

Université de Liège
Laboratoire de Pédagogie Expérimentale

LES REPRÉSENTATIONS EN ARBRES : PRÉSENTATION, CRITIQUES ET SUGGESTIONS

[TREE REPRESENTATIONS. A CRITICAL REVIEW AND SUGGESTIONS]

DANIEL DEFAYS

This review is devoted to the analysis of tree representation; hierarchical clustering and linear seriation are presented as special cases. For each type of tree, examples, properties and algorithms are given or evoked. Finally, representation of ordinal data is discussed. New suggestions are made.

La représentation en arbre est un mode de structuration des données de plus en plus fréquemment rencontré en psychologie et en sciences de l'éducation. Il n'est pas inutile à ce stade de développement de ce type d'analyse des données de répertorier de manière critique les principaux apports des dernières années. Notre intention est de présenter et d'illustrer différents modes de représentation en arbre; certains modèles évoqués sont originaux.

Le premier paragraphe est consacré à l'étude d'un arbre utilisé en classification hiérarchique: le dendrogramme; dans le deuxième, divers types d'arbres, pour la plupart déjà connus, sont passés en revue. La sériation linéaire, cas particulier d'une représentation en arbre, est évoquée dans le troisième paragraphe. Nous terminons en présentant des méthodes particulièrement adaptées à des données peu structurées (non quantitatives).

LE DENDROGRAMME OU L'ARBRE ULTRAMÉTRIQUE

Les techniques de classification hiérarchique offrent souvent une alternative aux représentations géométriques classiques. Un mode de présentation très utilisé et très parlant d'une classification hiérarchique est le dendrogramme. Présentons d'emblée un exemple: le dendrogramme associé à la classification hiérarchique définie par la distance ultramétrique du Tab. 1 est donnée par la Fig. 1.

Comme le montre la Fig. 1, un dendrogramme est un graphe, c'est-à-dire un ensemble de sommets (représentés ici par des points) reliés par des arcs. Mais ce graphe n'est pas quelconque! En effet, on peut toujours, sur ce graphe, passer d'un sommet à un autre en suivant des arcs (éventuellement à contre sens), propriété appelée connexité; le nombre d'arcs est toujours égal au nombre de sommets moins un. Les graphes qui possèdent ces deux propriétés sont appelés des

	x_1	x_2	x_3	x_4
x_1	0	1	2	3
x_2		0	2	3
x_3			0	3
x_4				0

TAB. 1. UNE DISTANCE ULTRAMÉTRIQUE — AN ULTRAMETRIC DISTANCE



FIG. 1. CLASSIFICATION HIÉRARCHIQUE DES DONNÉES PRÉSENTÉES DANS LE TAB. 1. — HIERARCHICAL CLUSTERING OF DATA GIVEN IN TAB. 1

arbres. Ainsi, les deux graphes repris dans la Fig. 2 ne sont pas des arbres; le premier parce qu'il n'est pas connexe, le second parce qu'il a trop d'arcs par rapport aux sommets.

Qualifier un dendrogramme d'arbre n'est pas suffisant! L'arbre repris dans la Fig. 1 n'est pas quelconque. Il possède, entre autres, un sommet privilégié: du point situé au sommet du graphe on peut

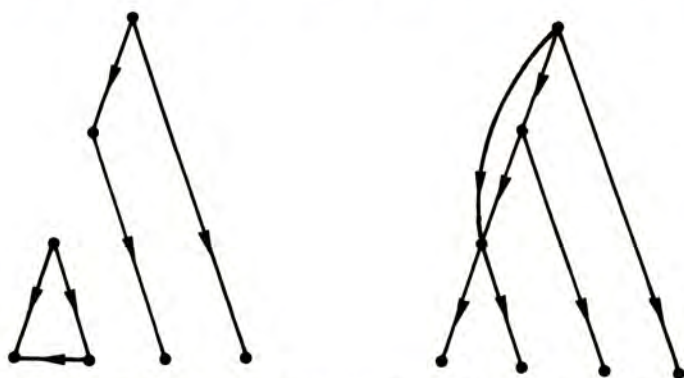


FIG. 2. DES GRAPHES QUI NE SONT PAS DES ARBRES — GRAPHS WHICH ARE NOT TREES

atteindre tous les autres sommets en empruntant des arcs et en respectant leur orientation. Ce point est appelé la racine de l'arbre. Les sommets dont ne part aucun arc sont dits terminaux. Un arbre qui possède une racine est une arborescence. Le dendrogramme est une arborescence dont tous les sommets terminaux, et uniquement eux, représentent les éléments de l'ensemble de départ. Une dernière caractérisation est encore nécessaire : l'arbre doit être fourchu ; en d'autres mots, de tout sommet non terminal doit partir au moins deux arcs. Une telle arborescence est quelquefois appelée un arbre ultramétrique.

Dans un dendrogramme, les sommets non terminaux représentent des classes d'éléments. Comme l'indique la Fig. 3, la classe associée à chaque sommet est l'ensemble des sommets terminaux auxquels on peut accéder à partir de ce sommet.

