Statistique et analyse des données

Jean Pierre Barthelemy
Nhuan Xuan Luong

Représentations arborées de mesures de dissimilarité

Statistique et analyse des données, tome 11, n° 1 (1986), p. 20-41
<http://www.numdam.org/item?id=SAD_1986__11_1_20_0>

Statistique et Analyse des Données
1986 - Vol. 11 n°1 - pp.20-41

REPRESENTATIONS ARBOREES DE MESURES DE DISSIMILARITE

Jean Pierre BARTHELEMY

ENST Département d'Informatique
46, rue Barrault
75634 PARIS CEDEX 13

Nhuan Xuan LUONG

URL9-INLF
Faculté des Lettres et
Sciences Humaines 98 bd
Herriot 06007NICE CEDEX

Résumé : *Dans ce travail nous rappellerons d'abord quelques propriétés des distances arborées et nous évoquons les algorithmes classiques d'approximation d'une dissimilarité par une telle distance. Nous présenterons ensuite un nouvel algorithme qui détermine une représentation arborée à partir de considérations "topologiques". Cet algorithme permet de traiter d'assez grands tableaux de données sur un micro-ordinateur.*

Abstract : *This paper is threefold: first we recall some basic properties of the additive tree representation; then we evoke the classical algorithms for the additive tree approximation of a dissimilarity matrix; eventually a "topological" point of view is used to develop a new algorithm allowing one to deal with same rather big data sets even with a personnal computer.*

Indices de classification STMA : 06-010, 06-900 .

Mots clés : *Arbre additif , algorithme de représentation, dissimilarité, distance arborée, classification hiérarchique, représentation arborée.*

Manuscrit reçu le 7.8.1985, révisé le 12.6.1986

0 INTRODUCTION

Les arbres ultramétriques (ou dendrogrammes) constituent le modèle métrique le plus courant en taxionomie. Ceux-ci sont des arbres plantés (i.e. avec un sommet distingué appelé racine), aux arêtes valuées par des réels positifs et tels que les longueurs du chemin de la racine à chacune des feuilles (ou sommets terminaux) sont égales. Définissons la distance entre deux sommets d'un arbre valué comme la longueur du chemin qui les joint. Dans un dendrogramme, tous les sommets d'une même classe, i.e. dominés par un même noeud, sont équidistants entre eux (cf figure 1). La distance d entre les feuilles vérifie alors l'inégalité ultramétrique

pour toutes feuilles x,y,z : $d(x,z) \leq \text{Max}(d(x,y),d(y,z))$.

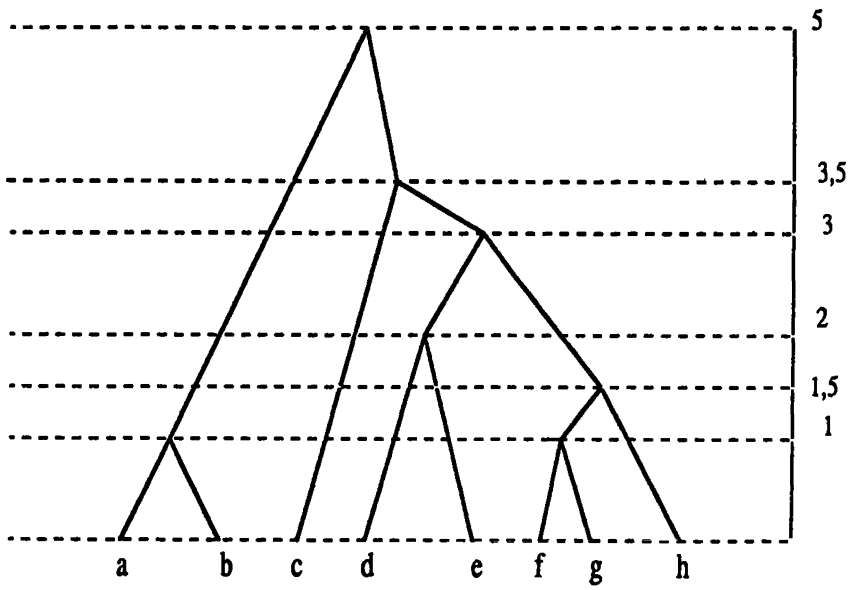


figure 1 : arbre ultramétrique sur l'ensemble $\{a,b,c,d,e,f,g,h\}$.

Il y a au moins deux avantages au modèle ultramétrique:
1) La notion de classe se laisse définir sans ambiguité. Puisque l'arbre est planté, ces classes se lisent, de la plus grossière aux plus fines en parcourant ses noeuds de la racine aux feuilles.
2) On obtient un indice donnant un niveau de formation d'une classe. Il s'agit simplement de la distance d'une feuille à un noeud; la propriété d'équidistance évoquée plus haut rendant cette définition légitime (cf figure 1).

Ce type de classification, parfois qualifiée d'aristotélicienne (mais c'est, bien sûr, le nom de Porphyre qui devrait être associé aux descriptions à l'aide d'arbres hiérarchiques), s'est trouvé, dans certains domaines et pour certains objectifs, supplanté par des modèles plus souples et moins infidèles. Ces modèles, qualifiés, selon les contextes, d'arbres phylogénétiques (non orientés), d'arbres additifs, de distances arborées ("tree metric"), d'arbres valués partiellement étiquetés, d'arbres de Buneman, de X-arbres, de "tree codes", sont, au niveau technique, apparus d'abord en traitement de l'information à travers de problème de codage et de décodage (Zaretskii, 1965; Smolenskii, 1963 ; cette tradition demeure d'ailleurs encore très vivante, cf. Dewdney, 1979 ; Chaiken, Dewdney & Slater, 1983). Notons que, pratiquement à la même époque, on les rencontre en recherche opérationnelle (Hakimi & Yau, 1964 ; Patrinos & Hakimi, 1972), pour des problèmes d'approximation de réseau.

Au delà de la technique mathématique mise en jeu, on pourrait évoquer les idées de philosophes, comme Wittgenstein, 1953, qui désirèrent substituer à la notion, trop rigide, de classe celle d'"air de famille". Cette notion n'était d'ailleurs pas neuve. On la retrouve, par exemple, chez Galton, 1879, à propos de ses "photographies mentales mélangées".

