DU DESSIN DES ARBRES RESULTANT DE CLASSIFICATIONS HIERARCHIQUES

Gérard Thauront
Institut national de recherche sur les transports et leur sécurité
2, avenue du Général Malleret-Joinville
F-94114 ARCUEIL CEDEX
tél. 33 (1) 47 40 71 09 fax 33 (1) 45 47 56 06
thauront@inrets.fr

1. Introduction

Les programmes de classification arborescente, par exemple de classifications ascendantes hiérarchiques (CAH), fournissent comme résultat une distance ultramétrique sur les objets à classer. Cette distance ultramétrique est représentable sous forme d'un arbre hiérarchique.

Les programmes de CAH offrent des représentations graphiques prédéfinies, pas toujours satisfaisantes, au point que la rédaction d'une étude " soignée " nécessite de redessiner les arbres de résultat.

Ce travail présente :

- des considérations générales sur le dessin des arbres de classification ;
- des représentations alternatives ;
- des représentations adaptées à des cas particuliers.

Certaines de ces représentations sont programmées et accessibles par FTP anonyme.

L'étude a fait l'objet d'un exposé aux XXVIII journées de statistiques de l'ASU à Québec en mai 1996.

2. Trois visions du même modèle

La structure fournie par les programmes de classification hiérarchique est très riche. Elle peut être regardée selon différents points de vue selon que l'on veut mettre en évidence :

- la forme : **structure** d'ordre sur un arbre ;
- l'aspect classificatoire : système de **classes** emboîtées ;
- la notion de ressemblance : **distance** ultramétrique ;

Nous allons d'expliquer ces trois points de vues à l'aide de la plus ancienne et la plus fameuse des classifications hiérarchiques: celle des êtres vivants.

2.1 La vision en structure arborescente

C'est, par exemple, la vision de l'évolutionniste qui s'intéresse à la façon dont une population primitive s'est diversifiée en phylums distincts. C'est bien sûr la structure d'arbre qui va présenter le mieux cette vision : Un nœud de l'arbre pourra être identifié à " l'ancêtre commun " à toutes les feuilles qui en sont issues. Les branches issues de ce nœud représentant des lignées distinctes.

L'ordre sur les branches de l'arbre est ici le temps et on place volontiers dans ce cas les feuilles vers la droite pour respecter le sens " culturel " du temps.

2.2 La vision en classes ou partitions emboîtées

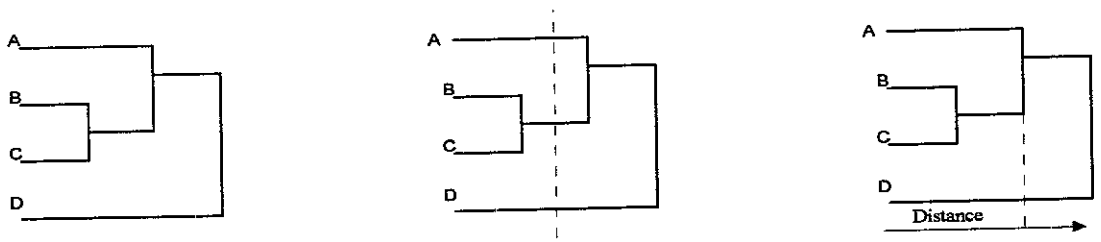
C'est typiquement la vision du taxinomiste qui, même s'il utilise la même structure (le même arbre hiérarchique) que son collègue évolutionniste, va interpréter un nœud comme l'**ensemble** des branches qui le composent.

2.3 La vision en ressemblance/dissemblance : distance

Mais cette richesse opératoire se paye par une perte de précision sur les distances initiales.

La distance ultramétrique, place tout *canis canis* à même distance d'un *canis lupus*, mais on sait bien qu'en fait le *loup gris* ressemble plus au (est plus prêt du) *berger allemand* qu'au *loulou de Poméranie*.

3. Dessin classique : vision généraliste



La structure d'**arbre hiérarchique** apparaît bien sûr sur le dessin de l'arbre.