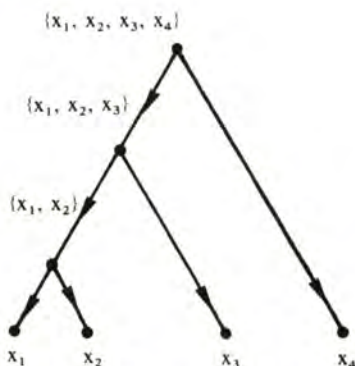


FIG. 3. LES DIFFÉRENTES CLASSES DE LA CLASSIFICATION HIÉRARCHIQUE DONNÉE DANS LA FIG. 1 — THE DIFFERENT CLASSES OF THE HIERARCHICAL CLUSTERING GIVEN IN FIG. 1

Il est facile de montrer que dans un arbre et donc, a fortiori, dans un arbre ultramétrique, il est toujours possible de se déplacer d'un sommet à un autre en empruntant des arcs (indépendamment de leur orientation) et ce d'une seule manière si l'on convient de ne pas emprunter deux fois le même arc. Ce parcours est appelé une chaîne. Ainsi la chaîne (elle est unique) liant x_1 à x_3 , dans le graphe de la Fig. 3, est constitué des arcs

$$(\{x_1, x_2\}, x_1), (\{x_1, x_2, x_3\}, \{x_1, x_2\}), (\{x_1, x_2, x_3\}, x_3)$$

Si on appelle longueur d'une chaîne, la somme des longueurs des arcs qui la constituent, on peut montrer qu'il est possible de calibrer la longueur des arcs d'un dendrogramme de manière à ce que pour toute paire de sommets terminaux, la longueur de la chaîne qui les joint soit égale à la distance qui les sépare. Reprenons notre exemple initial, en indiquant à côté de chaque arc, sa longueur. Le dendrogramme obtenu est représenté dans la Fig. 4. Dans un dendrogramme, la distance d'une classe à tous ses éléments est la même.

On a donc représenté une matrice de distances par un arbre ultramétrique valué (un graphe est dit valué lorsqu'une longueur ou un poids est associé à chacun de ses arcs) de telle manière que la distance

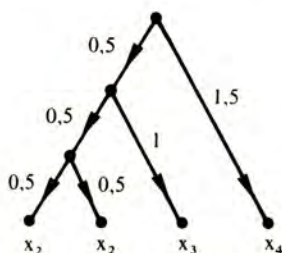


FIG. 4. UN ARBRE VALUÉ — A WEIGHTED TREE

entre deux éléments soit toujours égale à la longueur de la chaîne qui les joint.

Ce mode de structuration est particulièrement intéressant. Il permet, en effet, de regrouper les éléments en classes hiérarchisées et de visualiser sur un plan la hiérarchie de ces classes, tout en conservant toute l'information métrique contenue dans la matrice de distances. Ainsi, par exemple, la classification hiérarchique (obtenue par la complete link method) des huit premiers des 24 tests administrés par Holzinger et Swineford à 145 sujets de Chicago est représentée

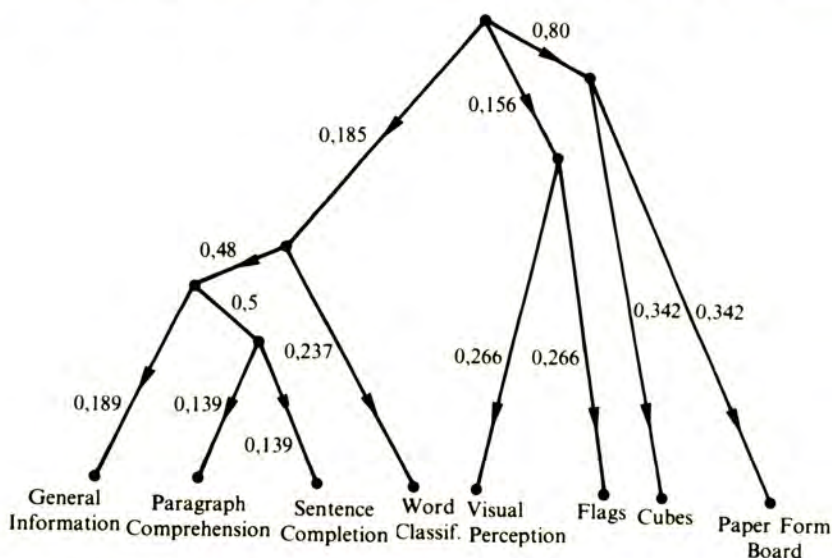


FIG. 5. CLASSIFICATION HIÉRARCHIQUE DE HUIT TESTS PSYCHOLOGIQUES — HIERARCHICAL CLUSTERING OF EIGHT PSYCHOLOGICAL TESTS

au Tab. 5. (B. Baker et L. J. Hubert, 1976). Malheureusement, comme on peut s'y attendre, ce type de structuration n'est pas souvent possible. Pour que la représentation soit adéquate, la distance doit satisfaire à l'inégalité ultramétrique :

$$\forall x, y, z, \quad d(x, y) \leq \max \{d(x, z), d(z, y)\}$$

Cette inégalité est très restrictive puisqu'elle impose, entre autres, que tous les triangles soient isocèles.

La matrice des dissimilarités à partir de laquelle la classification hiérarchique des 8 tests psychologiques a été obtenue, ne satisfait pas parfaitement à l'inégalité ultramétrique. Les algorithmes de classification hiérarchique sont, du reste, destinés à faire violence aux dissimilarités non ultramétriques pour les rendre ultramétriques. Encore faut-il que la distortion ainsi introduite ne soit pas trop importante. Une importante littérature est consacrée à ces différents algorithmes. Le lecteur intéressé trouvera dans Defays (1978b) une revue de la question.

LES REPRÉSENTATIONS EN ARBRES

Lorsqu'une classification hiérarchique n'est pas possible ou pas désirée, les statisticiens ont imaginé des structurations de données plus générales et moins contraignantes. Plutôt que représenter les données par un arbre ultramétrique, on peut plus simplement peut-être les représenter par un arbre. Différents modes de représentation que nous allons discuter maintenant sont, a priori, possibles.