Mais ce sont surtout les naturalistes et les psychologues qui ont, récemment, revivifié ce type de modèle.

Les premiers,intéressés par les phénomènes d'évolution et/ou de bifurcation, ont tendance depuis quelques temps à remplacer les arbres hiérarchiques par des arbres "phylogénétiques" non orientés (cf. Cavalli - Sforza & Edward, 1967; Waterman, Smith, Singh & Beyer, 1977; Robinson & Foulds, 1981 ...).

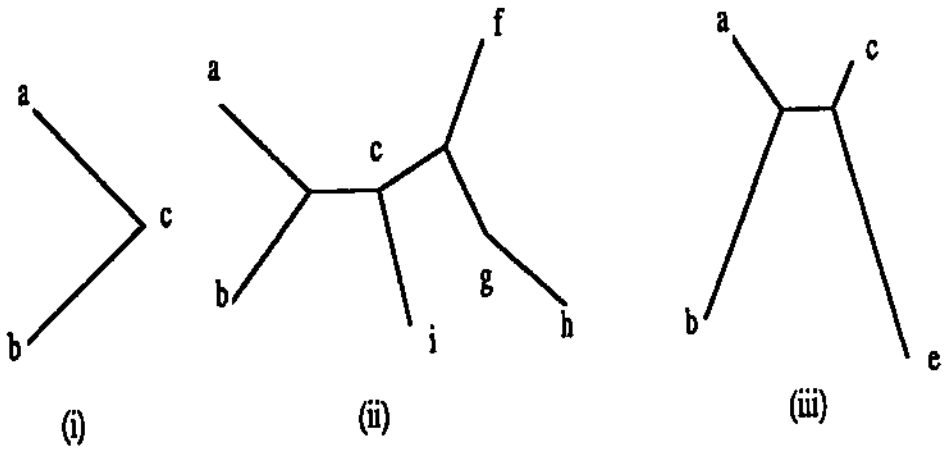


figure 2

Le schéma représenté sur la figure 2-(i) signifie que c est le "premier ancêtre commun à a et b ". L'absence d'orientation de l'arbre correspond à une absence d'hypothèse sur le "sens" de l'évolution. Ces arbres sont partiellement étiquetés (cf. figure 2-(ii)). Aux sommets étiquetés correspondent les "espèces" répertoriées; aux sommets sans étiquette correspondent les "chaînons manquants". En particulier, les feuilles de l'arbre représentent des espèces observables aujourd'hui (i.e. qui ne sont pas "ancêtres") et peuvent légitimement être supposées étiquetées. Dès lors se pose le problème de représenter par un arbre des proximités calculées à l'aide de caractères communs et spécifiques entre espèces. Un tel arbre fournira un modèle d'évolution. Au niveau de l'analyse des données d'intéressants résultats ont été obtenus récemment sur la comparaison et le consensus de ces arbres phylogénétiques (Estabrook, McMorris & Meacham, 1985; McMorris, 1985; Day, 1986).

Notons que ce type de représentation a été également utilisé, dans un esprit très proche pour la filiation et l'analyse des textes (Buneman, 1971; Luong & Novi, 1986).

En psychologie mathématique et en psychométrie, on considère les "arbres additifs" qui, formellement, correspondent aux arbres phylogénétiques (figure 2-(ii)). Cette pratique est à la mode depuis les travaux de Tversky, 1977, sur le modèle du contraste complétés par l'algorithme de Sattah & Tversky, 1977. Sa puissance vient du fait que ces arbres, venant en "sortie" d'une méthode d'analyse de la similitude rendent compte remarquablement de modèles et d'hypothèses usuels dans les travaux sur la mémoire et les catégories naturelles. L' abandon de la contrainte ultramétrique permet, en particulier, de différencier les objets d'une même classe selon un gradient de représentativité et de rendre compte de la notion de typicalité. (Cf. pour la psychologie mathématique : Colonius & Schulze, 1979, 1981; Abdi, Barthélemy & Luong, 1984; Abdi, 1985,... et pour la psychométrie : Carroll, 1976; Carroll, Clark & Desarbo, 1984; Carroll & Pruzansky, 1980; Cunningham, 1978,1980; De Soete, 1983 (a) et (b); De Soete, Desarbo, Furnas, 1984,...).

Le modèle métrique sous-jacent à ces diverses approches est celui d'une distance arborée: une distance arborée sur un ensemble X est une distance d sur X telle qu'il existe un arbre valué A dont X contient l'ensemble des feuilles, vérifiant:

d_{xy} = longueur du chemin de A entre x et y .

L'arbre A est alors appelé une représentation de d .

Le rapport entre proximité et classification devient alors ambigu - mais beaucoup plus riche - que dans le cas ultramétrique. Par exemple dans l'arbre de la figure 2-(iii), au sens de la distance. c est plus proche de a qu'il ne l'est de e . Cependant la "topologie" de cet arbre nous suggère deux classes $\{a,b\}$ et $\{c,e\}$!

Nous voyons là une illustration de la notion d'air de famille évoquée plus haut. Et nous ne résistons pas au plaisir de citer Wittgenstein "... innerhalb der Familie gibt es eine Familienähnlichkeit ; während es auch zwischen Mitgliedern verschiedener Familien eine Ähnlichkeit gibt..." (Vermischte Bemerkungen.) .

Le problème classique en taxionomie, de l'approximation d'une dissimilarité par une ultramétrique, peut être généralisé de la manière suivante: trouver une distance arborée qui rend compte au mieux des dissimilarités δ_{xy} .