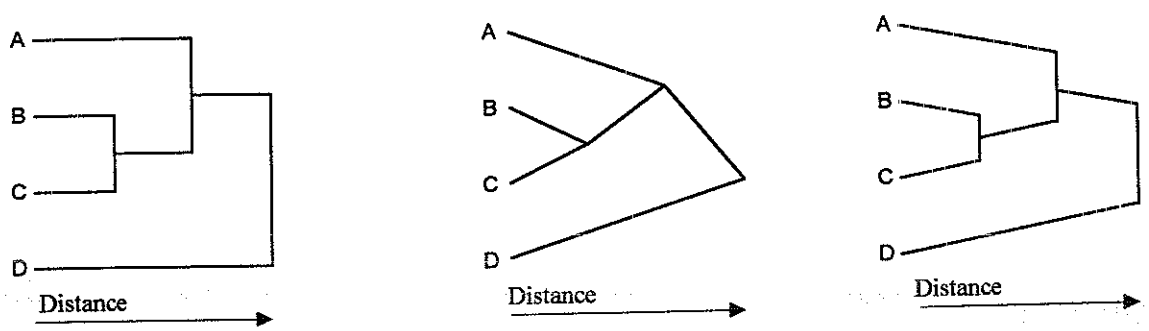
Pour faire apparaître une **partition** (en trois classes dans de dessin du milieu : {A} ; {B,C} ; {D}), il suffit de couper l'arbre (trait pointillé).

A droite, on a projeté un nœud sur l'axe supposé gradué pour lire la **distance** entre A et B (égale à la distance entre A et C).

4. Discussion sur des options du dessin de l'arbre

4.1 Carré ou pointu

On représente habituellement les arbres sous l'une de ces formes :



La forme " carrée " offre une bonne lisibilité. En effet, sur les dessins ci-dessus, seule la coordonnée horizontale est significative et représente la distance ultramétrique (ou indice d'agrégation).

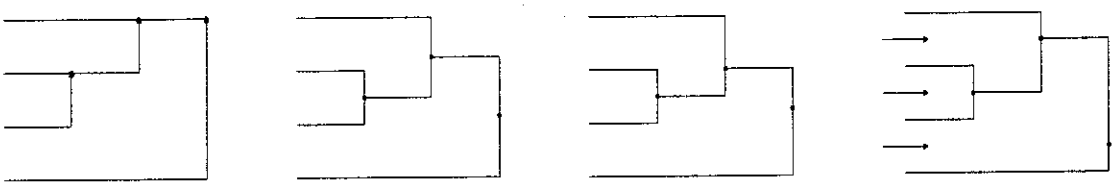
Sur l'arbre du milieu, " pointu " ou " en bec ", on a tendance à comparer la longueur des segments obliques au lieu de leur projection horizontale. Malgré tout, on peut préférer cette forme quand on s'intéresse plus à la structure de l'arbre, qu'aux valeurs de la distance ultramétrique.

La troisième forme en trapèze pourrait être choisie pour des raisons esthétiques. Pourrait-on donner une signification à l'angle des cotés ? De plus type de dessin est souvent utilisé pour représenter les " pyramides de Diday " et pourrait prêter à confusion.

Remarquons enfin que la forme carrée peut être " dessinée " sur une imprimante non graphique (fichier texte) Cf. le programme BAOBAB.

4.2 Position latérale des nœuds

Une autre discussion, moins futile qu'il n'y paraît, porte sur la position (en ordonnée) des nœuds.