L'alternative à l'arbre ultramétrique, la plus souvent rencontrée, est la suivante : représenter les données par un arbre valué dont certains sommets sont associés aux éléments de l'ensemble à structurer de manière à ce que, pour toute paire de sommets associés à ces éléments, la longueur de la chaîne qui les joint soit égale à la distance qui sépare les éléments. Ce mode de représentation est visiblement plus général que le précédent : on n'impose en effet ni l'existence d'une racine, ni les seuls sommets terminaux comme représentatifs des éléments de l'ensemble à structurer, ni le caractère fourchu de l'arbre. De plus ici, aucune orientation des arcs n'est, a priori, requise.

Différents auteurs (Buneman, 1974; Dobson, 1974, ...) ont montré qu'une représentation de ce type était possible, si et seulement si, la distance satisfaisait à la condition suivante, appelée condition des quatre points :

$$\forall x, y, u, v, \quad d(x, y) + d(u, v) \leq \max \{d(x, u) + d(y, v), d(x, v) + d(y, u)\}$$

Il est facile de montrer que l'inégalité ultramétrique implique la condition des quatre points qu'elle-même implique l'inégalité triangulaire. En fait, la condition des quatre points peut s'exprimer sous une forme équivalente mais plus parlante :

$$\forall x, y, u, v,$$

les deux plus grands des trois nombres

$$(d(x, y) + d(u, v)), (d(x, u) + d(y, v)), (d(x, v) + d(y, u))$$

sont égaux.

Il est facile de voir que cette condition est nécessaire. En effet, si a, b, c, d , sont quatre sommets de l'arbre, il y a toujours moyen de les permuter de manière à ce qu'ils s'organisent comme dans la figure

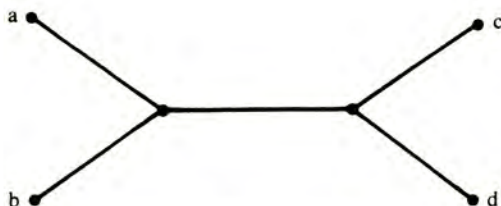


FIG. 6. LA CONFIGURATION LA PLUS GÉNÉRALE DE QUATRE OBJETS DANS UN ARBRE — MOST GENERAL CONFIGURATION OF FOUR OBJECTS IN A TREE

6 (des arêtes ont éventuellement une longueur nulle). Quelle que soit la longueur des arêtes, on aura toujours :

$$d(a, b) + d(c, d) \leq d(a, c) + d(b, d) = d(a, d) + d(b, c)$$

Le caractère suffisant de la condition a été démontré entre autres par Buneman (1974) et Dobson (1974). Donnons un exemple. A la matrice des distances, voir Tab. 2, on associe l'arbre valué, voir Fig. 7.

	x_1	x_2	x_3	x_4	x_5
x_1	0	1	2	3	2
x_2		0	2	3	2
x_3			0	3	2
x_4				0	1
x_5					0

TAB. 2. UNE DISTANCE ADMETTANT UNE REPRÉSENTATION EN ARBRE — A TREE METRIC

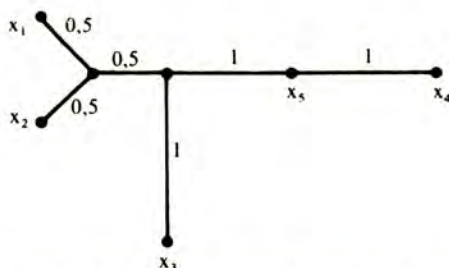


FIG. 7. REPRÉSENTATION EN ARBRE DES DONNÉES PRÉSENTÉES DANS LE TAB. 2 — TREE REALIZATION OF DATA GIVEN IN TAB. 2

Ce type de représentation des données est intéressant à maints égards. Il permet une représentation planaire parlante des distances entre éléments; il permet une mise en évidence de sommets intermédiaires non représentatifs d'éléments étudiés dont l'interprétation peut être intéressante; il permet, enfin, via la définition d'une racine de l'arbre et une orientation des arêtes, une interprétation en termes de caractéristiques communes ou différentes.

Si les deux premières affirmations sont évidentes, la troisième mérite une explication. Sattah et Tversky (1977) ont montré que si un graphe est une arborescence (ce qui nécessite dans notre cas le choix d'un sommet comme racine (on prend souvent le sommet le plus équidistant des sommets terminaux) et l'orientation des arêtes en conséquence), chaque arc peut s'interpréter comme l'ensemble des caractéristiques communes à tous les sommets subséquents (c'est-à-dire que l'on peut atteindre de l'arc en question, en respectant l'orientation). La longueur de l'arc donne une mesure de l'importance de cet ensemble de caractéristiques. L'exemple qui suit, extrait d'une analyse menée par Sattah et Tversky (1977), permettra de bien comprendre la fécondité de ce type d'approche.

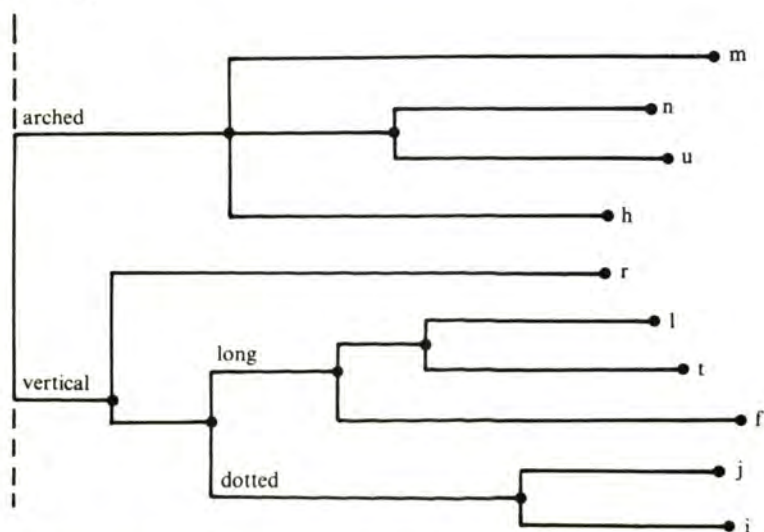


FIG. 8. REPRÉSENTATION EN ARBRE DE SIMILARITÉS ENTRE LETTRES — TREE REPRESENTATION OF LETTER SIMILARITY

On a demandé à 57 sujets suédois de chiffrer la similarité entre des paires de lettres minuscules. Le morceau d'arbre qui suit, obtenu à partir des données recueillies, peut s'interpréter comme sur la Fig. 8, en termes de caractéristiques communes ou différentes. La distance entre deux lettres est alors fonction de l'ensemble des caractéristiques qu'elles n'ont pas en commun.