1 CARACTERISATION DES DISTANCES ARBOREES.

Buneman, 1974, et Dobson, 1974, après Zaretskii, 1965, ont caractérisé les distances arborées:

Théorème des quatre points. Une mesure de dissimilarité δ sur X est une distance arborée si et seulement si elle vérifie la condition ci-dessous, dite condition des quatre points, pour tout $x,y,z,t \in X$:

$$\delta_{xy} + \delta_{zt} \leq \text{Max} (\delta_{xz} + \delta_{yt} , \delta_{xt} + \delta_{yz})$$

En fait, cette condition est souvent apparue, sous des variantes plus ou moins proches chez un certain nombre d'auteurs, au cours de la période 1969-1974 . Citons au moins les repères suivants:

- Simoes-Pereira, 1969, montre que δ est une distance arborée sur X si et seulement si la restriction de δ à tout sous-ensemble à quatre éléments de X est arborée.
- Buneman, 1971, obtient la condition des quatre points en examinant la structure de l'ensemble des bipartitions sur X , obtenues en gommant les arêtes d'un arbre dont l'ensemble des feuilles est contenu dans X .
- Patrinos et Hakimi, 1972, prouvent que δ est une distance arborée sur X si et seulement si pour tout x,y,z,t de X les deux plus grands des trois nombres $\delta_{xy} + \delta_{zt}$, $\delta_{xt} + \delta_{yz}$, $\delta_{xz} + \delta_{yt}$ sont égaux.

2 ALGORITHMES DE REPRESENTATION ARBOREE

A partir de la condition des quatre points se développent depuis 1974 deux familles d'algorithmes de représentation arborée. La première recherche directement la distance arborée, souvent par des méthodes de

programmation mathématique, déterminant une distance d qui minimise

$$\mathcal{L} = \sum_x \sum_y (\delta_{xy} - d_{xy})^2$$

sous de contraintes diverses. La seconde recherche d'abord une structure d'arbre la plus "appropriée", puis estime les distances de différentes manières. Notons que, restreint aux distances à valeurs entières et aux arbres valués par des entiers, minimiser $\mathcal{L}$ est un problème NP-difficile.

2-1 Recherches directes de la distance arborée.

2-1-1 Cunningham, 1978, présente un algorithme pour minimiser $\mathcal{L}$, avec un critère fondé sur la présomption que pour tout x,y,u,v de X la distance arborée d satisfait à

$$d_{xy} + d_{uv} \leq d_{xu} + d_{yv} = d_{xv} + d_{yu}$$

chaque fois que l'on a

$$\delta_{xy} + \delta_{uv} \leq \delta_{xu} + \delta_{yv} = \delta_{xv} + \delta_{yu} \ .$$

Cependant la distance arborée ainsi obtenue n'est pas nécessairement optimale au sens des moindres carrés (cf. par exemple Carroll & Pruzansky, 1980).

2-1-2 Carroll & Pruzansky, 1980, proposent une méthode hybride, utilisant le fait qu'une distance arborée peut être toujours décomposée en une somme de distances ultramétriques et de constantes (selon la terminologie de ces auteurs). Ces constantes constituent ce que Le Calvet, 1986, appelle une distance à centre. Cette méthode utilise les moindres carrés alternés où à chaque pas on se ramène à un problème de minimisation d'une expression du type

$$\sum_x \sum_y (\delta'_{xy} - d'_{xy})^2 \ , \text{ avec des contraintes ultramétriques.}$$

2-1-3 De Soete, 1983-a, utilise une méthode de pénalisation pour renforcer la condition des quatre points sur les dissimilarités. La fonction de pénalisation est

$$\mathcal{J} = \sum_{\Xi} (d_{xu} + d_{yv} - d_{xv} - d_{yu})^2$$

où Ξ est l'ensemble des x,y,u,v distincts vérifiant

$$d_{xy} + d_{uv} \leq \text{Min}(d_{xu} + d_{yv} , d_{xv} + d_{yu}) \quad .$$

La méthode revient à minimiser la fonction $\mathcal{L} + r\mathcal{J}$ pour une suite croissante de valeurs de r .

2-1-4 Les algorithmes précédents ont besoin de moyens de calcul puissants. Celui proposé par Cunnigham a une solution analytique qui nécessite l'inversion d'une matrice de type $\binom{n}{4} \times \binom{n}{4}$. Les deux autres demandent en plus une programmation assez complexe.

2-1-5 Roux, 1985, partant de la relation d'ordre : d d' si et seulement si pour tout x,y,z,t, avec x=y , z=t , $d_{xy} + d_{zt}$ $d'_{xy} + d'_{zt}$, s'inspire de son algorithme de réduction des triangles, Roux 1968, pour obtenir, par réduction des quadruplets, une distance arborée inférieure à une dissimilarité donnée.

2-1-6 On trouvera dans Brossier 1985 une étude des relations entre distances arborées, distances ultramétriques et distances à centre, ainsi qu'une utilisation algorithmique de la décomposition mentionnée en 2-1-2.

2-2 Choix de la structure de l'arbre.

2-2-1 Sattah & Tversky, 1977, déterminent d'abord la structure de l'arbre. C'est un arbre binaire, construit de manière itérative, fondé sur une une notion de voisinage dérivant de la condition des quatre points. Si x,y,u,v de X vérifient

$$(2\text{-}1) \quad \delta_{xy} + \delta_{uv} \leq \delta_{xu} + \delta_{yv} \quad \text{et} \quad \delta_{xy} + \delta_{uv} \leq \delta_{xv} + \delta_{yu}$$

on dit que x,y sont des "voisins lâches" relativement à $\{u,v\}$. A chaque pas d'itération, on examine tous les quadruplets des éléments de X pour en dégager une paire $\{x,y\}$ qui vérifie le plus grand nombre des relations (2-1) pour u,v de $X-\{x,y\}$. A l'itération suivante x et y viennent former un seul élément z dont les dissimilarités δ_{zu}, pout tout u de $X-\{x,y\}$, sont définies par $(\delta_{xu} + \delta_{yu})/2$.

On procède ainsi jusqu'à l'obtention complète de l'arbre. L'estimation de la longueur des arcs de cet arbre est obtenue par les moindres carrés.

2-2-2 Dans un travail récent (cf. Abdi, Barthélemy & Luong, 1984) nous avons introduit un algorithme également fondé sur cette notion de voisinage lâche. On appelle score d'une paire $\{x,y\}$ de X, noté $s(x,y)$, le nombre de paires u,v de $X - \{x,y\}$ vérifiant les inégalités (2-1). On calcule les scores de toutes les paires de X. On sélectionne une paire dont le score soit maximum, soit $\{x,y\}$. On regroupe x,y pour avoir un objet z qui les représente et on pose :

$$s(z,u) = (s(x,u) + s(y,u))/2 \quad \text{pout tout } u \text{ de } X - \{x,y\} .$$

On réitère le processus jusqu'à l'obtention complète de l'arbre. La structure de cet arbre binaire est, en quelque sorte, obtenue par une méthode d'agrégation par paires, avec le critère de la moyenne ("average linkage", cf. Sokal & Michener, 1958) sur le tableau des scores.

