RAIRO. RECHERCHE OPÉRATIONNELLE

ANDRÉ BATBEDAT

Sur les orientations d'une ultramétrique ou d'une hiérarchie

RAIRO. Recherche opérationnelle, tome 23, n° 4 (1989), p. 393-403

<http://www.numdam.org/item?id=RO_1989__23_4_393_0>

Recherche opérationnelle/Operations Research
(vol. 23, n° 4, 1989, p. 393 à 403)

SUR LES ORIENTATIONS D'UNE ULTRAMÉTRIQUE OU D'UNE HIÉRARCHIE (*)

par André BATBEDAT (1)

Résumé. — *Un algorithme de Diday permet de construire les orientations d'une ultramétrique ou d'une hiérarchie. On en déduit quelques propriétés remarquables de ces orientations et on montre qu'elles ne s'étendent pas au cadre directement plus général des prépyramides avec les dissimilarités associées.*

Mots clés : Ultramétrique; hiérarchie; orientation.

Abstract. — *A Diday's algorithm gives the orientations of an ultrametric or an hierarchy. We deduce some remarkable properties for these orientations and we show that they cannot be extended to the directly more general context of prepyramids and associated dissimilarities.*

Keywords : Ultrametric; hierarchy; orientation.

INTRODUCTION

Nous allons établir dans cet article quelques propriétés remarquables des orientations d'une ultramétrique ou d'une hiérarchie. L'unité de présentation sera réalisée par le recours systématique à un algorithme de Diday [5] appelé PROCHE. Ensuite pour situer la portée mathématique de ces propriétés, chacune d'elles sera accompagnée d'un contre-exemple dans le cadre immédiatement plus général des pyramides [6]; pour cela, nous nous référons à l'article [1] de l'auteur.

Considérons un ensemble X (ici X est fini) : une famille K de parties de X (les paliers de K) est appelée un hypergraphe sur X. Un des problèmes classiques pour K est de déterminer s'il existe une chaîne C sur X (une

(*) Reçu avril 1988.
(1) Institut de Mathématiques, Université des Sciences et Techniques du Languedoc, place Eugène-Bataillon, 34060 Montpellier Cedex, France.

Recherche opérationnelle/Operations Research, 0399-0559/89/04 393 11/$ 3.10

orientation de K) pour laquelle tout palier de K est un C-intervalle : Booth/Lueker ont proposé en 1976 un algorithme pour résoudre ce problème. Puis en 1985, Dubost/Oubina ont élaboré un autre algorithme. Enfin, PROXEL de [1] résoud aussi ce problème. Tout ceci est détaillé dans [1].

Pour le cas particulier des hiérarchies, on utilisera PROCHE.

Dans ce contexte mathématique, il est évidemment intéressant de mettre en valeur des propriétés des orientations pour des sous-classes importantes (ici, les hiérarchies).

Maintenant dans les applications, les orientations jouent un rôle clef pour la visualisation plane : un dendrogramme hiérarchique rejette les croisements parasites et, pour cela, il doit être présenté par une orientation. Ceci se prolonge aux dendrogrammes pyramidaux qui ont été introduits par Diday [6] afin d'étendre le hiérarchique sous contrainte de visualisation plane. Remarquons qu'*a priori* la notion de croisement parasite pour un dendrogramme plan se situe dans un contexte de planarité pour le graphe du diagramme de Hasse associé : il n'est pas directement évident que la solution se trouve dans les orientations (nous avons même des contre-exemples dès que l'on sort du cadre strict des dendrogrammes). Ce problème a été résolu par Batbedat/Oubina : on a démontré que, pour la visualisation plane, l'optimum est réalisé par les pyramides orientées. *Voir* [2].

Dans les applications, les hiérarchies sont souvent soumises à des contraintes extérieures qui limitent les orientations disponibles : il est donc utile de connaître des propriétés des orientations. Une contrainte classique, issue des méthodes de regroupement, provient de la volonté de rassembler deux paliers dans une hiérarchie en formation. Une autre contrainte se trouve dans les situations de consensus où l'on souhaite présenter plusieurs hiérarchies par une orientation commune. On trouvera d'autres contraintes pour les orientations des hiérarchies dans [2].

1. MISE EN PLACE

1.1. Dans tout cet article, nous avons un ensemble X avec n éléments (les singletons) $n>1$.

Une chaîne C est la suite des singletons pour un ordre total sur X. Au rang q, $1 \leqq q < n$, nous avons la section commençante C_q, de x_1 jusqu'à x_q. Ceci est prolongé en $q=0$ par la convention que ce C_q est vide.