La longueur des différents arcs peut aussi s'interpréter; le fait que l'arc «longueur» soit plus petit que l'arc «avec un point» signifie que pour les sujets interrogés la caractéristique «longueur» a moins de poids que la caractéristique «avec un point» quand il s'agit d'évaluer des similarités entre lettres.

Pour que la structure en arbre soit vraiment intéressante, il convient évidemment qu'il n'y ait qu'un seul arbre associé à une distance satisfaisant à la condition des quatre points. En effet, si plusieurs arbres différents sont associés à une même matrice de distances, aucun ne peut prétendre représenter plus qu'un autre la structure des données et l'interprétation devient fort délicate. Est-ce le cas? Si la question est posée de manière aussi directe, la réponse est malheureusement non. Les deux arbres qui suivent, par exemple, peuvent tout aussi bien être associés à la matrice de distances donnée dans le Tab. 2.

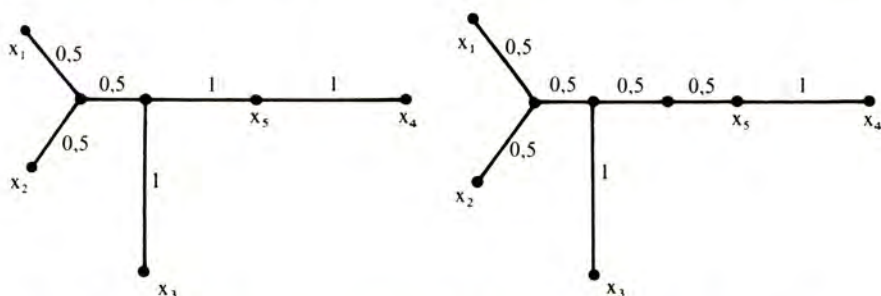


FIG. 9. DEUX ARBRES DIFFÉRENTS QUI REPRÉSENTENT LES DONNÉES DU TAB. 2 — TWO DIFFERENT TREES WHICH MAY REPRESENT DATA OF TAB. 2

Ce qui distingue ces deux arbres est la présence dans le second d'un sommet supplémentaire juste avant x_5 . Ce sommet ne joue cependant pas un rôle fondamental, puisqu'il peut être supprimé sans altérer l'adéquation de la représentation. Il est donc utile de spécifier que les sommets intermédiaires, c'est-à-dire les sommets non associés aux éléments de l'ensemble de départ, aient tous plus de deux sommets adjacents (sinon on peut les supprimer). Sous cette condition, on peut montrer que la représentation est unique.

Le problème qui se pose en pratique est d'associer un arbre valué à une dissimilarité δ qui ne satisfait pas parfaitement à la condition des quatre points. Différents algorithmes ont été proposés. Ils se décomposent généralement en deux phases: une première phase où l'arbre seul est déterminé, une seconde où la longueur des arcs est estimée. Seule la première étape est réellement problématique. Exposons brièvement une des solutions proposées. Elle est basée sur la constatation suivante: si quatre points x, y, u, v sont disposés comme on voit dans la Fig. 10 l'inégalité suivante est satisfaite:

$$d(x, y) + d(u, v) \leq d(x, u) + d(y, v) = d(x, v) + d(y, u).$$

Il suffit alors, pour tout ensemble de 4 points $\{x, y, u, v\}$, de les rebaptiser de manière à ce que

$$\delta(x, y) + \delta(u, v) \leq \delta(x, u) + \delta(y, v) \leq \delta(x, v) + \delta(y, u),$$

ce qu'il y a toujours moyen de faire. Les points x et y (resp. u et v) sont qualifiés de voisins. Tous les ensembles de quatre points sont considérés tour à tour. Les deux points les plus souvent qualifiés de



FIG. 10. CONFIGURATION DE QUATRE POINTS DANS UN ARBRE — CONFIGURATION OF FOUR POINTS IN A TREE

voisins sont joints et remplacés par un point intermédiaire z (situé entre x et y dans l'arbre) dans le reste de l'algorithme. La dissimilarité entre un point quelconque w et un nouveau point z ainsi formé à partir d'une paire $\{x, y\}$ est

$$\frac{\delta(x, w) + \delta(y, w)}{2}.$$

L'analyse est alors recommencée sur la nouvelle matrice de dissimilarités ainsi obtenue. On procède de manière itérative jusqu'à ce que le nombre d'éléments soit réduit à trois. Les trois derniers points sont alors réunis en étoile. Le centre de l'étoile est souvent pris comme racine de l'arbre lorsqu'une interprétation en termes de caractéristiques est désirée.