Pour l'estimation des arcs, on introduit un point de dissimilarité moyen g, avec

$$\delta_{gu} = (\sum_{v \in X} \delta_{vu})/n \quad \text{, pour tout } u \text{ de } X .$$

On interprète (abusivement!) les dissimilarités dans un espace euclidien, et avec l'hypothèse que z, représentant de $\{x,y\}$, se rapproche de g, on peut ainsi déterminer "géométriquement" les distances sur l'arbre.

3 UN NOUVEL ALGORITHME DE REPRESENTATION ARBOREE.

Nous introduisons d'abord la notion de groupement qui généralise les voisinages de 2-1. Ces groupements ont une double utilité; ils accélèrent les algorithmes fondés sur un principe de fusion (cf 2) et ils permettent d'obtenir des arbres non binaires.

Nous caractérisons ensuite ces groupements par une propriété remarquable qui nous permettra d'obtenir une procédure de reconstitution de l'arbre. L'algorithme de représentation arborée s'inspirera directement de cette procédure.

3-1 Groupement sur un arbre.

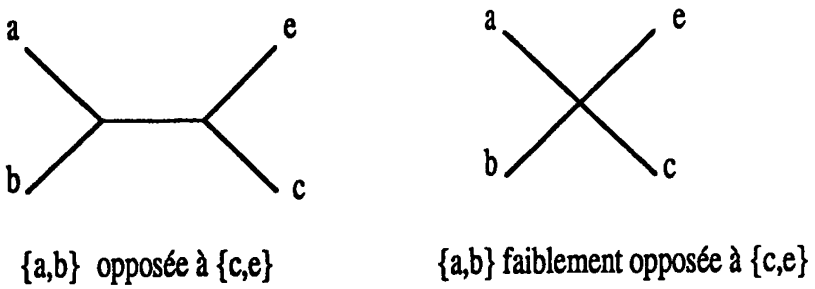
Soit A un arbre. Supposons que $X = \{ a,b,c,... \}$ est l'ensemble des n feuilles de A ($n > 3$). Considérons quatre feuilles distinctes a,b,c,e de X. Si l'on a

(1) $d_{ab} + d_{ce} < d_{ac} + d_{be} = d_{ae} + d_{bc}$

on dit que $\{a,b\}$ et $\{c,e\}$ sont <u>opposées</u>. On le note par ab/ce .
Si l'on a

(2) $d_{ab} + d_{ce} = d_{ac} + d_{be} = d_{ae} + d_{bc}$

on dit que $\{a,b\}$ et $\{c,e\}$ sont <u>faiblement opposées</u>.
La figure 3 illustre ces situations.



<u>figure 3</u> : chaque arc représente un chemin dans l'arbre.

Soient a,b deux éléments fixés de X :
on appelle <u>score</u> de a,b et on note $\S(a,b)$, le nombre de paires $\{c,e\}$ d'éléments de X vérifiant (1) . Une telle paire est dite <u>scorante</u> à a,b ,
on appelle <u>score affaibli</u> de a,b , noté $s(a,b)$, le nombre de paires $\{c,e\}$ d'éléments de X vérifiant (1) ou (2) . Une telle paire $\{c,e\}$ est dite <u>faiblement scorante</u> à a,b .

<u>Définition</u>. Soit G un sous-ensemble de X , $|G| \geq 2$, on dit que G est un <u>groupement</u> si l'on a une des deux conditions suivantes :
(3) Pour tout $a,b \in G$ et pour tout $x,y \in \bar{G}$, $\bar{G} = X-G$, on a ab/xy ; si de plus $a,b \in G$ et $c,e \in X$ sont tels que ab/ce alors $c \in \bar{G}$ et $e \in \bar{G}$.
(4) Pour tout $a,b \in G$ il n'existe pas de $x,y \in X$ tel que ab/xy .

Remarquons que (4) indique une dégénérescence: $G = X$ (et l'arbre est une étoile).

Propriété.

Soit G un sous-ensemble à k éléments de X , G est un groupement si et seulement si pour tout a,b ∈ G , avec a≠b :

(5) $\hat{s}(a,b) = (n-k)(n-k-1)/2$ et

(6) $s(a,b) = (n-2)(n-3)/2$.

Preuve: C'est clair pour k=n (l'arbre est une étoile) et trivial pour k=2 (s et ŝ coïncident). Supposons donc $2 < k < n$ et considérons un groupement G à k éléments. En vertu de (3) $\hat{s}(a,b)$ est le nombre de paires de X-G . Donc $\hat{s}(a,b) = (n-k)(n-k-1)/2$. Par ailleurs, pour a,b,c (distincts) dans G et pour $x \in X$, on a jamais ab/cx (toujours en vertu de (3)). En permutant les éléments a,b,c et x on en déduit que, nécessairement

$$d_{ab} + d_{cx} = d_{ac} + d_{bx} = d_{ax} + d_{bc} .$$

Il s'ensuit que toute paire $\{x,y\}$ d'éléments de X distincts de a et b sera faiblement scorante à $\{a,b\}$. D'où (6).

Réciproquement, soit G un sous-ensemble à k éléments de X ,vérifiant (5) et (6) . D'après (6) toute paire de X- $\{a,b\}$ est faiblement scorante à $\{a,b\}$. Supposons que l'on a a,b.c ∈ G et $x \in X$ tels que ab/cx. Les trois assertions suivantes sont alors fausses : ac/bx , ax/bc , $d_{ab} + d_{cx} = d_{ac} + d_{bx} = d_{ax} + d_{bc}$.

On en déduit que $s(a,c) < (n-2)(n-3)/2$. Ce qui contredit (6). Il vient donc ab/xy et a,b ∈ G entraîne $x \in \bar{G}$ et $y \in \bar{G}$. Comme par ailleurs, $s(a,b) = (n-k)(n-k-1)/2$ les seules paires scorantes pour $\{a,b\} \subseteq G$ seront les paires d'éléments de $\bar{G}$. La condition (3) est vérifiée et G est un groupement.