Une dissimilarité δ est une application qui attache à chaque paire $\{x, y\}$ une valeur réelle strictement positive $\delta(x, y)$.

Pour une chaîne C, le C-demi-tableau T de δ présente les valeurs de δ par les $T(i, j)$ pour i de 1 à $(n-1)$, pour j de $(i+1)$ à n.

Exemple : Voici un C-demi-tableau pour la chaîne C: $x<y<z<t$:

	y	z	t
x	2	3	4
y	.	1	1
z	.	.	1

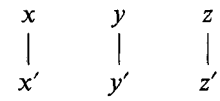
Ce demi-tableau est un peu particulier car chaque ligne est croissante (au sens large) et chaque colonne est décroissante. D'une façon générale, ceci est la définition d'un Robinson.

1.2. On rappelle qu'une dissimilarité est ultramétrique (plus simplement : ultra) si tout triplet non équilatéral est isocèle avec la base plus petite.

Exemple 1 : Voici une ultra :

	y	z	t
x	1	2	3
y	.	2	3
z	.	.	3

Exemple 2 : Dans le graphe suivant on value par 1 les arêtes et par 2 les non-arêtes : nous obtenons une ultra.



1.3. Soit δ une dissimilarité, x un singleton et Y une partie non vide de X ne contenant pas x : on dit qu'un y de Y est proche de x dans Y si pour tout $z \in Y$: $\delta(x, y) \leqq \delta(x, z)$.

Soit U une partie non vide de X, U disjointe de Y : on dit qu'un y de Y est proche de U dans Y si y est proche de chaque x de U.

Relativement à δ, on distingue deux importants types de chaînes :

(*) C est de proximité si pour tout q, $x(q+1)$ est proche de C_q dans $(X-C_q)$.

(**) C est de proche en proche si pour tout q, $x(q+1)$ est proche de x_q dans $(X-C_q)$.

LEMME (facile) : (*i*) *Toute chaîne de proximité est aussi de proche en proche (la réciproque n'est pas vraie).*

(ii) *Pour toute dissimilarité* δ *et pour tout singleton* x, *l'algorithme PROCHE que voici construit une chaîne de proche en proche commençant par* x :

pour $1 \leqq q < (n-1)$, *choisir* $x(q+1)$ *proche de* x_q *dans* $(X-C_q)$.

2. L'ALGORITHME PROXEL DE [1]

On se réfère aux résultats de [1].

Relativement à une dissimilarité δ et à une partie Y de X, on dit que $y \in Y$ est élevé dans Y si pour tous z, t dans Y :

$\delta(z, t) \leqq \text{Max}(\delta(y, z), \delta(y, t))$. On voit que cette notion d'élèvement met en valeur des singletons qui ont « un comportement ultramétrique ». D'où les chaînes de type élevé : C est élevée si pour tout q de 1 à $(n-1)$, x_q est élevé dans $(X-C(q-1))$.

Le théorème 3.6 de [1] dit qu'une chaîne C est de proximité élevée relativement à une dissimilarité δ, ssi le C-demi-tableau de δ est Robinson. Alors δ est appelée une pyra d'orientation C.

L'algorithme associé est appelé PROXEL (pour : proximité élevée) : il reconnaît si une dissimilarité est pyra et dans ce cas il construit ses orientations.

Commentaire : Nous avons maintenant deux notions d'orientation : une pour les hypergraphes (introduction) et celle-ci pour les dissimilarités. Nous verrons plus loin que ceci est tout à fait compatible.

Exemple : La dissimilarité de l'exemple 1.1 est pyra et ses orientations sont C et C^{0P} (son opposée).

3. L'ALGORITHME PROCHE DE [5] POUR LES ULTRAS

A la fin de [5], Diday explicite l'algorithme PROCHE pour une ultra et dit que cela construit une chaîne de plus courte longueur. Auparavant il avait fait le lien avec les demi-tableaux Robinson, ce qui donne le résultat suivant :

Pour une ultra, PROCHE construit une orientation.

En particulier ceci justifie un résultat connu par ailleurs : toute ultra est pyra.

Ainsi PROXEL peut aussi construire les orientations d'une ultra, mais PROCHE est plus simple et sans branche morte.

Soit α une ultra (notation reprise ensuite) : comme alternative à la démonstration de [5], il est intéressant de montrer que toute chaîne de proche en proche est de proximité élevée (autrement dit que pour α, PROCHE joue le rôle de PROXEL). Tout d'abord, pour une ultra la propriété d'élévement est toujours réalisée. D'autre part, soit C une chaîne de proche en proche, puis des singletons (notés ici par leur rang dans C).