Dans la seconde étape, il y a lieu de déterminer la longueur des arcs de l'arbre ainsi construit de manière à ce que

$$\sum_{x, y} (d(x, y) - \delta(x, y))^2.$$

soit un minimum (d est la distance dans l'arbre, δ est la dissimilarité initiale). Ce problème se ramène à un problème classique des moindres carrés qui peut se résoudre par exemple au moyen des techniques classiques de l'algèbre linéaire. Si des longueurs d'arcs sont négatives, on les prend égales à 0. Cet algorithme, baptisé **ADDTREE** a été mis au point par Sattah et Tversky (1977). Il permet d'associer à une dissimilarité un arbre valué. Encore faut-il que la dissimilarité soit mesurée sur une échelle d'intervalle ou de rapport pour que la procédure ait un sens. En effet, elle nécessite de sommer des dissimilarités, opération qui n'est a priori possible que pour des données quantitatives. Sensibles à cette limitation, les auteurs proposent une

variante pour données ordinales. Remarquons toutefois que si les dissimilarités ne sont définies qu'à une transformation monotone près, on peut, sans leur faire violence, leur associer des arbres valués très différents.

La Fig. 11 donne deux arbres valués différents pour lesquels l'ordre des distances $d(x, y)$, $d(y, z)$, $d(x, z)$ est le même.

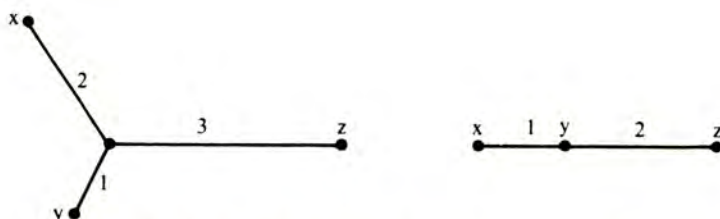


FIG. 11. DEUX REPRÉSENTATIONS DIFFÉRENTES DE DISSIMILARITÉS DÉFINIES À UNE TRANSFORMATION MONOTONE PRÈS — TWO DIFFERENT TREE REALIZATIONS OF THE SAME ORDINAL DISSIMILARITIES

Le type de représentation en arbre (valué ou non) qui nous avons examiné jusqu'ici nous semble donc inapproprié à l'étude des données ordinales. Par contre, si on impose à l'arbre représentatif des données de ne pas avoir plus de sommets que d'éléments à représenter, c'est-à-dire si les sommets intermédiaires ne sont plus tolérés, cette objection n'est plus possible : la représentation, si elle est possible, est unique (seules les longueurs d'arêtes sont indéterminées à une transformation monotone près)¹. Examinons plus en détail ce nouveau mode de présentation des données. Il est manifestement plus restrictif, moins général que le précédent. Il permet la description des données sans recours à des éléments hypothétiques non observés. La condition requise pour qu'une représentation par un arbre valué, avec correspondance biunivoque entre sommets de l'arbre et éléments à représenter, soit possible est plus sévère que la condition des quatre points.

En notant $x/z/y$ chaque fois que $d(x, z) + d(z, y) = d(x, y)$, on peut montrer que les 2 conditions suivantes sont nécessaires et suffisantes :

- pour tous x, y , soit il existe z tel que $x/z/y$, soit, pour tout u , on a $x/y/u$ ou $u/x/y$
- pour tous x, y, u, v tels que $x/y/u$ et $y/u/v$, on a $x/u/v$.

Elles impliquent entre autres la condition des quatre points. Donnons un exemple de ce type de représentation. La matrice des distances est donnée dans le Tab. 3.

¹ Ceci peut se démontrer par récurrence sur le nombre de sommets. La démonstration un peu longue est admise.

	x_1	x_2	x_3	x_4	x_5
x_1	0	1	2	1	0,5
x_2		0	2	1	0,5
x_3			0	1	1,5
x_4				0	0,5
x_5					0

TAB. 3. UNE DISTANCE ADMETTANT UNE REPRÉSENTATION EN ARBRE PARTICULIÈRE — A PARTICULAR TREE METRIC

L'arbre qui lui est associé est donné dans la Fig. 12.

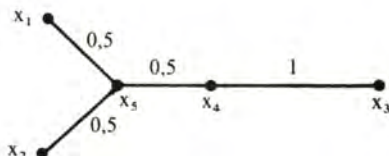


FIG. 12. REPRÉSENTATION EN ARBRE DES DONNÉES PRÉSENTÉES DANS LE TAB. 3 — TREE REALIZATION OF DATA GIVEN IN TAB. 3

Si la métrique ne satisfait pas parfaitement aux égalités requises, des approximations sont possibles. L'unicité de ce type de représentation est un corollaire immédiat d'un théorème d'unicité que nous avons mentionné dans le cas précédent.

Si la matrice des dissimilarités n'est plus symétrique, une autre variante de représentation arborescente a été proposée par Cunningham (1978). Remarquons que ce cas n'est pas aussi rare que l'on voudrait bien le croire. Tversky (1977) a montré que si un élément avait plus qu'un autre un statut de référent, leur dissimilarité n'était pas symétrique : ainsi, par exemple, la Corée du Nord est jugée, par un ensemble de sujets, plus semblable à la Chine Rouge que la Chine Rouge à la Corée du Nord. La généralisation proposée par Cunningham permet de tenir compte de cette dissymétrie. Dans la représentation qu'il propose, les paires de sommets qui dans les arbres précédents étaient joints par une arête sont maintenant joints par deux arcs de sens opposés et de longueurs éventuellement différentes. Un arbre bidirectionnel valué est donc un arbre valué où les arêtes ont été dédoublées. La longueur des arcs est ajustée de manière à ce que pour tout couple de sommets, la longueur du chemin qui les joint (en respectant cette fois l'orientation des arcs) soit égale à leur dissimilarité. Les sommets intermédiaires sont tolérés. Pour que cette représentation soit possible, les conditions suivantes sont requises :

- $\forall x, y, z, d(x, y) + d(y, z) + d(z, x) = d(y, x) + d(x, z) + d(z, y)$;
- si $d'(x, y) = d(x, y) + d(y, x)$, alors d' doit satisfaire à la condition des quatre points.