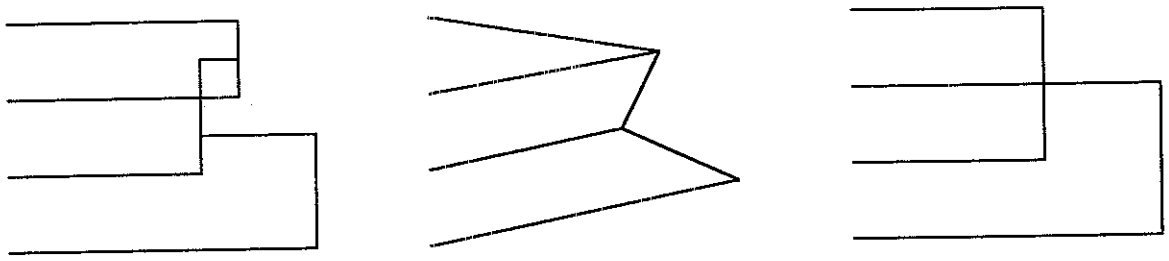
- Sur le dessin de gauche, les nœuds sont à une extrémité des segments verticaux (en haut). Cette position, trop souvent rencontrée dans les publications, ne s'explique guère que par la paresse du programmeur.
- Sur le second dessin, les nœuds sont au milieu des segments verticaux. C'est la première idée qui vient.
- Sur le troisième, ils sont au centre de gravité des sous-arbres qu'ils supportent (comme dans un mobile de Calder). C'est une fausse bonne idée !
- Sur le dessin de droite, les nœuds sont **à la bonne place,** entre les sous-arbres qu'ils rassemblent.

En effet dans un arbre binaire agrégeant **n** objets, les **n-1** nœuds s'intercalent **exactement** entre les éléments terminaux.

Par contre, si l'arbre n'est pas binaire, la position du nœud n'est plus définie.

4.3 Cas particulier des indices non monotones

Nous quittons ici le strict domaine des arbres hiérarchiques, mais c'est une situation que l'on rencontre, par exemple, le cas des CAH sous contrainte de contiguïté.



La forme carré me semble plus facile à lire que la forme en bec, chaque décroissance de l'indice étant marquée par un croisement sur le dessin.

Dans le cas on peut considérer que les niveaux atteints avant une décroissance ne sont pas significatifs, et qu'il faut mieux les supprimer en dessinant un arbre hiérarchique non binaire (dessin de droite).

5. Cas particulier d'un ordre latéral imposé

5.1 Rappels

Lorsqu'on représente une distance ultramétrique sous forme d'un arbre, on dispose d'une certaine liberté sur l'ordre latéral des objets. En effet, à chaque fois qu'on agrège deux sous-arbres, on peut choisir arbitrairement celui des deux qu'on place au dessus.

Malgré tout, les chances pour qu'un ordre a priori soit compatible avec une ultramétrique donnée sont infimes. Parmi les **n!** permutations, seulement $\mathbf{2^n}$ sont compatibles avec le tracé d'arbre binaire donné.

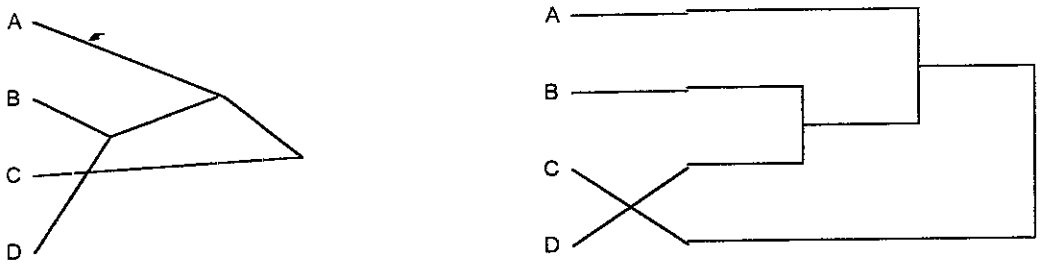
5.2 Croisement d'une ultramétrique et d'un ordre

Lorsqu'on dispose d'un ordre sur les objets classés, il est tentant de l'intégrer au dessin pour avoir une vision croisée de ces deux structures et pour améliorer la lisibilité du graphique dans le cas où cet ordre est " naturel ".

Certains ont développé des algorithmes relativement complexes pour calculer l'ordre latéral, compatible avec une ultramétrique donnée, **le plus proche** d'un ordre donné.

Je propose dans le dessin de gauche une autre approche, simple à mettre en œuvre et à mon avis plus informative, consistant à respecter l'ordre latéral donné et à observer là ou les branches de l'arbre se croisent.

Je préfère dans ce cas la forme " en bec ", mais dans certains cas, la position des nœuds peut poser un problème.