$i<q<(q+1)<j$, dans l'hypothèse où $\delta(i, j)<\delta(i, q+1)$: alors :

$\delta(i, q)\leqq\delta(i, j)<\delta(i, q+1)=\delta(j, q+1)=\delta(q, q+1)\leqq\delta(j, q)$, ce qui est absurde.

Exemple 1 : Déroulons PROCHE sur l'exemple 1 de 1.2 : Choix — on prend y — alors x — alors z — alors t.

Exemple 2 : Déroulons PROCHE sur l'ensemble 2 de 1.2 : Choix — on prend y — alors y' — choix — on prend x' — alors x — choix — on prend z — alors z'.

4. LES DISQUES D'UN SINGLETON POUR UNE ULTRA

On se donne une ultra α et un singleton y. Pour chaque x autre que y on définit le disque X_x de y en x : ses éléments sont x et les z tels que $\alpha(x, z)<\alpha(x, y)$.

LEMME : *Dans ces conditions, les disques X_x constituent une partition de $(X-y)$.*

Preuve : Fixons un x. Pour tout z de X_x : $\alpha(y, z)=\alpha(y, x)$ (puisque la base du triplet est strictement plus petite). Ensuite pour t de X_x, autre que x ou z, $\alpha(t, z)\leqq\max(\alpha(x, t), \alpha(x, z)<\alpha(y, x)$. Par conséquent t est dans X_z.

Suivons le déroulement de PROCHE sur α, vu par le singleton y :

PR 1 : y peut rentrer au rang 1.

PR 2 : Sinon, entre un x autre que y. Alors y ne peut pas rentrer avant que le disque X_x ne soit installé en section commençante. Ainsi y peut renter au rang $(|X_x|+1)$.

PR 3 : Sinon, entre un u autre que y. Alors y ne peut pas rentrer tant que le disque X_u n'est pas entièrement rentré. Ainsi y peut rentrer au rang $(|X_x|+|X_u|+1)$.

PR 4 : etc.

• *Cas particulier :* Notons ici $m = \text{Min}|X_x|$, pour X_x parcourant les disques de y. Alors après le rang 1, le premier rang auquel y a accès dans une orientation de α est $(1+m)$.

PROPRIÉTÉ : Une pyra β est ultra ssi pour tout singleton y il existe une orientation commençant par y (... finissant par y).

Preuve : Si β est ultra, on déroule PROCHE commençant par y.

Dans l'autre sens, tout singleton est élevé dans X, donc β est ultra.

COMMENTAIRE : Cette première propriété des orientations des ultras est simple, importante et de plus caractéristique pour cette sous-classe de pyras. Elle montre que dans un dendrogramme hiérarchique les positions extrêmes des singletons, prises isolément, n'ont aucune signification particulière.

Dans les exemples 1 et 2 du 3, nous avons construit des orientations en imposant y au premier rang.

Pour la pyra non ultra de l'exemple du 1, il n'existe pas d'orientation commençant par y : ici les positions extrêmes des singletons x et t ont une signification.

5. LE PROBLÈME DES GROUPES CONNEXES

En classification hiérarchique, on construit souvent un dendrogramme par « regroupements successifs » : Jusqu'à la fin de la construction, les groupes formés restent connexes; à la fin, nous avons l'ultra du dendrogramme, présentée par une orientation (ce contexte est limpide quand on connaît la bijection de Benzécri/Johnson).

Dans [6], Diday a proposé une méthode ascendante de regroupement pyramidal qui a été étudiée par Bertrand [4] : dans ce cadre plus général les problèmes de connexité sont parfois difficiles; nous citons Bertrand : « ... pour

réaliser un tel algorithme, certains choix doivent être précisés (en particulier, comment choisir l'ordre entre les paliers agrégés ? ».

Dans ce contexte de connexité, donnons-nous une pyra β et un groupe G de singletons avec $|G|>1$, puis cherchons s'il existe une orientation C de β dans laquelle G est rassemblé selon un intervalle.

L'algorithme PROXEL s'adapte aisément : lorsqu'un singleton de G rentre dans la chaîne en formation, tous les autres doivent suivre.

Si β est une ultra, on fait de même avec PROCHE.

Exemple : Dans l'exemple 2 du 3 nous avons construit une orientation pour $G=\{y, y', x'\}$. On voit que PROCHE accepte aussi bien x, y ou z', à la place de x'.