A la matrice de dissimilarités, non symétrique, donnée dans le Tab. 4.

	x_1	x_2	x_3	x_4	x_5
x_1	0	1	2	3	2
x_2	1,5	0	2,5	3,5	2,5
x_3	2	2	0	2,5	1,5
x_4	4	4	3,5	0	1
x_5	3	3	2,5	1	0

TAB. 4. UN TABLEAU DE DISSIMILARITÉS DISSYMETRIQUE — DISSYMETRIC DISSIMILARITIES

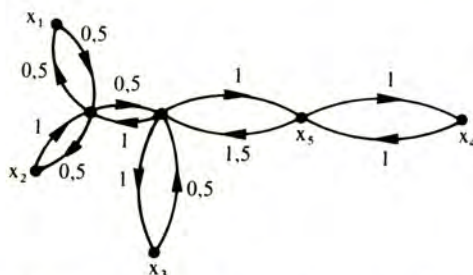


FIG. 13. REPRÉSENTATION EN ARBRE DES DONNÉES PRÉSENTÉES DANS LE TAB. 4 — TREE REALIZATION OF DATA GIVEN IN TAB. 4

il convient d'associer l'arbre bidirectionnel de la Fig. 13. Ici encore, on peut, lorsque les conditions de représentation ne sont pas parfaitement satisfaites, associer au moyen d'algorithmes appropriés, un arbre bidirectionnel à une dissimilarité.

LA SÉRIATION LINÉAIRE

Un dernier type de représentation en arbre, largement utilisée, est à mentionner : la représentation par un arbre linéaire. Un arbre linéaire est un arbre très particulier : il est composé d'un seul chemin sur lequel tous les sommets sont alignés. Ce type de représentation est plus souvent appelé *sériation linéaire*. Il impose donc de disposer les éléments à représenter sur un seul axe de manière à ce que les distances sur cet axe soient égales aux dissimilarités initiales. L'intérêt dans certains contextes d'une *sériation linéaire* n'est plus à démontrer. Elle permet d'ordonner tous les éléments sans exception. Ce mode de représentation simple et parlant est largement utilisé (Hodson et al. (1971), Hubert (1974)...). Il s'agit manifestement d'un cas particulier des représentations en arbres rencontrées jusqu'ici. Différents auteurs ont déterminé les conditions nécessaires et suffisantes pour que cela soit possible (Krantz et al. (1971), Orth (1975)...):

- $\forall x, y, z$ on a soit $x/y/z$, soit $x/z/y$, soit $y/x/z$;
- $\forall x, y, u, v$ tels que $x/y/u$ et $y/u/v$, on a $x/u/v$.

	x_1	x_2	x_3	x_4
x_1	0	1	2	4
x_2		0	1	3
x_3			0	2
x_4				0

TAB. 5. UNE DISTANCE COMPATIBLE AVEC UNE SÉRIATION DES ÉLÉMENTS — LINEAR TREE METRIC

La dissimilarité donnée dans le tableau ci-dessus satisfait à ces deux conditions. Une sériation linéaire est donc possible. Elle est donnée dans la Fig. 14.

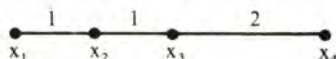


FIG. 14. REPRÉSENTATION EN ARBRE DES DONNÉES PRÉSENTÉES DANS LE TAB. 5 — TREE REALIZATION OF DATA GIVEN IN TAB. 5

Lorsque les données ne satisfont pas parfaitement aux conditions précédentes, des algorithmes permettent d'obtenir malgré tout une sériation des éléments. Nous avons, par exemple, présenté une méthode permettant à partir d'une matrice de dissimilarités donnée d'obtenir la meilleure sériation possible au sens des moindres carrés. En clair, si δ est la dissimilarité initiale et d la distance à déterminer, la méthode permet d'obtenir le d qui minimise

$$\sum_{(x,y)} (d(x,y) - \delta(x,y))^2$$

Le principe de la méthode est simple. Comme dans l'algorithme de Sattah et Tversky que nous avons exposé, il s'agit de déterminer l'arbre linéaire dans un premier temps, c'est-à-dire l'ordre de rangement des éléments à représenter sur l'axe, puis dans un deuxième temps, la longueur des arêtes, c'est-à-dire les espacements optimaux qu'il convient de ménager entre les éléments. L'ordre de rangement optimal ($x_1, x_2, x_3, \dots, x_n$) est celui pour lequel

$$\sum_{i=1}^n \left(\sum_{j>i} \delta(x_i, x_j) - \sum_{j<i} \delta(x_i, x_j) \right)^2$$

est maximum. Il s'obtient, par exemple, au moyen de l'algorithme en arbre exposé dans Defays (1978a). D'autres méthodes ont été proposées. Nous en exposerons trois assez différentes. Hubert (1974) propose la procédure suivante. Sa simplicité la rend intéressante :

1. prendre les deux éléments les plus dissemblables et les placer aux deux extrémités du continuum;
2. prendre parmi les paires dont un élément est déjà placé sur le continuum, la paire d'éléments les plus dissemblables. L'objet

non encore placé est mis en séquence sur le continuum aussi loin que possible de l'autre objet de la paire, en s'interdisant cependant de l'intercaler entre des éléments déjà placés;

3. répéter la deuxième étape jusqu'à épuisement des éléments à ranger. Cette méthode possède plusieurs variantes que nous n'exposerons pas ici. Son défaut principal est de ne pas tenir compte de toute l'information contenue dans la matrice des données. La méthode proposée par Gelfand (Hodson et al., 1971) ne présente pas cet inconvénient. Comme précédemment, les calculs peuvent s'effectuer à la main lorsque le nombre d'éléments à classer n'est pas trop grand. La méthode consiste essentiellement à construire autant de rangements qu'il y a d'éléments à classer puis à prendre le rangement moyen. Plus précisément voici comment procéder; pour chaque élément x_i construire un ordre de la manière suivante: considérer les éléments tour à tour en fonction croissante de leur distance à x_i ; on met à la droite de x_i l'élément qui lui est le plus semblable et on ordonne les autres éléments comme suit: si l'ordre obtenu après l'introduction du k^e élément est $x_{i1} \ x_{i2} \ \dots \ x_{ik}$, x est placé à la gauche de x_{i1} si $d(x, x_{i1}) < d(x, x_{ik})$ et à droite de x_{ik} si $d(x, x_{ik}) < d(x, x_{i1})$; si $d(x, x_{i1}) = d(x, x_{ik})$ on examine les distances de x à x_{i2} et à x_{ik-1} .

