) On peut comparer directement le tableau des distances représentée par l'arbre additif et le tableau initial.

	Math.	Fr.	Hist.	Ed. Ph.
a	4	6	8	8
b	5	10	12	15
c	9	8	8	7
d	10	8	10	8
e	12	16	10	7
f	18	7	7	8

Tableau initial

	Math.	Fr.	Hist.	Ed. Ph.
a	4	6	8	8
b	5	7	9	9
c	9	9	11	7
d	10	10	12	8
e	12	12	14	7
f	12	7	7	14

Tableau représenté

On remarque de suite que la première ligne et la première colonne sont exactement représentées. Ceci est dû à l'algorithme qui conserve les distances issues des deux racines (a et Math).

D'une façon générale la comparaison est aisée et on peut mesurer la distorsion imposée par la condition (ar).

EN GUISE DE CONCLUSION

Tous les tableaux rectangulaires pouvant s'interpréter en terme de distance, de proximité, de ressemblance ou de similarité peuvent être représentés par ce type de méthode.

On peut penser aux tableaux exprimant directement une distance entre deux ensemble : par exemple, « a » est une industrie, ou un lieu de consommation en général, « x » est un centre de production de matière première, ou un lieu d'approvisionnement en général, et T_{ax} mesure une distance entre ces deux unités.

Mais on trouve beaucoup de tableaux exprimés en terme de proximité tels les tableaux de préférence, où « a » est un consommateur, « x » un produit et T_{ax} une appréciation. D'une façon générale rentrent dans cette catégorie tous les tableaux d'opinions (ensemble d'objets, ensemble de juges, et opinion).

Une catégorie particulière est formée par les tableaux de contingences, ou « a » est une modalité d'une variable, « x » une modalité de l'autre

variable et T_{ax} est le nombre d'unité statistiques appartenant à ces 2 catégories. Alors on peut considérer T_{ax} comme la mesure d'une proximité entre « a » et « x ».

Quand le tableau de donnée T_{ax} peut s'interpréter en ces termes alors l'utilisation de cette approche fournit de nombreux avantages :

— on obtient une représentation arborée simultanée des deux ensembles,
— on peut en extraire deux représentations correspondant aux classifications par chacun des 2 ensembles,
— dans l'approche classique il est nécessaire pour passer d'un tableau rectangulaire à un tableau de dissimilarité sur un ensemble de définir et de calculer un indice de dissimilarité ou de distance. La classification produite dépendant évidemment de ce choix il est intéressant de constater que dans cette approche il n'y a pas de choix arbitraires intermédiaires,
— la comparaison entre les données initiales et celles représentées est facile.

BIBLIOGRAPHIE

[1] Gildas BROSSIER. — Approximation optimale d'une matrice de distance par un arbre additif, *Math. et Sc. Humaines,* n° 91, pp. 5-21, 1985.

[2] C.H. COOMBS. — A theory of data, New-York, Wiley, 1964.

[3] G.W. FURNAS. — *Objectifs and their features : the metric representation of two class data.* Unpublished Ph. D., Standford University, 1980.

[4] G. DE SOETE, W. DE SARBO, G.W. FURNAS, J.D. CAROLL. — The estimation of ultrametric and path lenglh trees from rectangular proximity data. *Psychométrika,* Vol. 49, n° 3, pp. 289, 310, 1984.

[5] G. GOVAERT. — *Classification croisée.* Thèse d'Etat, université Paris 6, 1983.

[6] S. SATTAH, A. TVERSKY. — Additive similarity trees. *Psychometrika,* Vol. 42, n° 3, 1977.