Contre-exemple : Dans l'exemple 2 du 3, il n'existe pas d'orientation de cette ultra pour laquelle $G=\{x, y, z\}$ serait connexe.

Contre-exemple : Pour la pyra de l'exemple du 1, il n'existe pas d'orientation regroupant x et t.

PROPRIÉTÉ : soit α une ultra, x et y des singletons distincts : il existe une orientation de α regroupant x et y.

Preuve : On déroule PROCHE selon PR 2 du 4 : rentre x, puis tout le disque X_x, puis y, et PROCHE continue. Ensuite on remplace la section commençante X_x par son opposée, avec u au début : ceci est compatible avec PROCHE car $X_u=X_x$ (lemme du 4).

6. LES COUPLES EXTRÊMES DANS LES ORIENTATIONS D'UNE ULTRA

Considérons un dendrogramme hiérarchique présenté par la chaîne C : $x_1<\ldots<x_n$: alors pour l'ultra α de ce dendrogramme, la valeur sur la paire $\{x_1, x_n\}$ est maximale. On pose le problème de savoir si une telle paire extrême est soumise à d'autres contraintes (autrement dit si elle apporte d'autres informations).

PROPRIÉTÉ : Soit α une ultra de valeur maximale M, x un singleton quelconque et y tel que $\alpha(x, y) = M$: il existe une orientation de α avec x au début et y à la fin.

Preuve : Tout d'abord la propriété ultra garantit l'existence de y. Ceci étant, on déroule PROCHE selon PR 2 du 4, en faisant entrer y juste après le disque X_x. Ensuite on remarque que pour tout t hors de X_x et pour tout z dans X_x : $\alpha(z, t) = M$: on peut donc remplacer la section finisante $(X - X_x)$ par son opposée.

Contre-exemple : Dans le cas général pour une pyra, x ne peut pas être choisi arbitrairement.

Maintenant considérons la pyra de ce demi-tableau :

	x	z	y
t	1	3	4
x		2	4
z			3

Prenons x et y qui sont à distance maximale 4 : on vérifie qu'aucune des deux chaînes $x < z < t < y$ ou $x < t < z < y$, ne donne un Robinson.

7. POUR LES HIÉRARCHIES

Nous avons sollicité dans [1] une bijection BIJ entre les pyras et les prépyramides strictement indicées. Précisons que nous considérons maintenant des hypergraphes contenant tous les singletons, la partie pleine et ne possédant pas la partie vide. Une prépyramide P est un tel hypergraphe possédant une orientation au sens de l'introduction. Ensuite un indice strict s sur P est une application réelle strictement croissante, à valeurs strictement positives, définie sur les paliers non singletons.

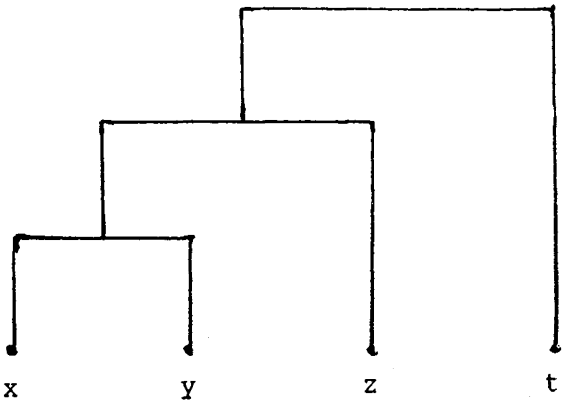
Nous utiliserons les deux propriétés suivantes de BIJ :

(*) BIJ respecte les orientations (c'est la compatibilité annoncée dans le 2).

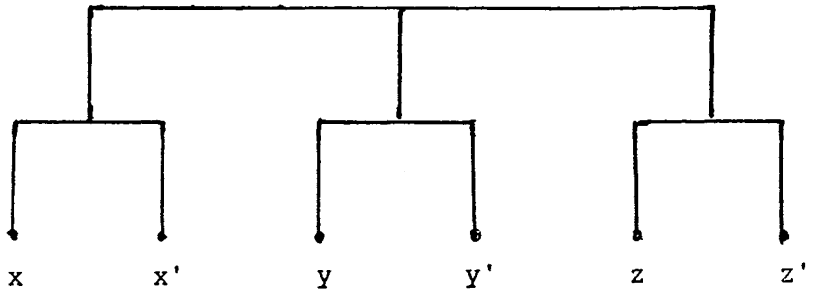
(**) La restriction de BIJ aux ultras est la bijection de Benzécri/Johnson, vers les hiérarchies strictement indicées.