3-2 Un algorithme de reconstitution d'un arbre.

Soit $D = (d_{xy})$ une matrice de distance arborée. Nous utilisons la notion de groupement pour reconstruire l'arbre de manière itérative.

3-2-1 Algorithme 1

i) Calcul des scores et des scores affaiblis de toutes les paires de X.

ii) Dégager les groupements, c'est à dire les sous-ensembles dont les scores et les scores affaiblis sont maximaux
($\mathcal{S} = (n-k)(n-k-1)/2$ et $s = (n-2)(n-3)/2$) .

iii) Estimer la longueur des arcs dans chaque groupement.
Chaque groupement est représenté par un objet "moyen".

iv) Avec les feuilles restantes et ces objets moyens réitérer le processus jusqu'à l'obtention complète de l'arbre.

3-2-2 Remarques.

- Soient deux feuilles u et v liées directement à un sommet intérieur p (i.e. avec une arête entre u et p et une arête entre v et p) :

$$d_{up} = (d_{uv} + d_{uh} - d_{vh})/2 \quad .$$

De cette manière, on évalue les arcs de chaque groupement.

- Un objet moyen z d'un groupement G , $|G| = k$, est déterminé par $d_{zu} = \left(\sum_{x \in G} d_{xu} \right)/k$, pour tout u de $X-G$.

- Il est clair qu'un arbre étant donné, il admet au moins un groupement. A chaque itération le nombre d'éléments à traiter va en diminuant : l'algorithme est convergent !

3-3 Un algorithme de représentation arborée.

3-3-1 Scores et groupements.

Les données constituent maintenant la matrice de dissimilarité $\Delta = (\delta_{xy})$. La notion d'opposition s'obtient à l'aide de : $\{x,y\}$ est opposée à $\{u,v\}$ lorsque

(7) $\delta_{xy} + \delta_{uv} < \text{Min}(\delta_{xu} + \delta_{yv} , \delta_{xv} + \delta_{yv})$.

On définit alors les scores, comme en 3-1 : le score de x,y est le nombre de paires opposées à $\{x,y\}$. Le score affaibli de x,y est le

nombre de paires $\{u,v\}$ d'éléments de X vérifiant (7) lorsque l'on y remplace l'inégalité stricte par l'inégalité large. Un groupement est alors un sous-ensemble de X dont les scores vérifient les conditions de la propriété 1 (i.e. (5) et (6)). D'autre part, soit un groupement G , $|G| = k$ et soient $a,b \in G$ et un x quelconque de $\bar{G}$, on peut montrer que

(8) $\hat{s}(a,b) - \hat{s}(a,x) \geq n-k-1$,

ainsi dans la constitution des groupements on peut utiliser la majoration (8) pour avoir un certain seuil de tolérance.

S'il n'existe pas de groupement à une étape donnée, on va grouper une paire dont le score est maximum. On peut ainsi utiliser les bases de l'algorithme précédent.

3-3-2 Algorithme 2.

i) Calcul des scores et des scores affaiblis de toutes les paires de X.

ii) Dégager les groupements. S'il n'en existe pas, agréger les paires dont le score est maximal.

iii) Estimer la longueur des arcs. Chaque groupement ou chaque paire agrégée est représenté par un objet moyen.

iv) Avec les objets restants et les objets moyens, réitérer le processus jusqu'à l'obtention complète de l'arbre.

3-3-3 Remarques.

- Si l'on a regroupé deux objets u,v , la longueur de l'arc up , p étant le sommet intérieur lié directement à u , est estimée par

$$d_{up} = \left(\delta_{uv} + \frac{1}{n-2} \sum_{h \neq u,v} (\delta_{uh} - \delta_{vh}) \right)/2$$

- Un objet moyen z d'un sous-ensemble G de X , $|G| = k$, est défini par

$$\delta_{zu} = \left(\sum_{x \in G} \delta_{xu} \right)/k \quad .$$

- Soit s' le score calculé d'une paire $\{x,y\}$ d'un sous-ensemble G de X, G est soit un groupement, soit une paire de score maximal. Soit s le score théorique, c'est à dire le score que doit avoir x,y s'ils appartiennent tous les deux à un groupement. On considère alors comme indice de qualité arborée, le rapport s'/s , appelé indice d'agrégation . Cet indice permet de mesurer la manière dont les objets vont s'agréger. Si s'/s = 1 alors $\{x,y\}$ respecte parfaitement la condition des quatre points dans tout quadruplet d'éléments de X où $\{x,y\}$ fait partie.

3-3-4 Commentaires.

Cet algorithme s'inscrit donc dans la lignée des algoritmes décrits en 2-2 .

Il a, dans cette famille, l'avantage d'être:

- exact sur les distances arborées;
- plus rapide en pratique (grâce à l'utilisation des groupements).

De plus, il permet d'obtenir des arbres non binaires et fournit des indices de qualité des noeuds de l'arbre dont l'utilisateur peut utilement tenir compte.

Enfin, une fois l'arbre construit, on peut toujours ré-évaluer ses arêtes de manière à ce que la distance arborée induite approche au mieux (mais relativement à cet arbre là) , au sens des moindres carrés. Il ne s'agit plus que d'un simple problème d'optimisation quadratique sous contrainte de positivité. On obtient ainsi une heuristique pour le problème (rappelons-le, NP-difficile!) de l'approximation directe.

4 EXEMPLES

Nous donnons deux exemples de représentation arborée obtenue à partir de l'algorithme 2 .

4-1 Les données de Friendly.

Dans la table 1 se trouve la matrice des proximités entre 18 substantifs obtenue par Friendly, 1979, dans une expérience de rappel libre (free recall). Celui-ci en propose une représentation hiérarchique (cf figure 4).