On obtient ainsi autant d'ordres qu'il y a d'éléments à classer. Comme les différentes sériations ainsi obtenues sont définies à leur orientation près, il est nécessaire de réorienter éventuellement certains ordres (en inversant complètement la séquence) de manière à ce que les différentes sériations obtenues aient toutes le même «sens». On calcule alors le rang moyen de chaque élément et un ordre résultant à partir de ces rangs moyens.

Le caractère optimal des deux dernières méthodes que nous venons de proposer n'est pas évident. Sibson (Hodson et al., 1971) suggère de procéder de manière plus systématique. Il propose de définir une mesure d'adéquation entre un ordre et une matrice de dissimilarités et de chercher l'ordre qui maximise cette mesure. Par exemple, si $(x_1, x_2, \dots, x_n)$ est un ordre de rangement des éléments, il importe pour que l'adéquation soit bonne, que pour tous $i \leq j \leq k$, les deux inégalités suivantes soient satisfaites:

$$\delta(x_i, x_j) \leq \delta(x_i, x_k)$$

et

$$\delta(x_j, x_k) \leq \delta(x_i, x_k).$$

Le nombre total d'inégalités de ce type vérifiées est une mesure d'adéquation possible entre l'ordre et la dissimilarité δ . Il faut alors permuter les éléments de manière à maximiser cette mesure.

Remarquons que toutes les méthodes présentées permettent uniquement d'obtenir un ordre de rangement des éléments mais pas la longueur des espacements qu'il faut ménager entre ceux-ci sur le continuum. Ce deuxième problème peut se résoudre, comme dans l'algorithme ADDTREE, par des techniques d'algèbre linéaire. Lorsque la

matrice des dissimilarités entre éléments à ranger est dissymétrique et que cette dissymétrie doit être prise en compte, le problème relève alors de l'analyse hiérarchique au sens de Guttman (Defays, 1976 et 1978c).

LES REPRÉSENTATIONS EN ARBRES A PARTIR DE DONNÉES NON QUANTITATIVES

Comme nous l'avons déjà signalé, la représentation en arbre valué peut être dépourvue de sens lorsque les dissimilarités ne sont pas définies sur une échelle d'intervalle ou de rapport. Les différentes conditions requises pour que ces types de représentation soient possibles font toutes intervenir des additions, opération qui n'a pas de sens que sur certaines échelles. Ainsi, si la dissimilarité est ordinale, la seule question qui ait un sens est la suivante : peut-on transformer de manière monotone cette dissimilarité en une dissimilarité qui satisfasse aux conditions requises pour une représentation en arbre ? Nous avons signalé que, pour la représentation en arbre la plus générale (avec sommets intermédiaires), il pouvait exister plusieurs transformations monotones débouchant sur plusieurs arbres différents. Ce qui rend la tentative sinon vaine du moins fortement dépourvue de sens. Nous avons suggéré un mode de représentation plus particulier échappant à cette critique. Malheureusement, le problème de la signification de l'addition des longueurs des arêtes de l'arbre valué construit subsiste, quelle que soit la manière dont il a été obtenu à partir des données. Est-ce à dire que dans des disciplines comme la psychologie et les sciences de l'éducation pour lesquelles les notions d'addition et de concaténation sont encore très difficiles voire impossibles à définir, les représentations en arbre soient à rejeter ? Remarquons d'abord que s'il est vrai que surestimer la structure des données doit inviter à observer la plus stricte prudence dans l'interprétation des résultats, il n'en est pas moins vrai que dans un grand nombre de situations en sciences humaines, ce type d'abus est toléré et même pratiqué avec certain fruit.

Pensons simplement aux foisons de tests psychométriques débouchant sur des mesures qui peuvent difficilement se targuer d'être des échelles d'intervalle mais qui sont fréquemment traitées comme telles. Il n'est pas absurde de supposer que souvent les échelles utilisées sont des échelles d'intervalle à des erreurs de mesure près et que les méthodes utilisées sont suffisamment robustes que pour « résister » aux distortions ainsi introduites.

Cette première argumentation de type analogique tient malheureusement plus du vœu pieux que de la démonstration scientifique rigoureuse. On peut heureusement montrer qu'une représentation en arbre est possible sans nécessiter la définition d'une quelconque opération d'addition ou de concaténation. Les théorèmes que nous allons donner ne sont malheureusement pas généralisables a priori aux arbres valués ; ils nécessitent le recours à des relations d'intermédiarité entre éléments.

Si l'on restreint ses ambitions à représenter au moyen d'un arbre non plus les distances entre les différents éléments mais plutôt leurs relations d'intermédiarité, le problème de la représentation admet une solution simple. Notre intention est de représenter les données au moyen d'un arbre de manière telle que chaque fois qu'un élément est réputé situé entre deux autres (relation d'intermédiarité qui est une donnée) le sommet qui lui est associé dans l'arbre soit situé sur la chaîne joignant les deux sommets associés à ces deux autres éléments. Réciproquement, lorsque trois sommets représentatifs d'éléments à structurer sont situés sur une même chaîne, les éléments correspondants doivent satisfaire à la relation d'intermédiarité adéquate. Ce mode de représentation des données n'est donc possible que via la définition d'une relation d'intermédiarité. Elle peut en fait s'obtenir soit directement par des jugements d'intermédiarité portés par des sujets, soit indirectement à partir de données multivariées ou de dissimilarités. La question qui se pose est : quelles conditions doivent satisfaire la relation d'intermédiarité pour qu'une représentation en arbre soit possible ?