Un dendrogramme hiérarchique se construit à partir d'un triple (H, s, C) : alors son ultra est BIJ(H, s) et C est une orientation de cette ultra. Dans l'autre sens, BIJ donne le dendrogramme d'une ultra orientée.

Exemple 1 : Voici le dendrogramme sur $x<y<z<t$, pour l'ultra de l'exemple 1 en 1.2 :



Exemple 2 : Voici le dendrogramme sur $x<x'<y<y'<z<z'$, pour l'ultra de l'exemple 2 en 1.2.



Étant donné une prépyramide, P, on la munit de l'indice cardinal i_c qui affecte à chaque palier le nombre de ses singletons. Cet indice est strict, donc nous avons la pyra β associée par BIJ. Ensuite, les propriétés des orientations de β passent aux orientations de P.

Ainsi on peut appliquer PROXEL pour les prépyramides et PROCHE pour les hiérarchies.

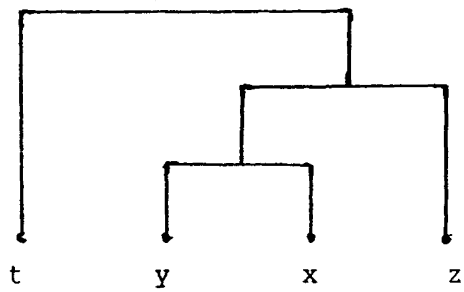
Maintenant traduisons pour les hiérarchies les propriétés des orientations des ultras : on note H une hiérarchie, x et y des singletons quelconques :

H 1 : Il existe une orientation de H qui commence par x (il existe une orientation de H qui finit par x).

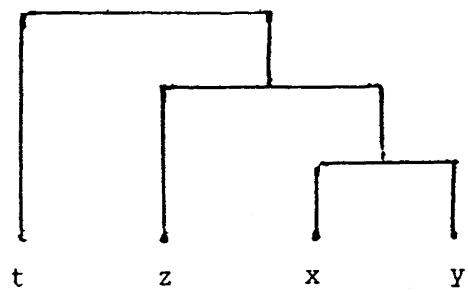
H 2 : Il existe une orientation de H dans laquelle x et y sont consécutifs.

H 3 : Il existe un singleton z pour lequel $\{x, z\}$ est contenue dans un unique palier (plein). Ensuite, il existe une orientation de H qui commence par x et qui finit par z.

Exemple 3 : Voici un dendrogramme de l'exemple 1, avec y et t consécutifs :



Exemple 4 : Voici un dendrogramme de l'exemple 1, commençant par t et finissant par y :



8. COMPLÉMENTS INFORMATIFS

Les pyramides de [6] sont les prépyramides stables par intersection non vide. Les pyras sont appelées indices pyramidaux dans [6]. Les représentations de [2] se font à trois niveaux fondamentaux : les ultras en hiérarchique, les pyras en pyramidal et les arbas en arboré.

La bijection BIJ a été étendue a toutes les dissimilarités (exposé de A. Batbedat dans le *Séminaire du Centre des Mathématiques Sociales*, le 16 mars 1987) : un résumé est disponible.

Sur les ultramétriques et les hiérarchies, Leclerc a travaillé en combinatoire et beaucoup d'auteurs (voir les livres de Jambu, Lerman, Roux, etc.) en classification.

BIBLIOGRAPHIE

1. A. Batbedat, *L'algorithme PROXEL pour les dissimilarités*, Math. Sci. Hum., vol. 102, 1988, p. 31-38.
2. A. Batbedat, Les approches pyramidales dans la classification arborée, (avec les programmes en PASCAL de J. P. Bordat), Ouvrage à paraître en 1990.
3. J. P. Benzecri, L'analyse des données, Dunod, Paris, 1973.
4. P. Bertrand, *Étude de la représentation pyramidale*, Thèse de 3e cycle, Université de Paris-Dauphine et I.N.R.I.A.-Rocquencourt, 1986.
5. E. Diday, *Croisements, ordres et ultramétriques : application à la recherche de consensus en classification automatique*, Métron, Roma, vol. XLIII-N, 1985, p. 3-20.
6. E. Diday, Une représentation visuelle des classes empiétantes : les pyramides, I.N.R.I.A.-Rocquencourt, vol. 291, 1984.
7. S. C. Johnson, *Hierarchical Clustering Schemes*, Psychometrica, vol. 22, 1967, p. 241-254.