100																
91	93															
88	91	100														
86	88	92	91													
88	89	94	94	100												
55	57	57	57	52	51											
62	65	62	61	60	56	95										
63	66	65	65	61	60	90	92									
62	64	63	63	56	58	94	94	93								
58	61	62	62	56	57	89	92	96	90							
55	57	59	59	51	53	93	90	92	91	98						
69	67	56	56	61	63	60	60	57	62	48	49					
69	68	60	58	65	64	61	62	60	62	56	57	96				
70	69	60	58	64	64	60	61	59	62	54	55	94	97			
68	67	60	59	64	65	65	65	64	67	60	60	87	95	93		
66	65	54	53	53	58	66	66	65	70	58	60	88	86	85	92	
66	65	57	56	60	62	63	64	62	67	58	57	91	93	92	97	94

dog	1	*face*	7	*celery*	13
cat	2	*mouth*	8	*spinach*	14
lion	3	*arm*	9	*lettuce*	15
tiger	4	*finger*	10	*potato*	16
sheep	5	*leg*	11	*rice*	17
goat	6	*foot*	12	*yam*	18

table 1: les données de FRIENDLY

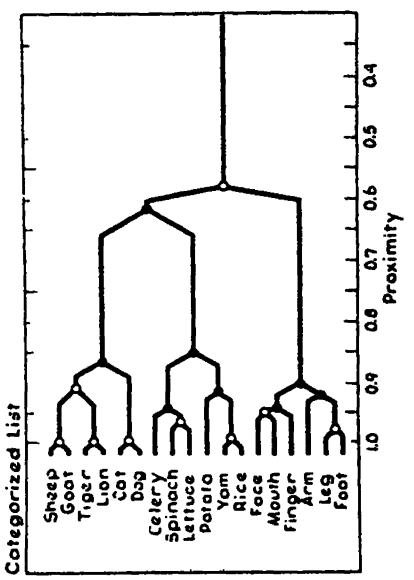


figure 4 : représentation hiérarchique.

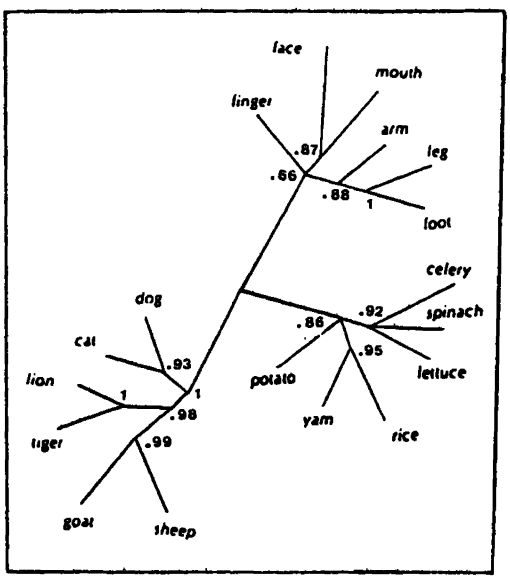


figure 5 : représentation arborée par l'algorithme 2

Reprises par Abdi, Barthélemy & Luong, 1984, à l'aide de l'algorithme présenté en 2-2-2, ces données ont été traitées, pour cet article, par l'algorithme 2 (3-2-2) .

On obtient alors la représentation arborée de la figure 5 . Les chiffres indiqués sont les indices de qualité arborée des sommets intérieurs. Nous distinguons dans cette représentation nettement trois groupes, chacun correspond à une des trois familles naturelles: "animaux", "végétaux" et "corps humain". La formation du groupe des animaux est nette, avec des indices d'agrégation proches de 1. Le groupe des végétaux est assez homogène. Le groupe du corps humain est moins bien structuré, avec certains indices d'agrégation assez faibles.

4-2 Données climatologiques.

Nous citons sommairement un travail fait en commun avec D. Joly (cf. Joly & Luong, 1985), sur des données provenant des observations climatiques collectées au Spitsberg (cf. Joly,1980).

Les observations sont à intervalles réguliers, chacune d'elles est composée des descripteurs météorologiques habituels. Elles sont classées en 24 catégories appelées ambiances climatiques instantanées (ACI), voir la figure 6. On va étudier les modalités du temps par le biais des liaisons des ACI les unes vers les autres. Nous représentons ici des représentations arborées des liaisons postérieures.

La figure 7 montre les ACI à six heures d'intervalle. On y remarque trois grandes familles :

- la famille 1 : le "mauvais temps" accompagné de vents faibles ou moyens,
- la famille 2 : le "beau temps" accompagné de vents faibles ou moyens.
- la famille 3 : "les vents forts", ainsi que certains ACI isolées qui s'agrègent à proximité du centre de l'arbre.

La figure 8 donne des liaisons postérieures à intervalle de 12, 18, 24, 48 et 120 heures. Nous remarquons que les structures des liaisons à 6 heures sont conservées à 12 heures, en dehors du fait que la famille 1 est dorénavent celle du mauvais temps associé exclusivement aux vents

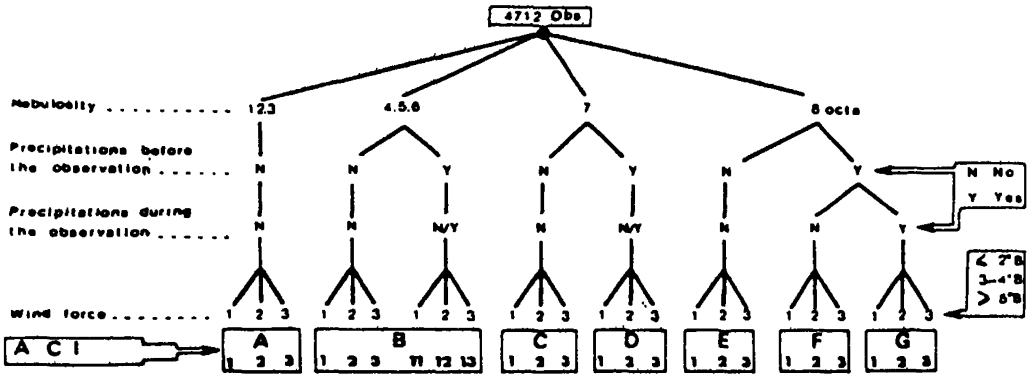


figure 6 : Ambiances climatiques instantanées.