Convenons de noter $a/b/c$ chaque fois que b est situé entre a et c . Une relation ternaire est appelée une relation d'intermédiarité si elle satisfait aux deux conditions qui suivent :

- $a/b/c$ et $b/a/c$ si et seulement si $a = b$;
- $a/b/c$ si et seulement si $c/b/a$.

Considérons maintenant les trois propriétés suivantes d'une relation d'intermédiarité :

- (i) pour tous a, b, c, d , lorsqu'on a $a/b/c$ on a toujours $a/b/d$ ou $d/b/c$;
- (ii) pour tous a, b , soit il existe c tel que $a/c/b$, soit, pour tout d , $a/b/d$ ou $b/a/d$;
- (iii) pour tous a, b, c , on a soit $a/b/c$, soit $b/a/c$, soit $a/c/b$.

Nous avons établi les résultats suivants (Defays, 1979). Si les trois propriétés (i) (ii) (iii) sont satisfaites, une représentation par un arbre linéaire est possible ; si seulement (i) et (ii) sont satisfaites une représentation par un arbre est possible, mais par un arbre comportant autant de sommets (et non pas plus) qu'il y a d'éléments à représenter ; si la propriété (i) est la seule satisfaite la représentation la plus générale, par un arbre comportant au moins autant de sommets que d'éléments à représenter, est possible. Si le nombre de sommets intermédiaires (c'est-à-dire non représentatifs d'éléments) est minimum, la représentation est unique.

D'autres représentations en arbres à partir de données ordinales ont été proposées. Ainsi Colonius et Schulze (1977) ont étudié des représentations en arbre définies non plus à partir de relations sur des triplets d'éléments comme nous l'avons fait, mais à partir de relations sur des quadruplets d'éléments.

Nous terminerons par une remarque d'ordre général : le choix judicieux d'une méthode d'analyse des données est chose importante sinon fondamentale : la présentation des données crée un type de lecture

qui lui même débouche sur des hypothèses théoriques. Les représentations en arbre offrent une alternative aux représentations euclidiennes habituelles. Si ces dernières ont prouvé leur efficacité dans l'étude des structures de type produit (décomposition factorielle), les arbres semblent plus particulièrement appropriés à l'étude des structures hiérarchiques et des processus évolutifs.

RÉFÉRENCES

- BAKER, B., & HUBERT, L. A graph theoretic approach to goodness of fit in complete-link hierarchical clustering. *Journal of the American Statistical Association*, 1976, 355, 870-878.
- BENZECRI, J.P. *L'analyse des données. 1. La taxinomie*, Paris : Dunod, 1976.
- BERGE, C. *Graphes et hypergraphes*. Paris : Dunod, 1970.
- BUNEMAN, P. A note on the metric properties of trees. *Journal of Combinatorial Theory*, 1974, B17 48-50.
- COLONIUS H., & SCHULZE H.H. *Trees constructed from empirical relations*. Unpublished manuscript, University of Technology Brunswick, 1977.
- CUNNINGHAM, J.P. Free trees and bidirectional trees as representations of psychological distance. *Journal of Mathematical Psychology*, 1978, 17, 165-188.
- DEFAYS, D. Relations floues et analyse hiérarchique des questionnaires. *Mathématiques et Sciences Humaines*, 1976, 55, 45-60.
- DEFAYS, D. A short note on a method of seriation. *British Journal of Mathematical and Statistical Psychology*, 1978a, 31, 49-53.
- DEFAYS, D. Les méthodes de l'analyse hiérarchique. Un essai de classification. *Revue Belge de Statistique, d'Informatique et de R.O.*, 1978b, 19-44.
- DEFAYS, D. Analyse hiérarchique des préférences et généralisations de la transitivité. *Mathématiques et Sciences Humaines*, 1978c, 61, 5-27.
- DEFAYS, D. Tree representations of ternary relations. *Journal of Mathematical Psychology*, 1979.
- DOBSON, A. Unrooted trees for numerical taxonomy. *Journal of Applied Probability*, 1974, 11, 32-42.
- FARRIS, J.S. Methods for computing Wagner trees. *Systematic Zoology*, 1970, 19, 83-92.
- HAKIMI, S.L., & YAU, S.S. Distance matrix of a graph and its realizability. *Quarterly of Applied Mathematics*, 1964, 22, 305-317.
- HODSON, F.R., KENDALL, D.G., & TÁUTU, P. (Eds.). *Mathematics in the archaeological and historical sciences*, Edinburgh : Edinburgh University Press, 1971.
- HUBERT, L. Some applications of graph theory and related non metric techniques to problems of approximate seriation : the case of symmetric proximity measures. *British Journal of Mathematical and Statistical Psychology*, 1974, 27, 133-153.
- KRANTZ, H., LUCE, R.D., SUPPES, P., & TVERSKY, A. *Foundations of measurement. Volume 1*. New York : Academic Press, 1971.
- ORTH, B. *Necessary and sufficient conditions for representing similarity data by a one-dimensional metric space*. Paper presented at the 6th European Mathematical Psychology Group Meeting, University of Stirling, September 1975.
- SATTAH, S., & TVERSKY, A. Additive similarity trees. *Psychometrika*, 1977, 42, 319-345.
- TVERSKY, A. Features of similarity. *Psychological Review*, 1977, 84, 327-352.