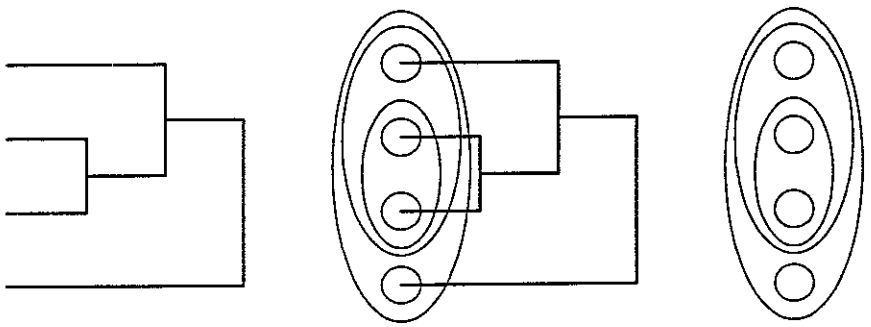


Dans la représentation de droite on applique un algorithme " d'ordre latéral le plus proche ", et on complète de dessin avec la bijection entre le résultat et l'ordre latéral a priori ; ce qui permet d'estimer le résultat.

6. Représentation alternative

6.1 Représentation en diagramme de Venn emboîtés

C'est une représentation qui met en évidence la vision des classes:



Cette représentation possède à la fois des qualités et des défauts.

Qualités :

- Totalement intuitive, elle est directement compréhensible sans besoin de règles de lecture ;
- Elle peut se faire sur un " fond " bidimensionnel. (Qui n'a jamais représenté des classes en faisant des " patates " sur un plan factoriel ?)

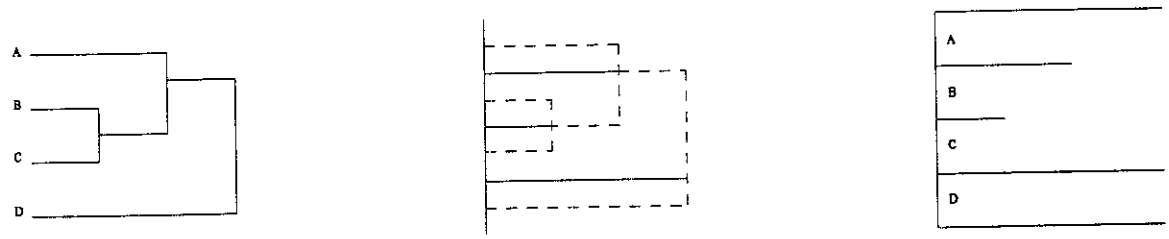
Défauts :

- Elle est uniquement qualitative, sauf si l'on fait intervenir des épaisseurs de traits (peu lisibles) ou des cotes sur les " courbes de niveaux " ;
- Elle est difficile à créer automatiquement par logiciel. Elle est habituellement dessinée à main levée.

6.2 Représentation alternative en peigne

Voici maintenant une représentation qui met l'accent sur la vision des distances.

Le graphique de droite est construit en faisant comme une empreinte de l'arbre habituel.



La distance ultramétrique entre deux objets est donnée par la plus longue barre qu'il faut " sauter " pour aller d'un point à l'autre. Cette représentation est **exactement** équivalente à la représentation classique en arbre. Elle est particulièrement adaptée aux très grands listages dont elle permet une lecture locale.

Elle peut être dessinée sur une imprimante non graphique. Cf. le programme SAVANE.

7. Programmes informatiques

Des sous programme en FORTRAN existent pour dessiner certaines de ces représentations sur des imprimantes non graphiques. D'autres, en cours de développement, s'appuient sur la norme graphique GKS. Ils sont installés sur le FTP anonyme de l'Inrets.

Accès :	ftp.inrets.fr	
chemin :	pub/graphique/arbre	
sous programme :	**baobab.f**	dessin de l'arbre classique
	savane.f	variante en peigne
	labour.f	préparation des données pour un dessin graphique
	elague.f	transformation d'une structure avec " inversions " en une structure hiérarchique non binaire.