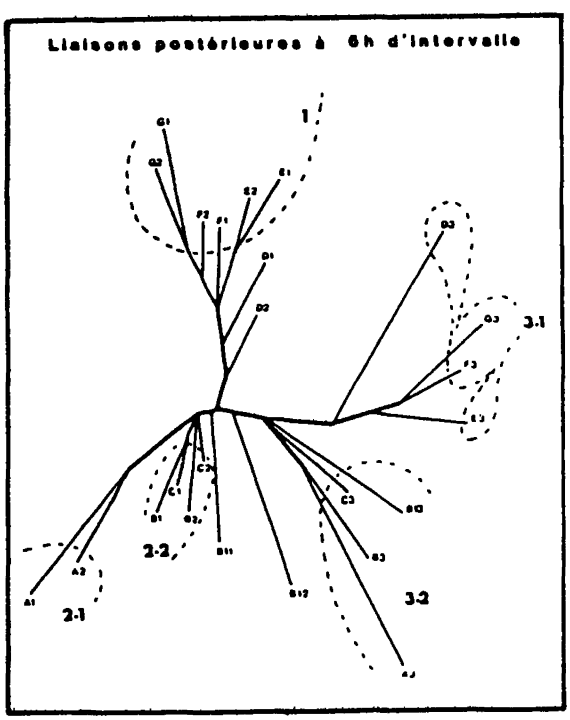


figure 7 :
liaisons postérieures des ACI à 6 heures d'intervalle.

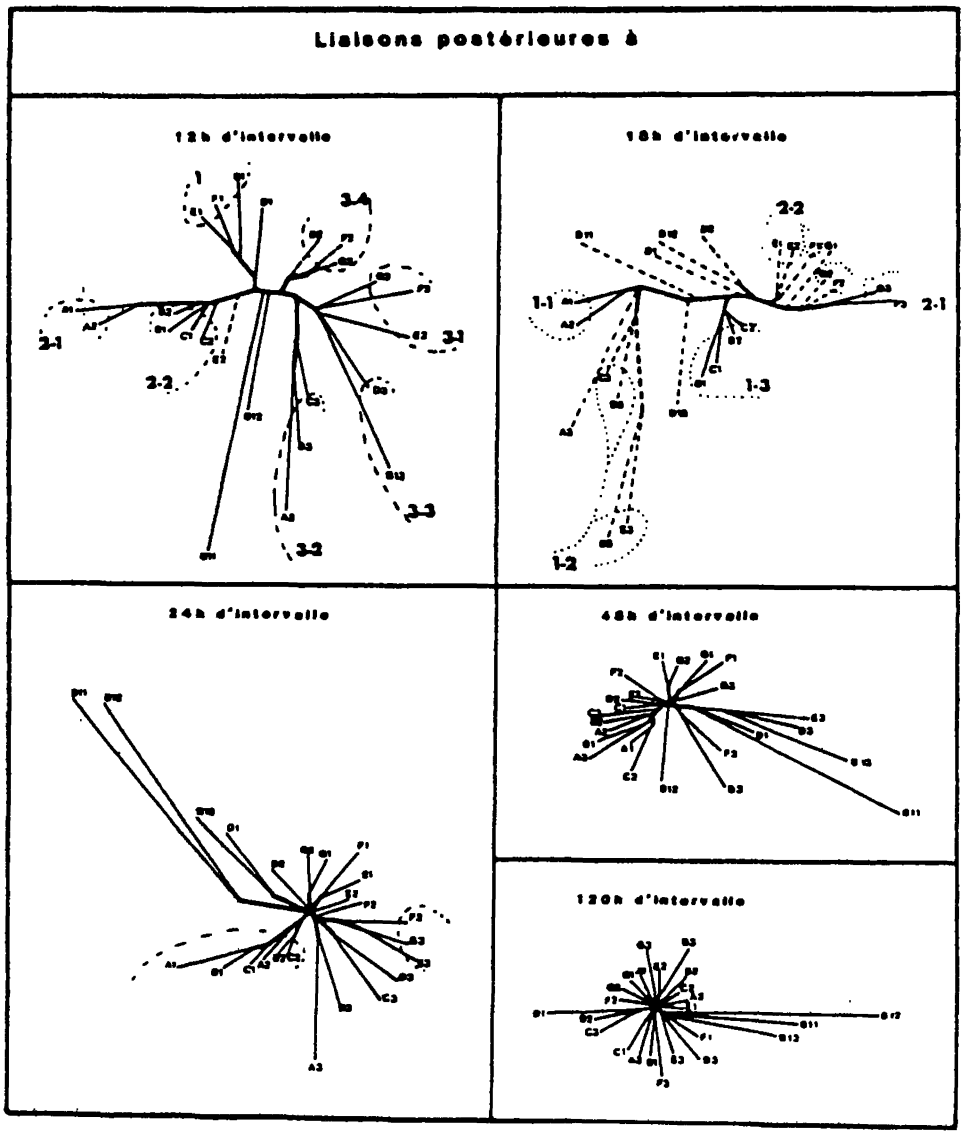


figure 8 : liaisons postérieures à intervalle de 12, 18, 24, 48 et 120 heures .

faibles. A 18ʜle mauvais temps n'est plus séparé en deux types: il s'oppose directement maintenant au beau temps; les indices d'agrégation sont souvent faibles (les arcs en pointillé indique des indices inférieurs à 0,8) . A partir des liaisons de plus de 24 heures, l'arbre prend la forme de plus en plus nette d'une étoile, les ACI tendant à s'agréger ensemble au coeur de l'arbre.

4-3 Note.

Les logiciels correspondant à cette partie ont été réalisés par l'un des auteurs de cet article (Luong) ainsi que par H. Abdi, A. Guénoche. Ces logiciels sont disponibles sur micro-ordinateurs, la version de Luong est écrite en Pascal (turbo), celles de Abdi, Guénoche, Luong, en Basic.

BIBLIOGRAPHIE

(1)H. ABDI , 1985, "Représentations arborées de l'information verbatim", Bulletin de Psychologie; 1; 11.
(2) H.ABDI, J.P. BARTHELEMY, LUONG X.,1984, "Tree representations of associative structures in semantic and episodic memory research" in Trends in Mathematical Psychology,(E. Degreef & J. Van Buggenhaut eds.), Elsevier Science Publishers, North-Holland; 3-31.
(3) G. BROSSIER, 1985, "Approximation des dissimilarités par les arbres additifs", Math. et Sciences Humaines; 91; 5-21.
(4) P. BUNEMAN , 1971, "The recovery of trees from measures of dissimilarity" in Mathematics in Archeological and Historical Sciences (Hobson, Kendal & Tautu eds.), Edinburgh University Press; 387-395.
(5) P.BUNEMAN., 1974, " A note on metric properties of trees", Journ. Comb. Theory (B); 17; 48-50.
(6)J.D. CARROLL., 1976, "Spatial, non-spatial and hybrid models for scaling", Psychometrika; 41, 4; 439-463.

(7) J.D. CARROLL, L.A. CLARK, W.S. DESARBO, 1984, "The representation of three way proximity data by single and multiple structure models", Journ. of Classification; 1; 25-74.
(8) J.D. CARROLL, S. PRUZANSKY, 1980, "Discrete and hybrid scaling models" in Similarity and Choice (Lanterman, Freger eds.); Hans Huber (Bern).
(9) L.L. CAVALLI-SFORZA, A.W.F. EDWARDS, 1967, "Phylogenetic analysis models and estimation procedures", Ame. Journ. of Human Genetic; 19; 233-257.
(10) S. CHAIKEN, A.K. DEWDNEY, A.W.F. SLATER, 1983, " An optimal diagonal tree code", SIAM J. Alg. Disc. Math.; 4,1; 424-429.
(11) H. COLONIUS, H.H. SCHULZE, 1979, "Representation nichnumerischer Anlichkeit durch Baumstrukturen", Psychologische Beitrage; 21; 98-111.
(12) H.COLONIUS, H.H. SCHULZE, 1981, "Tree structure for proximity data", British Journ. of Math. and Stat. Psychology; 17; 167-180.
(13) J.P. CUNNINGHAM, 1978, "Free trees as representation of psychological distances", Journ. of Math. Psychology; 17; 165-188.
(14) J.P. CUNNINGHAM, 1980, "Trees as memory representations for simple visual patterns", Memory and Cognition; 8; 593-605.
(15) W.H.E. DAY, 1986, "Analysis of quartet dissimilarity measures between undirected phylogenetic trees", prēprint, Uni. de Montréal.
(16) G.DE SOETE, 1983-a, "A least squares algorithm for fitting additive trees to proxility data", Psychometrika; 48; 621-626.
(17) G. DE SOETE, 1983-b, "Are nonmetric additive tree representations meaningful ?", Quality and Quantity; 17; 475-478.
(18) G. DE SOETE, W.S. DESARBO, G.W. FURNAS, J.D. CARROLL, 1984, "The estimation of ultrametric and path length trees from rectangular proxility data", Psychometrika; 49, 3 ; 289-310.
(19) A.K. DEWDNEY, 1979, "Diagonal tree codes", Information and Control; 40; 234-239
(20)A.J. DOBSON, 1974, "Unrooted trees for numerical taxonomy", Journ. of Applied Proba.; 11; 32-42

(21) G.F. ESTABROOK, F.R. MCMORRIS, C. MEACHAM, 1985,"Comparison of of undirected phylogenetic trees based on subtrees of four evolutionary units", Systematic Zoology; 34, 2; 193-200.

(22) M. FRIENDLY, 1979,"Methods for finding graphic representation of associative memory structures" in RUFF C.F. (eds) "Memory organisation and structure", New-York, Academic Press.

(23) F. GALTON, 1879, "Composite portraits, made by combining those of many different persons ...", Journal of the Anthropol. Inst.: 8; 132-144

(24) S.L. HAKIMI, S.S. YAU, 1964, "Distance matrix of a graph and its realizability", Quaterly of App. Math.; 22; 305-317.

(25) D. JOLY, 1980,"Essai de modélisation des variations thermiques,..", Thèse du 3ème cycle, EPHESS, Paris 2 tomes, 526 et 163 pages.

(26) D. JOLY, X. LUONG, 1984, "Etude de la succesion des types de temps fondé sur un nouvel algorithme de représentation arborée", IGU Working Group on Systems Analysis and Mathematical Models. Besançon Symposium.

(27)G. LE CALVE, 1986, "Distance à centre", Statistique et Analyse des Données.

(28) X. LUONG, M. NOVI, 1986, "Représentation arborée des données textuelles", in Méth. Quant. et Informatiques... Champion-Slatkine.

(29)F.R. MCMORRIS, 1985, "Axioms for consensus functions on undirected phylogenetic trees", Math. Biosciences; 74; 17-21

(30) A.N. PATRINOS, S.L. HAKIMI, 1972, "The distance matrix of a graph and its tree realisation", Quaterly of App. Math.; 30; 255-269.

(31) D.F. ROBINSON, L.R. FOULDS, 1981, "Comparison of Phylogenetic Trees", Math. Biosciences; 53; 131-147.

(32) M. ROUX, 1968, "Un algorithme pour construire une hiérarchie particulière". Thèse de 3ème cycle, Paris.

(33) M. ROUX, 1985, " Représentation d'une distance par un arbre aux arêtes additives", Colloque de l'INRIA, Versailles, à paraître dans Data Analysis and Informatics IV , North-Holland.

(34)S. SATTAH, A. TVERSKY, 1977, "Additive Similarity Trees", Psychometrika; 42, 3; 319-345.

(35) J.M.S. SIMOES-PEREIRA, 1967, "A note of tree realizability of a distance matrix", Journ. Comb. Theory (B); 6; 303-310.

(36) R.R. SOKAL, C.D. MICHENER, 1958, "A statistical method for evaluation systematic relation ships", University of Kansas Scie. Bulletin; 38; 1409-1438.
(37)Y.A. SMOLENSKII, 1969, "A method for linear recording of graphs", USSR Comput. Math. and Math. Phys.; 2; 396-397.
(38) A. TVERSKY, 1977, " Feature of similarity", Psychological Review; 84; 327-352.
(39)M.S. WATERMAN, T.F. SMITH, M. SINGH, W.A. BEYER, 1977, "Additive evolutionnary trees", Journal of Theorical Biology; 64; 199-213.
(40) L. WITTGENSTEIN, 1953, "Philosophische Untersuchunnen", Oxford; Blackwell.
(41) K. ZARETSKII, 1965, "Construction a tree on the basis of a set of distances between the hanging certices", Upekki Math. Nauk.; 20; 90-92 (en Russe).